Donald E. Kirk

Optimal Control Theory

An Introduction

Optimal Control Theory

An Introduction

Donald E. Kirk

Professor Emeritus of Electrical Engineering
San José State University
San José, California

Dover Publications, Inc.
Mineola, New York

Bibliographical Note

This Dover edition, first published in 2004, is an unabridged republication of the thirteenth printing of the work originally published by Prentice-Hall, Inc., Englewood Cliffs, New Jersey, in 1970.

Solutions Manual

Readers who would like to receive *Solutions to Selected Exercises* for this book may request them from the publisher at the following e-mail address: **editors@doverpublications.com**.

Library of Congress Cataloging-in-Publication Data

Kirk, Donald E., 1937–
Optimal control theory : an introduction / Donald E. Kirk.
p. cm.
Originally published: Englewood Cliffs, N.J. : Prentice-Hall, 1970 (Prentice-Hall networks series)
Includes bibliographical references and index.
ISBN 0-486-43484-2 (pbk.)
1. Control theory. 2. Mathematical optimization. I. Title.

QA402.3.K52 2004
003'.5—dc22

2003070111

Manufactured in the United States of America
Dover Publications, Inc., 31 East 2nd Street, Mineola, N.Y. 11501

Preface

Optimal control theory—which is playing an increasingly important role in the design of modern systems—has as its objective the maximization of the return from, or the minimization of the cost of, the operation of physical, social, and economic processes.

This book introduces three facets of optimal control theory—dynamic programming, Pontryagin's minimum principle, and numerical techniques for trajectory optimization—at a level appropriate for a first- or second-year graduate course, an undergraduate honors course, or for directed self-study. A reasonable proficiency in the use of state variable methods is assumed; however, this and other prerequisites are reviewed in Chapter 1. In the interest of flexibility, the book is divided into the following parts:

Part I: Describing the System and Evaluating Its Performance (Chapters 1 and 2)
Part II: Dynamic Programming (Chapter 3)
Part III: The Calculus of Variations and Pontryagin's Minimum Principle (Chapters 4 and 5)
Part IV: Iterative Numerical Techniques for Finding Optimal Controls and Trajectories (Chapter 6)
Part V: Conclusion (Chapter 7)

Because of the simplicity of the concept, dynamic programming (Part II) is presented before Pontryagin's minimum principle (Part III), thus enabling

the reader to solve meaningful problems at an early stage, and providing motivation for the material which follows. Parts II and III are self-contained; they may be studied in either order, or either may be omitted without affecting the treatment in the other. The problems provided in Parts I through IV are designed to introduce additional topics as well as to illustrate the basic concepts.

My experience indicates that it is possible to discuss, at a moderate pace, Chapters 1 through 4, Sections 5.1 through 5.3, and parts of Sections 5.4 and 5.5 in a one-quarter, four-credit-hour course. This material provides adequate background for reading the remainder of the book and other literature on optimal control theory. To study the entire book, a course of one semester's duration is recommended.

My thanks go to Professor Robert D. Strum for encouraging me to undertake the writing of this book, and for his helpful comments along the way. I also wish to express my appreciation to Professor John R. Ward for his constructive criticism of the presentation. Professor Charles H. Rothauge, Chairman of the Electrical Engineering Department at the Naval Postgraduate School, aided my efforts by providing a climate favorable for preparing and testing the manuscript. I thank Professors Jose B. Cruz, Jr., William R. Perkins, and Ronald A. Rohrer for introducing optimal control theory to me at the University of Illinois; undoubtedly their influence is reflected in this book. The valuable comments made by Professors James S. Demetry, Gene F. Franklin, Robert W. Newcomb, Ronald A. Rohrer, and Michael K. Sain are also gratefully acknowledged. In proofreading the manuscript I received generous assistance from my wife, Judy, and from Lcdr. D. T. Cowdrill and Lcdr. R. R. Owens, USN. Perhaps my greatest debt of gratitude is to the students whose comments were invaluable in preparing the final version of the book.

DONALD E. KIRK

Carmel, California

Contents

I

Describing the System and Evaluating Its Performance

1

Introduction

Classical control system design is generally a trial-and-error process in which various methods of analysis are used iteratively to determine the design parameters of an "acceptable" system. Acceptable performance is generally defined in terms of time and frequency domain criteria such as rise time, settling time, peak overshoot, gain and phase margin, and bandwidth. Radically different performance criteria must be satisfied, however, by the complex, multiple-input, multiple-output systems required to meet the demands of modern technology. For example, the design of a spacecraft attitude control system that minimizes fuel expenditure is not amenable to solution by classical methods. A new and direct approach to the synthesis of these complex systems, called optimal control theory, has been made feasible by the development of the digital computer.

The objective of optimal control theory is *to determine the control signals that will cause a process to satisfy the physical constraints and at the same time minimize (or maximize) some performance criterion.* Later, we shall give a more explicit mathematical statement of "the optimal control problem," but first let us consider the matter of problem formulation.

1.1 PROBLEM FORMULATION

The axiom "A problem well put is a problem half solved" may be a slight exaggeration, but its intent is nonetheless appropriate. In this section, we

shall review the important aspects of problem formulation, and introduce the notation and nomenclature to be used in the following chapters.

The formulation of an optimal control problem requires:

1. A mathematical description (or model) of the process to be controlled.
2. A statement of the physical constraints.
3. Specification of a performance criterion.

The Mathematical Model

A nontrivial part of any control problem is modeling the process. The objective is to obtain the simplest mathematical description that adequately predicts the response of the physical system to all anticipated inputs. Our discussion will be restricted to systems described by ordinary differential equations (in state variable form).† Thus, if

$$x_1(t), x_2(t), \ldots, x_n(t)$$

are the *state variables* (or simply the *states*) of the process at time t, and

$$u_1(t), u_2(t), \ldots, u_m(t)$$

are *control inputs* to the process at time t, then the system may be described by n first-order differential equations

$$\begin{aligned}
\dot{x}_1(t) &= a_1(x_1(t), x_2(t), \ldots, x_n(t), u_1(t), u_2(t), \ldots, u_m(t), t) \\
\dot{x}_2(t) &= a_2(x_1(t), x_2(t), \ldots, x_n(t), u_1(t), u_2(t), \ldots, u_m(t), t) \\
&\vdots \\
\dot{x}_n(t) &= a_n(x_1(t), x_2(t), \ldots, x_n(t), u_1(t), u_2(t), \ldots, u_m(t), t).\ddagger
\end{aligned} \tag{1.1-1}$$

We shall define

$$\mathbf{x}(t) \triangleq \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

as the *state vector* of the system, and

† The reader will find the concepts much the same for discrete systems (see [A-1]).

‡ Note that $\dot{x}_i(t)$ is in general a nonlinear time-varying function a_i of the states, the control inputs, and time.

$$\mathbf{u}(t) \triangleq \begin{bmatrix} u_1(t) \\ u_2(t) \\ \cdot \\ \cdot \\ \cdot \\ u_m(t) \end{bmatrix}$$

as the *control vector*. The state equations can then be written

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t), \tag{1.1-1a}$$

where the definition of **a** is apparent by comparison with (1.1-1).

Figure 1-1 A simplified control problem

Example 1.1-1. The car shown parked in Fig. 1-1 is to be driven in a straight line away from point O. The distance of the car from O at time t is denoted by $d(t)$. To simplify the model, let us approximate the car by a unit point mass that can be accelerated by using the throttle or decelerated by using the brake. The differential equation is

$$\ddot{d}(t) = \alpha(t) + \beta(t), \tag{1.1-2}$$

where the control α is throttle acceleration and-β is braking deceleration. Selecting position and velocity as state variables, that is,

$$x_1(t) \triangleq d(t) \quad \text{and} \quad x_2(t) \triangleq \dot{d}(t),$$

and letting

$$u_1(t) \triangleq \alpha(t) \quad \text{and} \quad u_2(t) \triangleq \beta(t),$$

we find that the state equations become

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u_1(t) + u_2(t), \end{aligned} \tag{1.1-3}$$

or, using matrix notation,

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \mathbf{u}(t). \tag{1.1-3a}$$

This is the mathematical model of the process in state form.

Before we move on to the matter of physical constraints, let us consider two definitions that will be useful later. Let the system be described by Eq. (1.1-1a) for $t \in [t_0, t_f]$.†

DEFINITION 1-1

A history of control input values during the interval $[t_0, t_f]$ is denoted by **u** and is called a *control history*, or simply a *control.*

DEFINITION 1-2

A history of state values in the interval $[t_0, t_f]$ is called a *state trajectory* and is denoted by **x**.

The terms "history," "curve," "function," and "trajectory" will be used interchangeably. It is most important to keep in mind the difference between a *function* and *the value of a function.* Figure 1-2 shows a single-valued function of time which is denoted by x. The *value* of the function at time t_1 is denoted by $x(t_1)$.

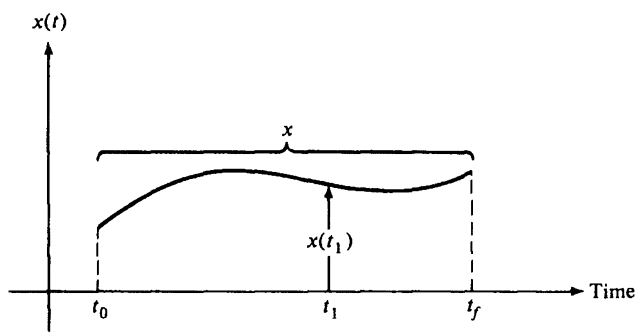


Figure 1-2 A function, x, and its value at time t_1, $x(t_1)$

Physical Constraints

After we have selected a mathematical model, the next step is to define the physical constraints on the state and control values. To illustrate some typical constraints, let us return to the automobile whose model was determined in Example 1.1-1.

Example 1.1-2. Consider the problem of driving the car in Fig. 1-1 between the points O and e. Assume that the car starts from rest and stops upon reaching point e.

† This notation means $t_0 \leq t \leq t_f$.

First let us define the state constraints. If t_0 is the time of leaving O, and t_f is the time of arrival at e, then, clearly,

$$\begin{aligned} x_1(t_0) &= 0 \\ x_1(t_f) &= e. \end{aligned} \tag{1.1-4}$$

In addition, since the automobile starts from rest and *stops* at e,

$$\begin{aligned} x_2(t_0) &= 0 \\ x_2(t_f) &= 0. \end{aligned} \tag{1.1-5}$$

In matrix notation these *boundary conditions* are

$$\mathbf{x}(t_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0} \quad \text{and} \quad \mathbf{x}(t_f) = \begin{bmatrix} e \\ 0 \end{bmatrix}. \tag{1.1-6}$$

If we assume that the car does not back up, then the additional constraints

$$\begin{aligned} 0 &\leq x_1(t) \leq e \\ 0 &\leq x_2(t) \end{aligned} \tag{1.1-7}$$

are also imposed.

What are the constraints on the control inputs (acceleration)? We know that the acceleration is bounded by some upper limit which depends on the capability of the engine, and that the maximum deceleration is limited by the braking system parameters. If the maximum acceleration is $M_1 > 0$, and the maximum deceleration is $M_2 > 0$, then the controls must satisfy

$$\begin{aligned} 0 &\leq u_1(t) \leq M_1 \\ -M_2 &\leq u_2(t) \leq 0. \end{aligned} \tag{1.1-8}$$

In addition, if the car starts with G gallons of gas and there are no service stations on the way, another constraint is

$$\int_{t_0}^{t_f} [k_1 u_1(t) + k_2 x_2(t)]\, dt \leq G \tag{1.1-9}$$

which assumes that the rate of gas consumption is proportional to both acceleration and speed with constants of proportionality k_1 and k_2.

Now that we have an idea of typical constraints that may be encountered, let us make these concepts more precise.

DEFINITION 1-3

A control history which satisfies the control constraints during the entire time interval $[t_0, t_f]$ is called an *admissible control.*

We shall denote the set of admissible controls by U, and the notation $\mathbf{u} \in U$ means that the control history $\mathbf{u}$ is admissible.

To illustrate the concept of admissibility Fig. 1-3 shows four possible acceleration histories for Example 1.1-2. $u_1^{(2)}$ and $u_1^{(4)}$ are not admissible;

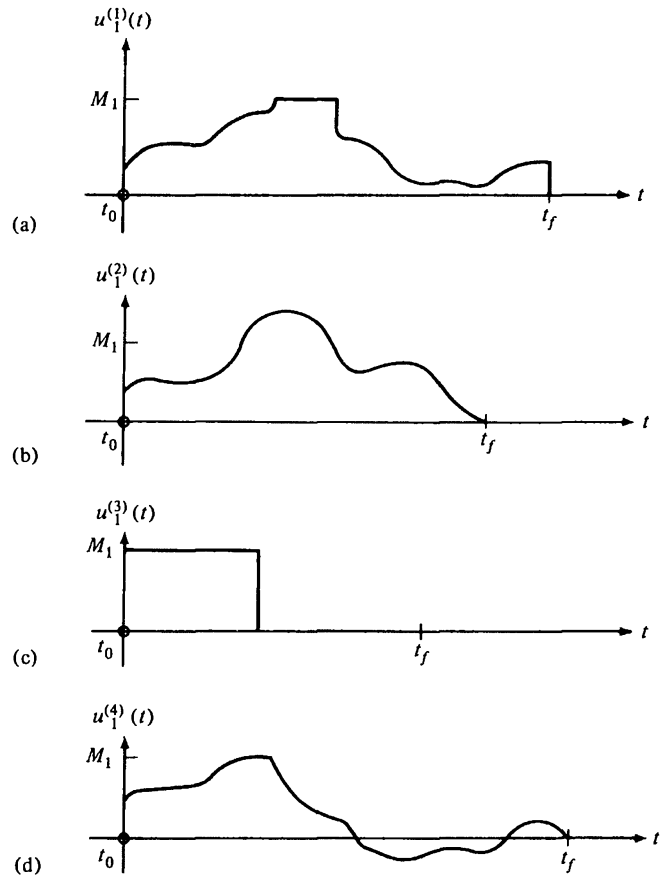


Figure 1-3 Some acceleration histories

$u_1^{(1)}$ and $u_1^{(3)}$ are admissible if they satisfy the consumed-fuel constraint of Eq. (1.1-9). In this example, the set of admissible controls U is defined by the inequalities in (1.1-8) and (1.1-9).

DEFINITION 1-4

A state trajectory which satisfies the state variable constraints during the entire time interval $[t_0, t_f]$ is called an *admissible trajectory.*

The set of admissible state trajectories will be denoted by X, and $\mathbf{x} \in X$ means that the trajectory $\mathbf{x}$ is admissible.

In Example 1.1-2 the set of admissible state trajectories X is specified by the conditions given in Eqs. (1.1-6), (1.1-7), and (1.1-9). In general, the final state of a system will be required to lie in a specified region S of the $(n + 1)$-dimensional state-time space. We shall call S the *target set*. If the final state and the final time are fixed, then S is a point. In the automobile problem of Example 1.1-2 the target set was the line shown in Fig. 1-4(a). If the automobile had been required to arrive within three feet of e with zero terminal velocity, the target set would have been as shown in Fig. 1-4(b).

Admissibility is an important concept, because it reduces the range of values that can be assumed by the states and controls. Rather than consider all control histories and their trajectories to see which are best (according to some criterion), we investigate only those trajectories and controls that are admissible.

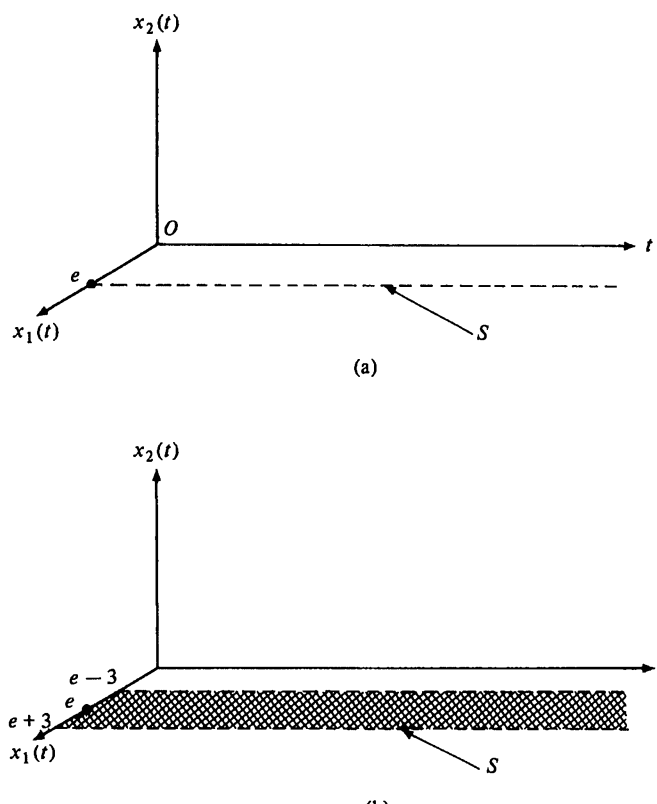


Figure 1-4 (a) The target set for Example 1.1-2. (b) The target set defined by $|x_1(t) - e| \leq 3$, $x_2(t) = 0$

The Performance Measure

In order to evaluate the performance of a system quantitatively, the designer selects a performance measure. An *optimal control* is defined as one that *minimizes* (or maximizes) the performance measure. In certain cases the problem statement may clearly indicate what to select for a performance measure, whereas in other problems the selection is a subjective matter. For example, the statement, "Transfer the system from point A to point B as quickly as possible," clearly indicates that elapsed time is the performance measure to be minimized. On the other hand, the statement, "Maintain the position and velocity of the system near zero with a small expenditure of control energy," does not instantly suggest a unique performance measure. In such problems the designer may be required to try several performance measures before selecting one which yields what he considers to be optimal performance. We shall discuss the selection of a performance measure in more detail in Chapter 2.

> **Example 1.1-3.** Let us return to the automobile problem begun in Example 1.1-1. The state equations and physical constraints have been defined; now we turn to the selection of a performance measure. Suppose the objective is to make the car reach point e as quickly as possible; then the performance measure J is given by
>
> $$J = t_f - t_0. \tag{1.1-10}$$

In all that follows it will be assumed that the performance of a system is evaluated by a measure of the form

$$J = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt, \tag{1.1-11}$$

where t_0 and t_f are the initial and final time; h and g are scalar functions. t_f may be specified or "free," depending on the problem statement.

Starting from the initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ and applying a control signal $\mathbf{u}(t)$, for $t \in [t_0, t_f]$, causes a system to follow some state trajectory; the performance measure assigns a unique real number to each trajectory of the system.

With the background material we have accumulated it is now possible to present an explicit statement of "the optimal control problem."

The Optimal Control Problem

The theory developed in the subsequent chapters is aimed at solving the following problem.

Find an *admissible control* $\mathbf{u}^*$ which causes the system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{1.1-12}$$

to follow an *admissible trajectory* $\mathbf{x}^*$ that *minimizes* the performance measure

$$J = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt. \tag{1.1-13}$$

$\mathbf{u}^*$ is called an *optimal control* and $\mathbf{x}^*$ an *optimal trajectory.*

Several comments are in order here. First, we may not know in advance that an optimal control *exists*; that is, it may be impossible to find a control which (a) is admissible and (b) causes the system to follow an admissible trajectory. Since existence theorems are in rather short supply, we shall, in most cases, attempt to find an optimal control rather than try to prove that one exists.

Second, even if an optimal control exists, it may not be *unique*. Nonunique optimal controls may complicate computational procedures, but they do allow the possibility of choosing among several controller configurations. This is certainly helpful to the designer, because he can then consider other factors, such as cost, size, reliability, etc., which may not have been included in the performance measure.

Third, when we say that $\mathbf{u}^*$ causes the performance measure to be minimized, we mean that

$$\begin{aligned} J^* &\triangleq h(\mathbf{x}^*(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\, dt \\ &\leq h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt \end{aligned} \tag{1.1-14}$$

for all $\mathbf{u} \in U$, which make $\mathbf{x} \in X$. The above inequality states that an optimal control and its trajectory cause the performance measure to have a value smaller than (or perhaps equal to) the performance measure for *any other* admissible control and trajectory. Thus, we are seeking the *absolute* or *global minimum* of J, not merely *local minima.* Of course, one way to find the global minimum is to determine all of the local minima and then simply pick out one (or more) that yields the smallest value for the performance measure.

It may be helpful to visualize the optimization as shown in Fig. 1-5. $u^{(1)}$, $u^{(2)}$, $u^{(3)}$, and $u^{(4)}$ are "points" at which J has local, or relative, minima; $u^{(1)}$ is the "point" where J has its global, or absolute, minimum.

Finally, observe that if the objective is to maximize some measure of system performance, the theory we shall develop still applies because this

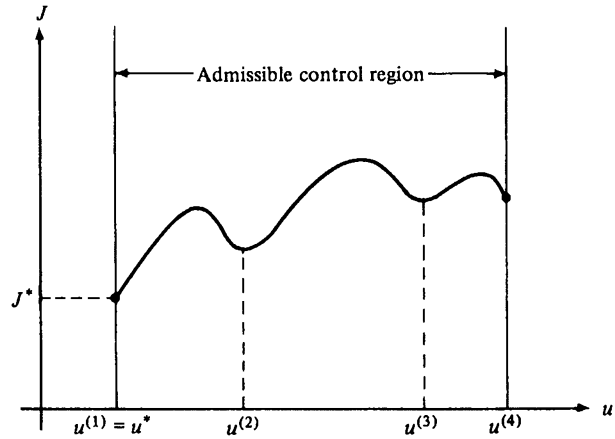


Figure 1-5 A representation of the optimization problem

is the same as minimizing the negative of this performance measure. Henceforth, we shall speak, with no lack of generality, of minimizing the performance measure.

Example 1.1-4. To illustrate a complete problem formulation, let us now summarize the results of Example 1.1-1, using the notation and definitions which have been developed.

The state equations are

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u_1(t) + u_2(t). \end{aligned} \tag{1.1-3}$$

The set of admissible states X is partially specified by the boundary conditions

$$\mathbf{x}(t_0) = \mathbf{0}, \qquad \mathbf{x}(t_f) = \begin{bmatrix} e \\ 0 \end{bmatrix}$$

and the inequalities

$$\begin{aligned} 0 &\leq x_1(t) \leq e \\ 0 &\leq x_2(t). \end{aligned} \tag{1.1-7}$$

The set of admissible controls U is partially defined by the constraints

$$\begin{aligned} 0 &\leq u_1(t) \leq M_1 \\ -M_2 &\leq u_2(t) \leq 0. \end{aligned} \tag{1.1-8}$$

The inequality constraint

$$\int_{t_0}^{t_f} [k_1 u_1(t) + k_2 x_2(t)]\, dt \leq G \tag{1.1-9}$$

completes the description of the admissible states and controls.

The solution to this problem (which is left as an exercise for the reader at the end of Chapter 5) is shown in Fig. 1-6 for the situation where $M_1 = M_2 \triangleq M$. We have also assumed that the car has enough fuel available to reach point e using the control shown.

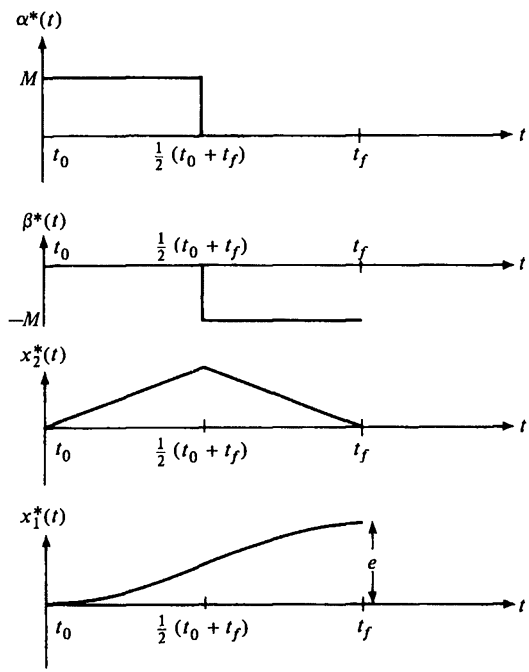


Figure 1-6 The optimal control and trajectory for the automobile problem

Example 1.1-5. Let us now consider what would happen if the preceding problem had been improperly formulated. Suppose that the control constraints had not been recognized. If we let

$$\alpha(t) + \beta(t) = e\frac{d}{dt}[\delta(t - t_0)] \tag{1.1-15}$$

where $\delta(t - t_0)$ is a unit impulse function that occurs at time t_0,† then

$$x_2(t) = e\,\delta(t - t_0) \tag{1.1-16}$$

and

$$x_1(t) = e\,\mathbb{1}(t - t_0) \tag{1.1-17}$$

[$\mathbb{1}(t - t_0)$ represents a unit step function at $t = t_0$]. Figure 1-7 shows the state trajectory which results from applying the "optimal" control in (1.1-15). Unfortunately, although the desired transfer from point O

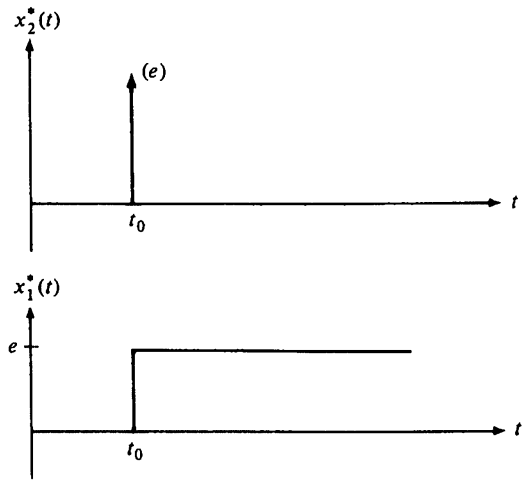


Figure 1-7 The optimal trajectory resulting from unconstrained controls

to point e is accomplished in infinitesimal time, the control required, apart from being rather unsafe, is physically impossible! Thus, we see the importance of correctly formulating problems before attempting their solution.

Form of the Optimal Control

DEFINITION 1-5

If a functional relationship of the form

$$\mathbf{u}^*(t) = \mathbf{f}(\mathbf{x}(t), t)‡ \tag{1.1-18}$$

† See reference [Z-1].

‡ Here we write $\mathbf{x}(t)$ instead of $\mathbf{x}^*(t)$ to emphasize that the control law is optimal for *all* admissible $\mathbf{x}(t)$, not just for some special state value at time t.

can be found for the optimal control at time t, then the function $\mathbf{f}$ is called the *optimal control law*, or the *optimal policy*.†

Notice that Eq. (1.1-18) implies that $\mathbf{f}$ is a rule which determines the optimal control at time t for *any* (admissible) state value at time t. For example, if

$$\mathbf{u}^*(t) = \mathbf{F}\mathbf{x}(t), \tag{1.1-19}$$

where $\mathbf{F}$ is an $m \times n$ matrix of real constants, then we would say that the optimal control law is linear, time-invariant feedback of the states.

DEFINITION 1-6

If the optimal control is determined as a function of time for a specified initial state value, that is,

$$\mathbf{u}^*(t) = \mathbf{e}(\mathbf{x}(t_0), t), \tag{1.1-20}$$

then the optimal control is said to be in *open-loop* form.

Thus the optimal open-loop control is optimal only for a *particular* initial state value, whereas, if the optimal control *law* is known, the optimal control history starting from *any* state value can be generated.

Conceptually, it is helpful to imagine the difference between an optimal control law and an open-loop optimal control as shown in Fig. 1-8; notice,

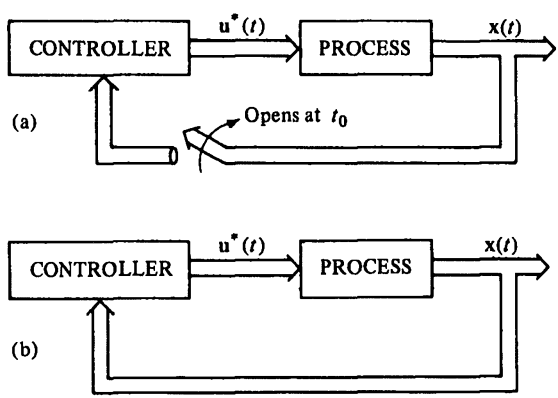


Figure 1-8 (a) Open-loop optimal control. (b) Optimal control law

however, that the mere presence of connections from the states to a controller does not, in general, guarantee an *optimal* control law.‡

† The terms *optimal feedback control*, *closed-loop optimal control*, and *optimal control strategy* are also often used.

‡ This is pursued further in reference [K-1].

Although engineers normally prefer closed-loop solutions to optimal control problems, there are cases when an open-loop control may be feasible. For example, in the radar tracking of a satellite, once the orbit is set very little can happen to cause an undesired change in the trajectory parameters. In this situation a pre-programmed control for the radar antenna might well be used.

A typical example of feedback control is in the classic servomechanism problem where the actual and desired outputs are compared and any deviation produces a control signal that attempts to reduce the discrepancy to zero.

1.2 STATE VARIABLE REPRESENTATION OF SYSTEMS

The starting point for optimal control investigations is a mathematical model in state variable form. In this section we shall summarize the results and notation to be used in the subsequent discussion. There are several excellent texts available for the reader who needs additional background material.†

Why Use State Variables?

Having the mathematical model in state variable form is convenient because

1. The differential equations are ideally suited for digital or analog solution.
2. The state form provides a unified framework for the study of nonlinear and linear systems.
3. The state variable form is invaluable in theoretical investigations.
4. The concept of state has strong physical motivation.

Definition of State of a System

When referring to the state of a system, we shall have the following definition in mind.

DEFINITION 1-7

The *state of a system* is a set of quantities $x_1(t), x_2(t), \ldots, x_n(t)$

† See [D-1], [O-1], [S-1], [S-2], [T-1], [W-1], [Z-1].

which if known at $t = t_0$ are determined for $t \geq t_0$ by specifying the inputs to the system for $t \geq t_0$.

System Classification

Systems are described by the terms *linear*, *nonlinear*, *time-invariant*,† and *time-varying*. We shall classify systems according to the form of their state equations.‡ For example, if a system is *nonlinear* and *time-varying*, the state equations are written

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t). \tag{1.2-1}$$

Nonlinear, *time-invariant* systems are represented by state equations of the form

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t)). \tag{1.2-2}$$

If a system is *linear* and *time-varying* its state equations are

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \tag{1.2-3}$$

where $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are $n \times n$ and $n \times m$ matrices with time-varying elements. State equations for *linear*, *time-invariant* systems have the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{1.2-4}$$

where $\mathbf{A}$ and $\mathbf{B}$ are constant matrices.

Output Equations

The physical quantities that can be measured are called the *outputs* and are denoted by $y_1(t), y_2(t), \ldots, y_q(t)$. If the outputs are *nonlinear*, *time-varying* functions of the states and controls, we write the output equations

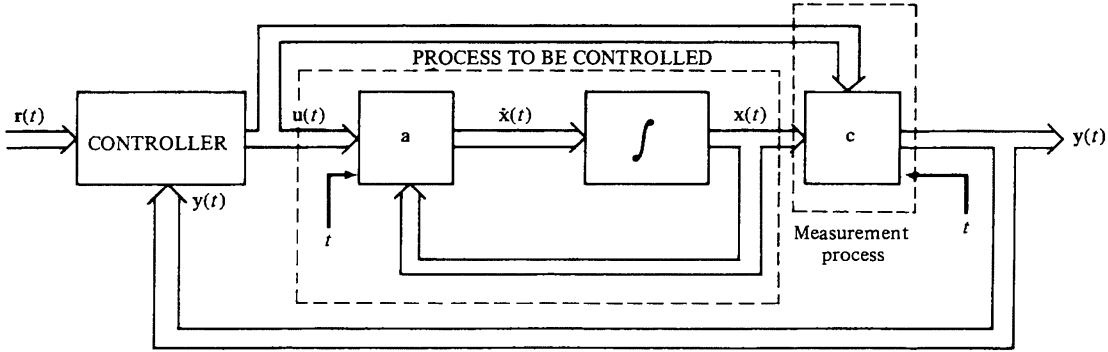
$$\mathbf{y}(t) = \mathbf{c}(\mathbf{x}(t), \mathbf{u}(t), t). \tag{1.2-5}$$

If the output is related to the states and controls by a *linear*, *time-invariant* relationship, then

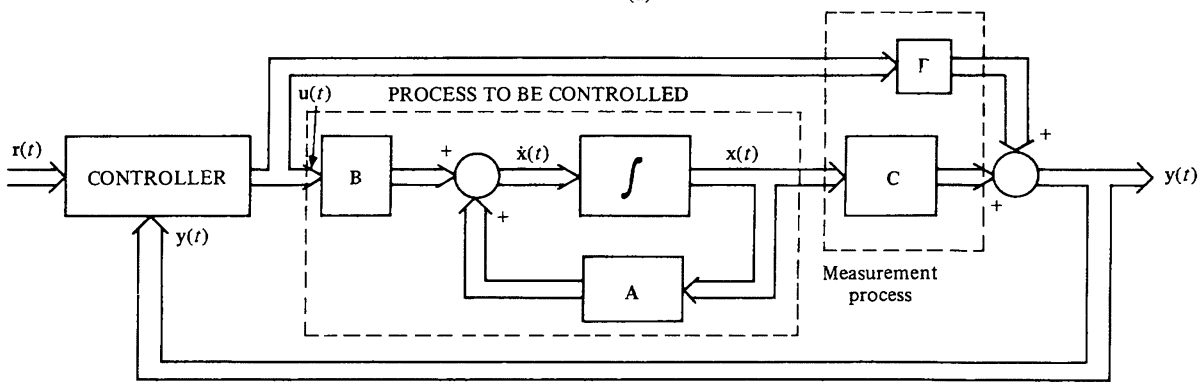
$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \tag{1.2-6}$$

where $\mathbf{C}$ and $\mathbf{D}$ are $q \times n$ and $q \times m$ constant matrices. A nonlinear, time-

† *Time-invariant*, *stationary*, and *fixed* will be used interchangeably.
‡ See Chapter 1 of [S-1] for an excellent discussion of system classification.



(a)



(b)

Figure 1-9 (a) Nonlinear system representation. (b) Linear system representation

varying system and a linear, time-invariant system are shown in Fig. 1-9. $\mathbf{r}(t)$, which has not been included in the state equations and represents any inputs that are not controlled, is called the *reference* or *command* input.

In our discussion of optimal control theory we shall make the simplifying assumption that the states are all available for measurement; that is, $\mathbf{y}(t) = \mathbf{x}(t)$.

Solution of the State Equations—Linear Systems

For linear systems the state equations (1.2-3) have the solution

$$\mathbf{x}(t) = \boldsymbol{\varphi}(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \boldsymbol{\varphi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \tag{1.2-7}$$

where $\boldsymbol{\varphi}(t, t_0)$ is the *state transition matrix*† of the system. If the system is time-invariant as well as linear, t_0 can be set equal to 0 and the solution of the state equations is given by any of the three equivalent forms

$$\mathbf{x}(t) = \mathscr{L}^{-1}\{[s\mathbf{I} - \mathbf{A}]^{-1}\mathbf{x}(0) + [s\mathbf{I} - \mathbf{A}]^{-1}\mathbf{B}\mathbf{U}(s)\}, \tag{1.2-8a}$$

$$\mathbf{x}(t) = \mathscr{L}^{-1}\{\boldsymbol{\Phi}(s)\mathbf{x}(0) + \mathbf{H}(s)\mathbf{U}(s)\}, \tag{1.2-8b}$$

$$\mathbf{x}(t) = \epsilon^{\mathbf{A}t}\mathbf{x}(0) + \epsilon^{\mathbf{A}t}\int_0^t \epsilon^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)\, d\tau, \tag{1.2-8c}$$

where $\mathbf{U}(s)$ and $\boldsymbol{\Phi}(s)$ are the Laplace transforms of $\mathbf{u}(t)$ and $\boldsymbol{\varphi}(t)$, $\mathscr{L}^{-1}\{\cdot\}$ denotes the inverse Laplace transform of $\{\cdot\}$, and $\epsilon^{\mathbf{A}t}$ is the $n \times n$ matrix

$$\epsilon^{\mathbf{A}t} \triangleq \mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \frac{1}{3!}\mathbf{A}^3t^3 + \cdots + \frac{1}{k!}\mathbf{A}^kt^k + \cdots. \tag{1.2-9}$$

Equation (1.2-8a) results when the state equations (1.2-4) are Laplace transformed and solved for $\mathbf{X}(s)$. Equation (1.2-8b) can be obtained by drawing a block diagram (or signal flow graph) of the system and applying Mason's gain formula.‡ Notice that $\mathbf{H}(s)$ is the transfer function matrix. The solution in (1.2-8c) can be found by classical methods. The equivalence of these three solutions establishes the correspondences

$$\epsilon^{\mathbf{A}t} = \mathscr{L}^{-1}\{\boldsymbol{\Phi}(s)\} = \mathscr{L}^{-1}\{[s\mathbf{I} - \mathbf{A}]^{-1}\} \triangleq \boldsymbol{\varphi}(t), \tag{1.2-10}$$

$$\begin{aligned}\epsilon^{\mathbf{A}t}\int_0^t \epsilon^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)\, d\tau &= \mathscr{L}^{-1}\{\mathbf{H}(s)\mathbf{U}(s)\} = \mathscr{L}^{-1}\{[s\mathbf{I} - \mathbf{A}]^{-1}\mathbf{B}\mathbf{U}(s)\} \\ &\triangleq \boldsymbol{\varphi}(t)\int_0^t \boldsymbol{\varphi}(-\tau)\mathbf{B}\mathbf{u}(\tau)\, d\tau.\end{aligned} \tag{1.2-11}$$

† $\boldsymbol{\varphi}(t, t_0)$ is also called the *fundamental matrix.*
‡ See [W-1].

Properties of the State Transition Matrix

It can be verified that the state transition matrix has the properties shown in Table 1-1 for all t, t_0, t_1, and t_2.

Table 1-1 PROPERTIES OF THE LINEAR SYSTEM STATE TRANSITION MATRIX

Time-invariant systems	*Time-varying systems*
$\varphi(0) = \mathbf{I}$	$\varphi(t, t) = \mathbf{I}$
$\varphi(t_2 - t_1)\varphi(t_1 - t_0) = \varphi(t_2 - t_0)$	$\varphi(t_2, t_1)\varphi(t_1, t_0) = \varphi(t_2, t_0)$
$\varphi^{-1}(t_2 - t_1) = \varphi(t_1 - t_2)$	$\varphi^{-1}(t_2, t_1) = \varphi(t_1, t_2)$
$\frac{d}{dt}\varphi(t) = \mathbf{A}\varphi(t)$	$\frac{d}{dt}\varphi(t, t_0) = \mathbf{A}(t)\varphi(t, t_0)$

Determination of the State Transition Matrix

For systems having a constant **A** matrix, the state transition matrix, $\varphi(t)$, can be determined by any of the following methods:

1. Inverting the matrix $[s\mathbf{I} - \mathbf{A}]$ and finding the inverse Laplace transform of each element.
2. Using Mason's gain formula to find $\Phi(s)$ from a block diagram or signal flow graph of the system [the ijth element of the matrix $\Phi(s)$ is given by the transmission $X_i(s)/x_j(0)$] and evaluating the inverse Laplace transform of $\Phi(s)$.
3. Evaluating the matrix expansion

$$\epsilon^{\mathbf{A}t} \triangleq \mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \frac{1}{3!}\mathbf{A}^3t^3 + \cdots + \frac{1}{k!}\mathbf{A}^kt^k + \cdots.\dagger \quad (1.2\text{-}9)$$

For high-order systems ($n > 4$), evaluating $\epsilon^{\mathbf{A}t}$ numerically (with the aid of a digital computer) is the most feasible of these methods.

For systems having a time-varying **A** matrix the state transition matrix can be found by numerical integration of the matrix differential equation

$$\frac{d}{dt}\varphi(t, t_0) = \mathbf{A}(t)\varphi(t, t_0) \quad (1.2\text{-}12)$$

with the initial condition $\varphi(t_0, t_0) = \mathbf{I}$.

† Although a digital computer program for the evaluation of this expansion is easy to write, the running time may be excessive because of convergence properties of the series. For a discussion of more efficient numerical techniques see [O-1], p. 315ff.

Controllability and Observability†

Consider the system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{1.2-13}$$

for $t \geq t_0$ with initial state $\mathbf{x}(t_0) = \mathbf{x}_0$.

DEFINITION 1-8

If there is a finite time $t_1 \geq t_0$ and a control $\mathbf{u}(t)$, $t \in [t_0, t_1]$, which transfers the state $\mathbf{x}_0$ to the origin at time t_1, the state $\mathbf{x}_0$ is said to be *controllable at time* t_0. If all values of $\mathbf{x}_0$ are controllable for all t_0, the system is *completely controllable*, or simply *controllable.*

Controllability is very important, because we shall consider problems in which the goal is to transfer a system from an arbitrary initial state to the origin while minimizing some performance measure; thus, controllability of the system is a necessary condition for the existence of a solution.

Kalman‡ has shown that a *linear, time-invariant* system is controllable if and only if the $n \times mn$ matrix

$$\mathbf{E} \triangleq \left[\mathbf{B} \vdots \mathbf{AB} \vdots \mathbf{A}^2\mathbf{B} \vdots \cdots \vdots \mathbf{A}^{n-1}\mathbf{B}\right]$$

has rank n. If there is only one control input ($m = 1$), a necessary and sufficient condition for controllability is that the $n \times n$ matrix $\mathbf{E}$ be nonsingular.

The concept of observability is defined by considering the system (1.2-13) with the control $\mathbf{u}(t) = \mathbf{0}$ for $t \geq t_0$.§

DEFINITION 1-9

If by observing the output $\mathbf{y}(t)$ during the finite time interval $[t_0, t_1]$ the state $\mathbf{x}(t_0) = \mathbf{x}_0$ can be determined, the state $\mathbf{x}_0$ is said to be *observable at time* t_0. If all states $\mathbf{x}_0$ are observable for every t_0, the system is called *completely observable*, or simply *observable.*

Analogous to the test for controllability, it can be shown that the *linear, time-invariant* system

$$\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{Bu}(t) \tag{1.2-14}$$

$$\mathbf{y}(t) = \mathbf{Cx}(t) \tag{1.2-15}$$

† See [K-2], [K-3].

‡ See [K-2].

§ If the system is linear and time-invariant, $\mathbf{u}$ can be any known function—see [Z-1], p. 502.

is observable if and only if the $n \times qn$ matrix

$$\mathbf{G} \triangleq \left[\mathbf{C}^T \vdots \mathbf{A}^T\mathbf{C}^T \vdots (\mathbf{A}^T)^2\mathbf{C}^T \vdots \cdots \vdots (\mathbf{A}^T)^{n-1}\mathbf{C}^T\right]$$

has rank n. If there is only one output ($q = 1$) $\mathbf{G}$ is an $n \times n$ matrix and a necessary and sufficient condition for observability is that $\mathbf{G}$ be nonsingular. Since we have made the simplifying assumption that all of the states can be physically measured ($\mathbf{y}(t) = \mathbf{x}(t)$), the question of observability will not arise in our subsequent discussion.

1.3 CONCLUDING REMARKS

In control system design, the ultimate objective is to obtain a controller that will cause a system to perform in a desirable manner. Usually, other factors, such as weight, volume, cost, and reliability also influence the controller design, and compromises between performance requirements and implementation considerations must be made. Classical design procedures are best suited for *linear, single-input, single-output systems with zero initial conditions.* Using simulation, mathematical analysis, or graphical methods, the designer evaluates the effects of inserting various physical devices into the system. By trial and error either an acceptable controller design is obtained, or the designer concludes that the performance requirements cannot be satisfied.

Many complex aerospace problems that are not amenable to classical techniques have been solved by using optimal control theory. However, we are forced to admit that optimal control theory does not, at the present time, constitute a generally applicable procedure for the design of simple controllers. The optimal control law, if it can be obtained, usually requires a digital computer for implementation (an important exception is the linear regulator problem discussed in Section 5.2), and *all* of the states must be available for feedback to the controller. These limitations may preclude implementation of the optimal control law; however, the theory of optimal control is still useful, because

1. Knowing the optimal control law may provide insight helpful in designing a suboptimal, but easily implemented controller.
2. The optimal control law provides a standard for evaluating proposed suboptimal designs. In other words, by knowing the optimal control law we have a quantitative measure of performance degradation caused by using a suboptimal controller.

REFERENCES

A-1 Athans, M., "The Status of Optimal Control Theory and Applications for Deterministic Systems," *IEEE Trans. Automatic Control* (1966), 580–596.

D-1 Derusso, P. M., R. J. Roy, and C. M. Close, *State Variables for Engineers.* New York: John Wiley & Sons, Inc., 1965.

K-1 Kliger, I., "On Closed-Loop Optimal Control," *IEEE Trans. Automatic Control* (1965), 207.

K-2 Kalman, R. E., "On the General Theory of Control Systems," *Proc. First IFAC Congress* (1960), 481–493.

K-3 Kalman, R. E., Y. C. Ho, and K. S. Narendra, "Controllability of Linear Dynamical Systems," in *Contributions to Differential Equations*, Vol. 1. New York: John Wiley & Sons, Inc., 1962.

O-1 Ogata, K., *State Space Analysis of Control Systems.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1967.

S-1 Schwarz, R. J., and B. Friedland, *Linear Systems.* New York: McGraw-Hill, Inc., 1965.

S-2 Schultz, D. G., and J. L. Melsa, *State Functions and Linear Control Systems.* New York: McGraw-Hill, Inc., 1967.

T-1 Timothy, L. K., and B. E. Bona, *State Space Analysis: An Introduction.* New York: McGraw-Hill, Inc., 1968.

W-1 Ward, J. R., and R. D. Strum, *State Variable Analysis* (A Programmed Text). Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1970.

Z-1 Zadeh, L. A., and C. A. Desoer, *Linear System Theory: The State Space Approach.* New York: McGraw-Hill, Inc., 1963.

PROBLEMS

1-1. The tanks A and B shown in Fig. 1-P1 each have a capacity of 50 gal. Both tanks are filled at $t = 0$, tank A with 60 lb of salt dissolved in water, and

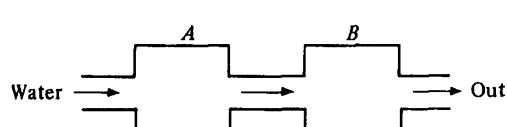
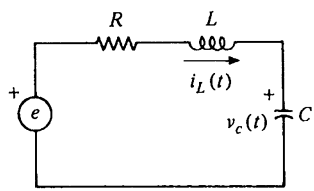


Figure 1-P1

tank B with water. Fresh water enters tank A at the rate of 8 gal/min, the mixture of salt and water (assumed uniform) leaves A and enters B at the rate of 8 gal/min, and the flow is incompressible. Let $q(t)$ and $p(t)$ be the number of pounds of salt contained in tanks A and B, respectively.

(a) Write a set of state equations for the system.
(b) Draw a block diagram (or signal flow graph) for the system.
(c) Find the state transition matrix $\boldsymbol{\varphi}(t)$.
(d) Determine $q(t)$ and $p(t)$ for $t \geq 0$.

1-2. (a) Using the capacitor voltage $v_c(t)$ and the inductor current $i_L(t)$ as states, write state equations for the RLC series circuit shown in Fig. 1-P2.



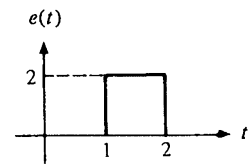


Figure 1-P2

(b) Find the state transition matrix $\boldsymbol{\varphi}(t)$ if $R = 3\ \Omega$, $L = 1$ H, $C = \frac{1}{2}$ F.
(c) If $v_c(0) = 0$, $i_L(0) = 0$, and $e(t)$ is as shown, determine $v_c(t)$ and $i_L(t)$ for $t \geq 0$.

1-3. (a) Write a set of state equations for the mechanical system shown in Fig. 1-P3. The applied force is $f(t)$, the block has mass M, the spring constant is K, and the coefficient of viscous friction is B. The displacement of the block, $y(t)$, is measured from the equilibrium position with no force applied.

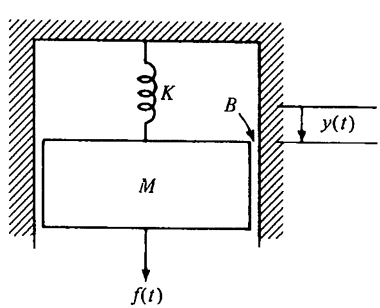


Figure 1-P3

(b) Draw a block diagram (or signal flow graph) for the system.
(c) Let $M = 1$ kg, $K = 2$ N/m, $B = 2$ N/m/sec, and determine the state transition matrix $\boldsymbol{\varphi}(t)$.
(d) If $y(0) = 0.2$ m, $\dot{y}(0) = 0$, and $f(t) = 2\epsilon^{-2t}$ N for $t \geq 0$, determine $y(t)$ and $\dot{y}(t)$ for $t \geq 0$.

1-4. Write a set of state equations for the electrical network shown in Fig. 1-P4.

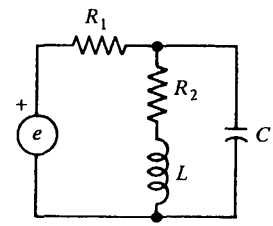


Figure 1-P4

1-5. Write state equations for the mechanical system in Fig. 1-P5. λ is the applied torque, I is the moment of inertia, K is the spring constant, and B is the coefficient of viscous friction. The angular displacement $\theta(t)$ is measured from the equilibrium position with no torque applied.

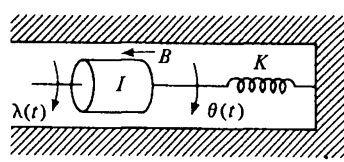


Figure 1-P5

1-6. A chemical mixing process is shown in Fig. 1-P6. Water enters the tanks at rates of $w_1(t)$ and $w_2(t)$ ft³/min, and $m(t)$ ft³/min of dye enters tank 1. $v_1(t)$ and $v_2(t)$ ft³ of dye are present in tanks 1 and 2 at time t. The tanks have cross-sectional areas α_1 and α_2. Assume that the flow rate between the two tanks, $q(t)$, is proportional to the difference in head with propor-

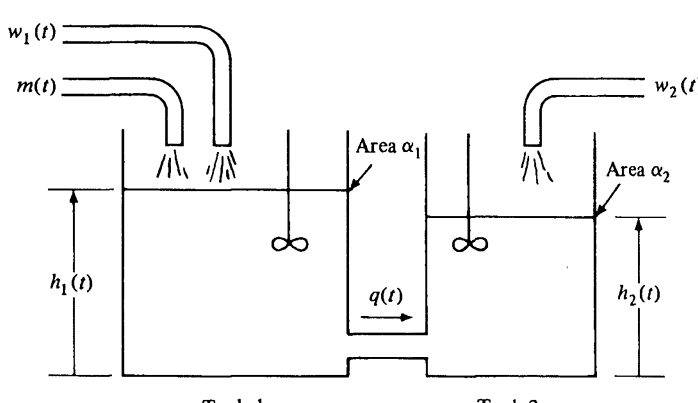


Figure 1-P6

tionality constant k ft^3/ft-min, and that the mixtures in the tanks are homogeneous. Write the differential equations of the system, using $h_1(t)$, $h_2(t)$, $v_1(t)$, and $v_2(t)$ as state variables.

1-7. Write a set of state equations for the electromechanical system shown in Fig. 1-P7. The amplifier gain is K_a, and the developed torque is $\lambda(t) = K_t i_f(t)$, where K_a and K_t are known constants.

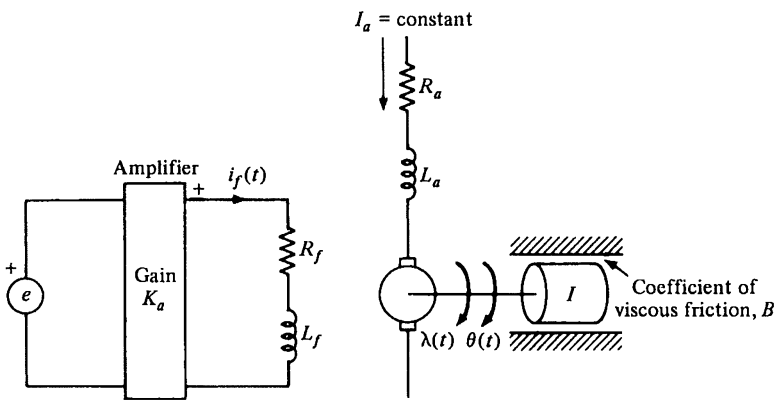


Figure 1-P7

1-8. Write a set of state equations for the mechanical system shown in Fig. 1-P8. The displacements $y_1(t)$ and $y_2(t)$ are measured from the equilibrium position of the system with no force applied.

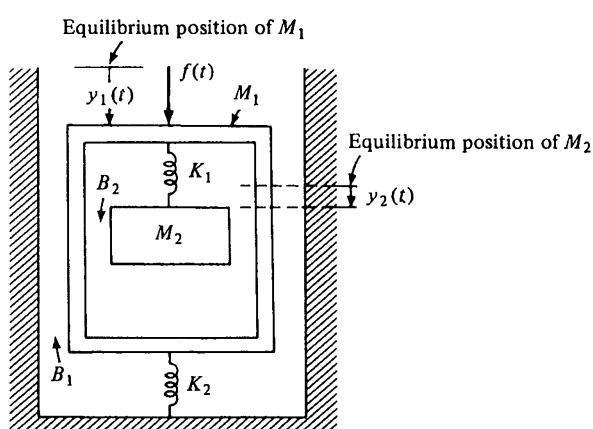


Figure 1-P8

1-9. Write a set of differential equations, in state form, for the coupled *RLC* network shown in Fig. 1-P9.

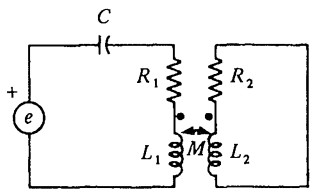


Figure 1-P9

1-10. Write a set of state equations for the network shown in Fig. 1-P10. $R_2(t)$ is a time-varying resistor, and the circuit also contains a nonlinear resistor.

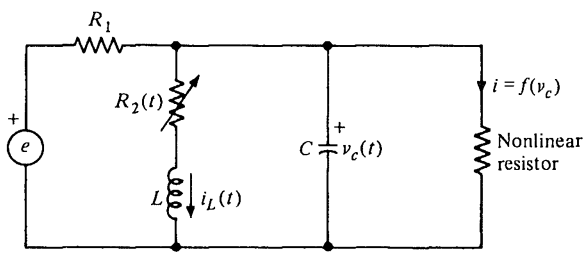


Figure 1-P10

1-11. Show that the state transition matrix satisfies the properties given in Table 1-1.
Hint:

$$\mathbf{x}(t) = \boldsymbol{\varphi}(t, t_i)\mathbf{x}(t_i)$$

is a solution of

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t).$$

1-12. Draw a block diagram, or signal flow graph, and write state and output equations that correspond to the transfer functions:

(a) $\dfrac{Y(s)}{U(s)} = \dfrac{5}{s[s+1]}$

(b) $\dfrac{Y(s)}{U(s)} = \dfrac{1}{s^2}$

(c) $\dfrac{Y(s)}{U(s)} = \dfrac{10}{s^3 + 5s^2 + 6s + 3}$

(d) $\dfrac{Y(s)}{U(s)} = \dfrac{8}{2s^4 + 6s^3 + 14s^2 + 7s + 1}$

(e) $\dfrac{Y(s)}{U(s)} = \dfrac{5[s+2]}{s[s+1]}$

(f) $\dfrac{Y(s)}{U(s)} = \dfrac{[s+1][s+2]}{s^2}$

(g) $\dfrac{Y(s)}{U(s)} = \dfrac{10[s^2 + 2s + 3]}{s^3 + 5s^2 + 6s + 3}$

(h) $\dfrac{Y(s)}{U(s)} = \dfrac{4}{[s+1][s+2]}$

(i) $\dfrac{Y(s)}{U(s)} = \dfrac{[s^2 + 7s + 12]}{s[s+1][s+2]}$

(j) $\dfrac{Y(s)}{U(s)} = \dfrac{8[s^3 + s + 2]}{2s^4 + 6s^3 + 14s^2 + 7s + 1}$

1-13. Find the state transition matrices $\varphi(t)$ for the systems (a), (b), (e), (f), (h), and (i) in Problem 1-12.

1-14. For each of the following systems determine:

(i) If the system is controllable.

(ii) If the system is observable.

(iii) The block diagram or signal flow graph of the system.

(a) $\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t); \qquad y(t) = x_1(t).$

(b) $\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t); \qquad y(t) = x_2(t).$

(c) The coupled circuit in Problem 1-9 with $M = 0$, $\mathbf{y}(t) = \begin{bmatrix} v_c(t) \\ i_{L_2}(t) \end{bmatrix}$.

(d) The coupled circuit in Problem 1-9 with $M = 0.5\text{H}$, $L_1 = 1.0\text{H}$, $L_2 = 0.5\text{H}$, $R_1 = 2.0\,\Omega$, $R_2 = 1.0\,\Omega$, $C = 0.5\text{F}$, and $y(t) = v_c(t)$.

(e) $\dot{\mathbf{x}}(t) = \begin{bmatrix} -2 & 0 & 1 \\ 0 & -1 & 0 \\ -3 & -4 & -2 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}\mathbf{u}(t); \qquad y(t) = x_1(t).$

(f) $\dot{\mathbf{x}}(t) = \begin{bmatrix} -2 & 0 & 1 \\ 0 & -1 & 1 \\ -3 & 0 & -2 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}\mathbf{u}(t); \qquad y(t) = x_1(t).$

(g) $\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}u(t);$

$y(t) = x_1(t); \; a_i \neq 0, i = 0, 1, 2, 3.$

1-15. What are the requirements for the system

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}u(t);$$

$$y(t) = [c_1 \quad c_2 \quad c_3 \quad c_4]\mathbf{x}(t)$$

to be:

(i) Controllable?

(ii) Observable?

Assume that λ_i, $i = 1, \ldots, 4$ are real and distinct.

2

The Performance Measure

Having already considered the modeling of systems and the determination of state and control constraints, we are now ready to discuss performance measures used in control problems. Our objective is to provide physical motivation for the selection of a performance measure.

Classical design techniques have been successfully applied to *linear, time-invariant, single-input single-output systems with zero initial conditions.* Typical performance criteria are system response to a step or ramp input—characterized by rise time, settling time, peak overshoot, and steady-state accuracy—and the frequency response of the system—characterized by gain and phase margin, peak amplitude, and bandwidth. Classical techniques have proved to be successful in many applications; however, we wish to consider systems of a more general nature with performance objectives not readily described in classical terms.

2.1 PERFORMANCE MEASURES FOR OPTIMAL CONTROL PROBLEMS

The "optimal control problem" is to find a control $\mathbf{u}^* \in U$ which causes the system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{2.1-1}$$

to follow a trajectory $\mathbf{x}^* \in X$ that minimizes the performance measure

$$J = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt. \tag{2.1-2}$$

Let us now discuss some typical control problems to provide some physical motivation for the selection of a performance measure.

Minimum-Time Problems

Problem: To transfer a system from an arbitrary initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ to a specified target set S in minimum time.

The performance measure to be minimized is

$$\begin{aligned} J &= t_f - t_0 \\ &= \int_{t_0}^{t_f} dt, \end{aligned} \tag{2.1-3}$$

with t_f the first instant of time when $\mathbf{x}(t)$ and S intersect. The automobile example discussed in Section 1.1 is a minimum-time problem. Other typical examples are the interception of attacking aircraft and missiles, and the slewing mode operation of a radar, or gun system.

Terminal Control Problems

Problem: To minimize the deviation of the final state of a system from its desired value $\mathbf{r}(t_f)$.

A possible performance measure is

$$J = \sum_{i=1}^{n} [x_i(t_f) - r_i(t_f)]^2. \tag{2.1-4}$$

Since positive and negative deviations are equally undesirable, the error is squared. Absolute values could also be used, but the quadratic form in Eq. (2.1-4) is easier to handle mathematically. Using matrix notation, we have

$$J = [\mathbf{x}(t_f) - \mathbf{r}(t_f)]^T [\mathbf{x}(t_f) - \mathbf{r}(t_f)], \tag{†2.1-5}$$

or this can be written as

$$J = \| \mathbf{x}(t_f) - \mathbf{r}(t_f) \|^2. \tag{2.1-5a}$$

$\| \mathbf{x}(t_f) - \mathbf{r}(t_f) \|$ is called the *norm* of the vector $[\mathbf{x}(t_f) - \mathbf{r}(t_f)]$.

† T denotes the matrix transpose.

To allow greater generality, we can insert a real symmetric positive semi-definite $n \times n$ weighting matrix $\mathbf{H}$† to obtain

$$J = [\mathbf{x}(t_f) - \mathbf{r}(t_f)]^T \mathbf{H} [\mathbf{x}(t_f) - \mathbf{r}(t_f)]. \tag{2.1-6}$$

This quadratic form is also written

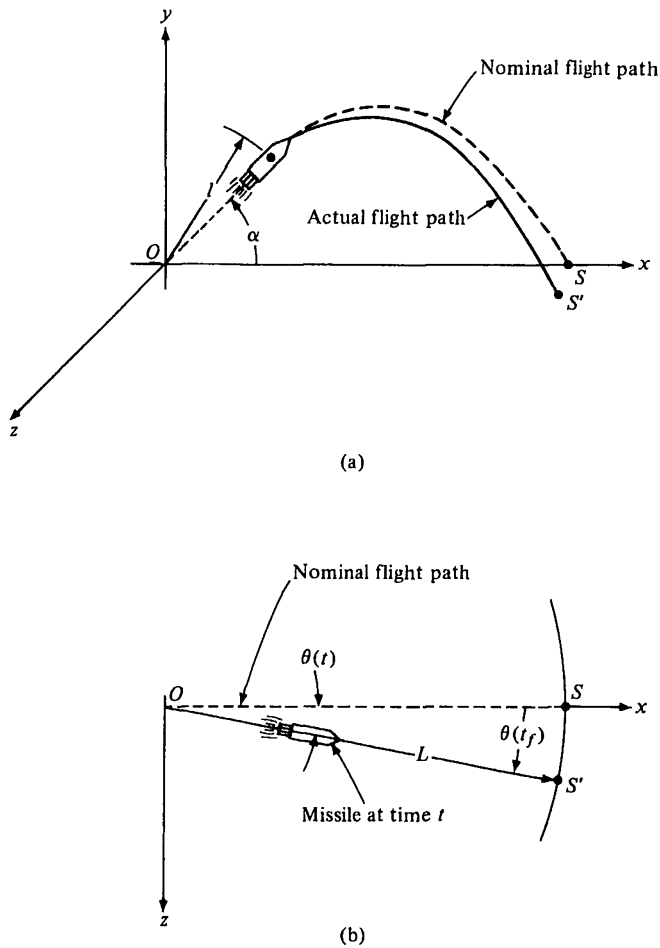


Figure 2-1 A ballistic missile aimed toward the target S

† A real symmetric matrix $\mathbf{H}$ is *positive semi-definite* (*or nonnegative definite*) if for all vectors $\mathbf{z}$, $\mathbf{z}^T\mathbf{H}\mathbf{z} \geq 0$. In other words, there are some vectors for which $\mathbf{H}\mathbf{z} = \mathbf{0}$ in which case $\mathbf{z}^T\mathbf{H}\mathbf{z} = 0$, and for all other $\mathbf{z}$, $\mathbf{z}^T\mathbf{H}\mathbf{z} > 0$.

$$J = \| \mathbf{x}(t_f) - \mathbf{r}(t_f) \|_{\mathbf{H}}^2. \tag{2.1-6a}$$

If $\mathbf{H}$ is the identity matrix,† (2.1-6) and (2.1-5) are identical.

Suppose that $\mathbf{H}$ is a diagonal matrix. The assumption that $\mathbf{H}$ is positive semi-definite implies that all of the diagonal elements are nonnegative. By adjusting the element values we can weight the relative importance of the deviation of each of the states from their desired values. Thus, by increasing h_{ii}‡ we attach more significance to deviation of $x_i(t_f)$ from its desired value; by making h_{jj} zero we indicate that the final value of x_j is of no concern whatsoever.

The elements of $\mathbf{H}$ should also be adjusted to normalize the numerical values encountered. For example, consider the ballistic missile shown in Fig. 2-1. The position of the missile at time t is specified by the spherical coordinates $l(t)$, $\alpha(t)$, and $\theta(t)$. l is the distance from the origin of the coordinate system, and α and θ are the elevation and azimuth angles. If $L = 5000$ miles and $l(t_f) = L$, an azimuth error at impact of 0.01 rad results in missing the target S by 50 miles! If the performance measure is

$$J = h_{11}[l(t_f) - 5000]^2 + h_{22}[\theta(t_f)]^2, \tag{2.1-7}$$

then we would select $h_{22} = [50/0.01]^2 \cdot h_{11}$ to weight equally deviations in range and azimuth. Alternatively, the variables θ and l could be normalized, in which case $h_{11} = h_{22}$.

Minimum-Control-Effort Problems

Problem: To transfer a system from an arbitrary initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ to a specified target set S, with a minimum expenditure of control effort.

The meaning of the term "minimum control effort" depends upon the particular physical application; therefore, the performance measure may assume various forms. For example, consider a spacecraft on an interplanetary exploration—let $u(t)$ be the thrust of the rocket engine, and assume that the magnitude of thrust is proportional to the rate of fuel consumption. In order to minimize the total expenditure of fuel, the performance measure

† The identity matrix is

$$\mathbf{I} \triangleq \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

‡ h_{ii} denotes the iith element of $\mathbf{H}$.

$$J = \int_{t_0}^{t_f} |u(t)|\, dt \tag{2.1-8}$$

would be selected. If there are several controls, and the rate of expenditure of control effort of the ith control is $c_i|u_i(t)|$, $i = 1, \ldots, m$ (c_i is a constant of proportionality), then minimizing

$$J = \int_{t_0}^{t_f} \left[\sum_{i=1}^{m} \beta_i |u_i(t)|\right] dt \tag{2.1-8a}$$

would minimize the control effort expended. The β_i's are nonnegative weighting factors.

As another example, consider a voltage source driving a network containing no energy storage elements. Let $u(t)$ be the source voltage, and suppose that the network is to be controlled with minimum source energy dissipation. The source current is directly proportional to $u(t)$ in this case, so to minimize the energy dissipated, minimize the performance measure

$$J = \int_{t_0}^{t_f} u^2(t)\, dt. \tag{2.1-9}$$

For several control inputs the general form of performance measure corresponding to (2.1-9) is

$$\begin{aligned} J &= \int_{t_0}^{t_f} [\mathbf{u}^T(t)\mathbf{R}\mathbf{u}(t)]\, dt \\ &= \int_{t_0}^{t_f} \|\mathbf{u}(t)\|_{\mathbf{R}}^2\, dt, \end{aligned} \tag{2.1-9a}$$

where $\mathbf{R}$ is a real symmetric positive definite† weighting matrix. The elements of $\mathbf{R}$ may be functions of time if it is desired to vary the weighting on control-effort expenditure during the interval $[t_0, t_f]$.

Tracking Problems

Problem: To maintain the system state $\mathbf{x}(t)$ as close as possible to the desired state $\mathbf{r}(t)$ in the interval $[t_0, t_f]$.

As a performance measure we select

$$J = \int_{t_0}^{t_f} \|\mathbf{x}(t) - \mathbf{r}(t)\|_{\mathbf{Q}(t)}^2\, dt, \tag{2.1-10}$$

† A real symmetric matrix $\mathbf{R}$ is positive definite if

$$\mathbf{z}^T\mathbf{R}\mathbf{z} > 0$$

for all $\mathbf{z} \neq \mathbf{0}$.

where $\mathbf{Q}(t)$ is a real symmetric $n \times n$ matrix that is positive semi-definite for all $t \in [t_0, t_f]$. The elements of the matrix $\mathbf{Q}$ are selected to weight the relative importance of the different components of the state vector and to normalize the numerical values of the deviations. For example, if $\mathbf{Q}$ is a constant diagonal matrix and q_{ii} is zero, this indicates that deviations of x_i are of no concern.

If the set of admissible controls is bounded, e.g., $|u_i(t)| \leq 1$, $i = 1, 2, \ldots, m$, then (2.1-10) is a reasonable performance measure; however, if the controls are not bounded, minimizing (2.1-10) results in controls with impulses and their derivatives. To avoid placing bounds on the admissible controls, or if control energy is to be conserved, we use the modified performance measure

$$J = \int_{t_0}^{t_f} [\,\|\mathbf{x}(t) - \mathbf{r}(t)\|^2_{\mathbf{Q}(t)} + \|\mathbf{u}(t)\|^2_{\mathbf{R}(t)}\,]\,dt. \tag{2.1-11}$$

$\mathbf{R}(t)$ is a real symmetric *positive definite* $m \times m$ matrix for all $t \in [t_0, t_f]$. We shall see in Section 5.2 that if the plant is linear this performance measure leads to an easily implemented optimal controller.

It may be especially important that the states be close to their desired values at the final time. In this case, the performance measure

$$J = \|\mathbf{x}(t_f) - \mathbf{r}(t_f)\|^2_{\mathbf{H}} + \int_{t_0}^{t_f} [\,\|\mathbf{x}(t) - \mathbf{r}(t)\|^2_{\mathbf{Q}(t)} + \|\mathbf{u}(t)\|^2_{\mathbf{R}(t)}\,]\,dt \tag{2.1-12}$$

could be used. $\mathbf{H}$ is a real symmetric positive semi-definite $n \times n$ matrix.

Regulator Problems

A regulator problem is the special case of a tracking problem which results when the desired state values are zero ($\mathbf{r}(t) = \mathbf{0}$ for all $t \in [t_0, t_f]$).

2.2 SELECTING A PERFORMANCE MEASURE

In selecting a performance measure the designer attempts to define a mathematical expression which when minimized indicates that the system is performing in the most desirable manner. Thus, choosing a performance measure is a translation of the system's physical requirements into mathematical terms. In particular, suppose that two admissible control histories which cause admissible state trajectories are specified and we are to select the better one. To evaluate these controls, perform the test shown in Fig.

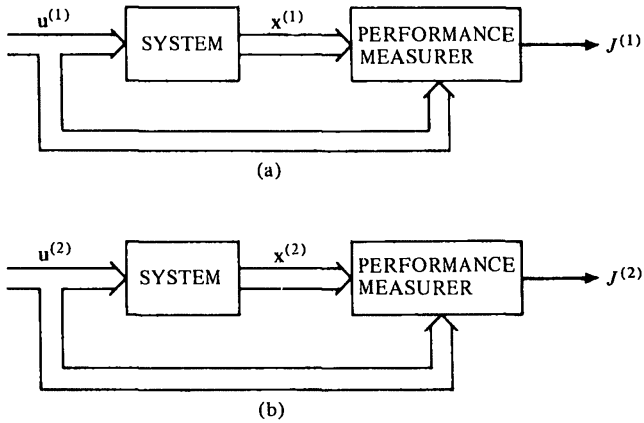


Figure 2-2 Evaluating two specified control histories

2-2. First, apply the control $\mathbf{u}^{(1)}$ to the system and determine the value of the performance measure $J^{(1)}$; then repeat this procedure with $\mathbf{u}^{(2)}$ applied to obtain $J^{(2)}$. If $J^{(1)} < J^{(2)}$, then we designate $\mathbf{u}^{(1)}$ as the better control; if $J^{(2)} < J^{(1)}$, $\mathbf{u}^{(2)}$ is better; if $J^{(1)} = J^{(2)}$ the two controls are equally desirable. An alternative test is to apply each control, record the state trajectories, and then subjectively decide which trajectory is better.

If the performance measure truly reflects desired system performance, the trajectory selected by the designer as being "more to his liking" should yield the smaller value of J. If this is not the case, the performance measure or the constraints should be modified.

Example 2.2-1. Figure 2-3 shows a manned spacecraft whose attitude is to be controlled by a gas expulsion system. As a simplification, we shall

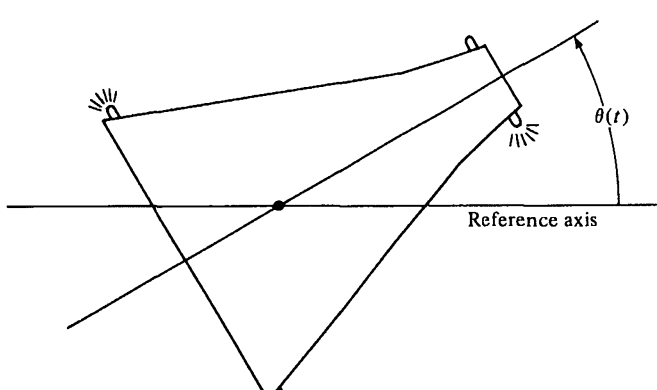


Figure 2-3 Attitude control of a spacecraft

consider only the control of the pitch angle $\theta(t)$. The differential equation that describes the motion is

$$I\frac{d^2}{dt^2}[\theta(t)] = \lambda(t), \tag{2.2-1}$$

where I is the angular moment of inertia and $\lambda(t)$ is the torque produced by the gas jets. Selecting $x_1(t) \triangleq \theta(t)$ and $x_2(t) \triangleq \dot{\theta}(t)$ as state variables, and $u(t) \triangleq \lambda(t)/I$ as the control gives the state equations

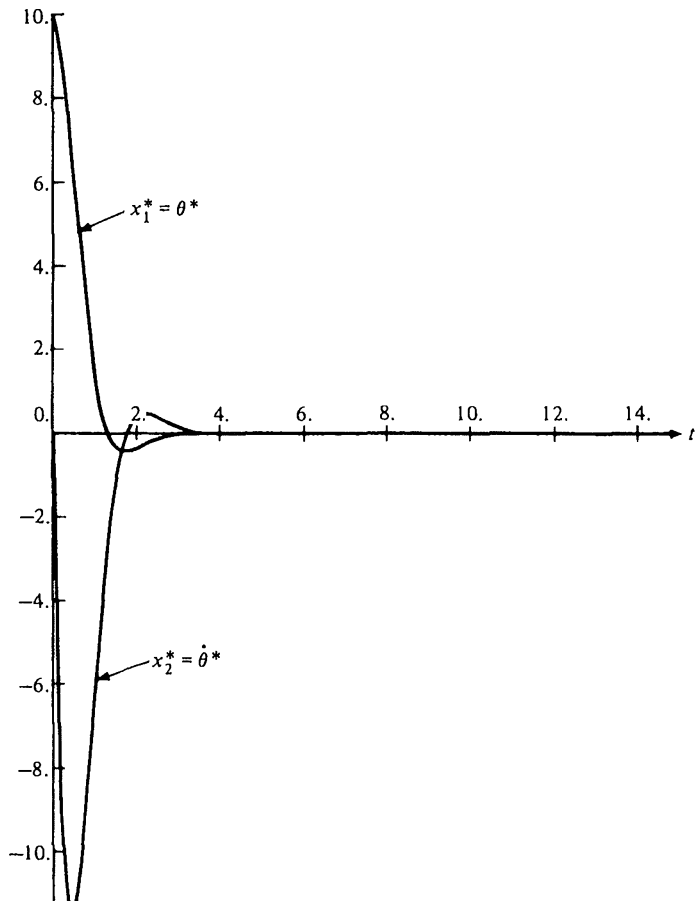


Figure 2-4(a) Position and velocity as functions of time

$$\mathbf{Q} = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}; R = .1; \mathbf{x}(0) = \begin{bmatrix} 10 \\ 0 \end{bmatrix}$$

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u(t). \end{aligned} \tag{2.2-2}$$

The primary objective of the control system is to maintain the angular position near zero. This is to be accomplished with small acceleration.

As a performance measure we select

$$J = \int_0^\infty [q_{11}x_1^2(t) + q_{22}x_2^2(t) + Ru^2(t)]\, dt, \tag{2.2-3}$$

where $q_{11}, q_{22} \geq 0$, and $R > 0$ are weighting factors. In Figs. 2-4, 2-5, 2-6, and 2-7 the optimal trajectories for $q_{11} = 4.0$, $q_{22} = 0$, and several

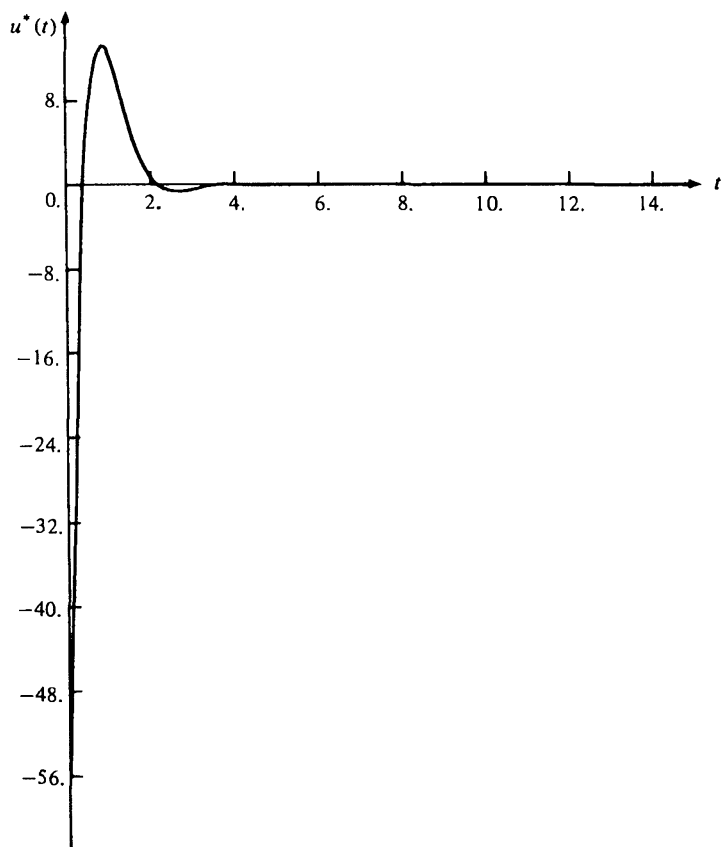


Figure 2-4(b) Acceleration as a function of time

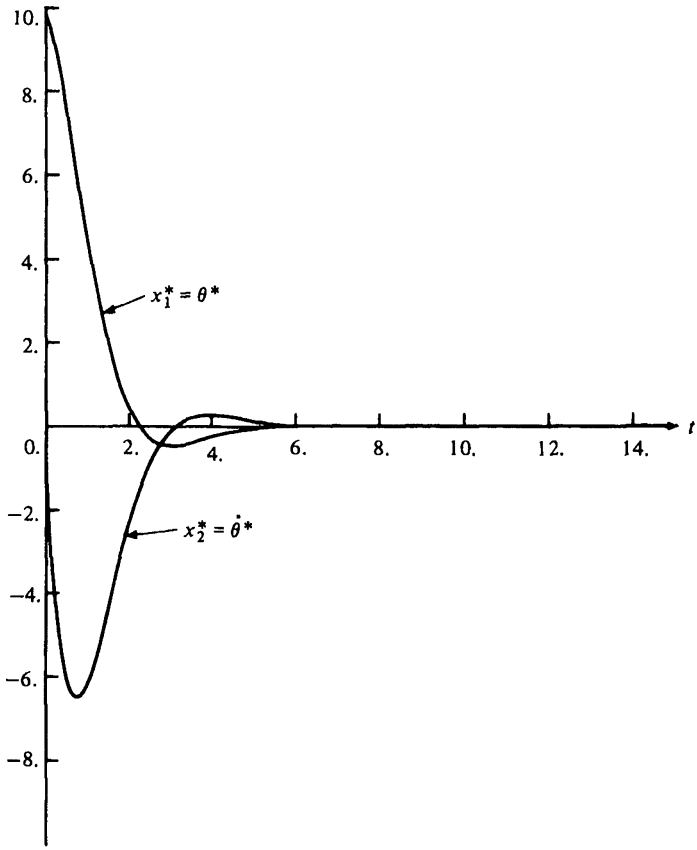


Figure 2-5(a) Position and velocity as functions of time

$$\mathbf{Q} = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}; R = 1.; \mathbf{x}(0) = \begin{bmatrix} 10 \\ 0 \end{bmatrix}$$

values of R are shown.† Increasing R places a heavier penalty on acceleration and control energy expenditure. All of these trajectories are optimal, each for a different performance measure. If we are most concerned about reducing the angular displacement to zero quickly, then the trajectory in Fig. 2-4 would be our choice. The astronauts, however, would probably prefer the trajectory shown in Fig. 2-7 because of the much smaller accelerations.

We must be very careful when interpreting the numerical value of the

† These trajectories were obtained by using the techniques discussed in Sections 3.12 and 5.2.

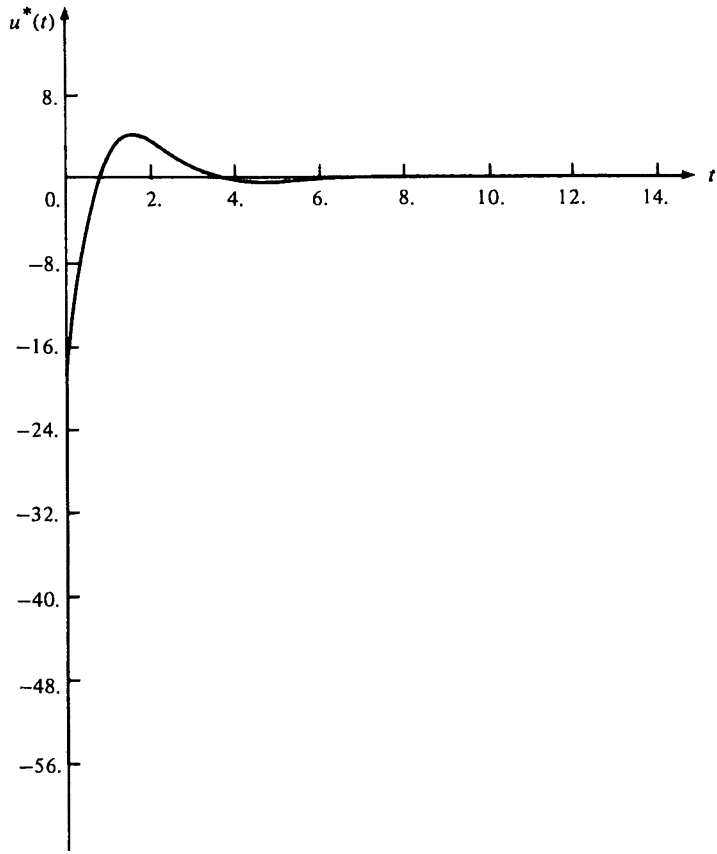


Figure 2-5(b) Acceleration as a function of time

minimum performance measure. By multiplying every weighting factor in the performance measure by a positive constant k, the value of the measure would be k times its original value, but the optimal control and trajectory would remain exactly the same. In fact, it may be possible to adjust the weighting factors by different amounts and still retain the same optimal control and trajectory.†

The physical interpretation of the value of the performance measure is also a factor to be considered. The minimum value of a performance measure such as elapsed time or consumed fuel has a definite physical significance; however, for problems in which the performance measure is a weighted

† See Chapter 8 of reference [S-2].

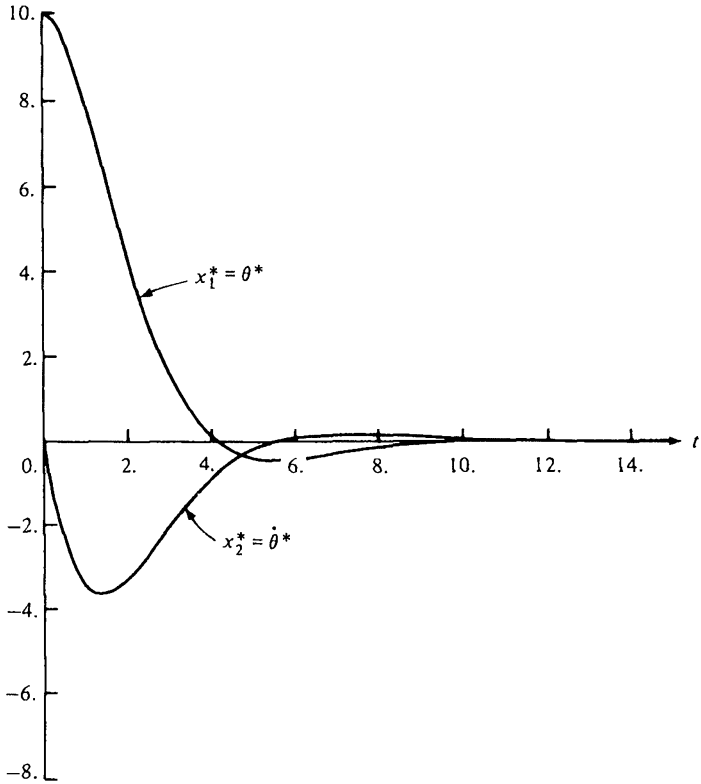


Figure 2-6(a) Position and velocity as functions of time

$$\mathbf{Q} = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}; R = 10; \mathbf{x}(0) = \begin{bmatrix} 10 \\ 0 \end{bmatrix}$$

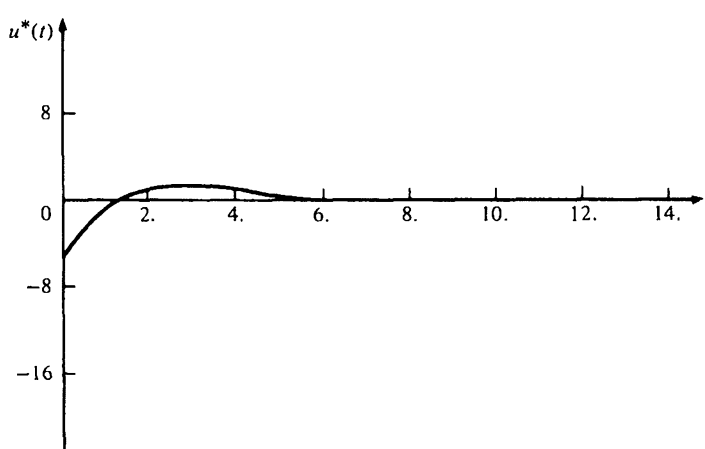


Figure 2-6(b) Acceleration as a function of time

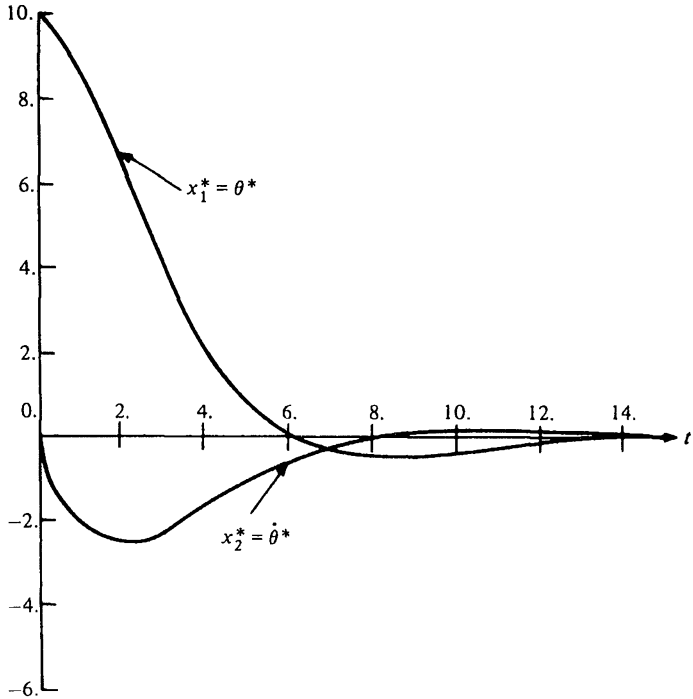


Figure 2-7(a) Position and velocity as functions of time

$$\mathbf{Q} = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}; R = 50; \mathbf{x}(0) = \begin{bmatrix} 10 \\ 0 \end{bmatrix}$$

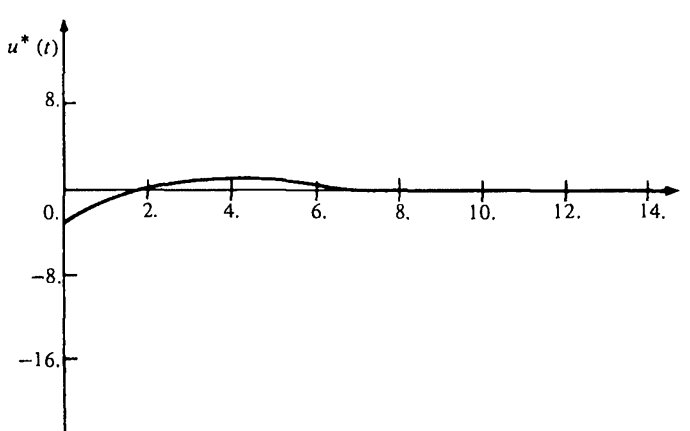


Figure 2-7(b) Acceleration as a function of time

combination of different physical quantities—as in the preceding spacecraft example—the numerical value of the performance measure does not represent a physically meaningful quantity.

2.3 SELECTION OF A PERFORMANCE MEASURE: THE CARRIER LANDING OF A JET AIRCRAFT

The following example, which is similar to a problem considered by Merriam and Ellert [M-1], illustrates the selection of a performance measure. The problem is to design an automatic control system for landing a high-speed jet airplane on the deck of an aircraft carrier.

The jet aircraft is shown in Fig. 2-8. The x direction is along the velocity vector of the aircraft, and the y and z directions are as shown. α is the angle of attack, θ is the pitch angle, and γ is the glide path angle.

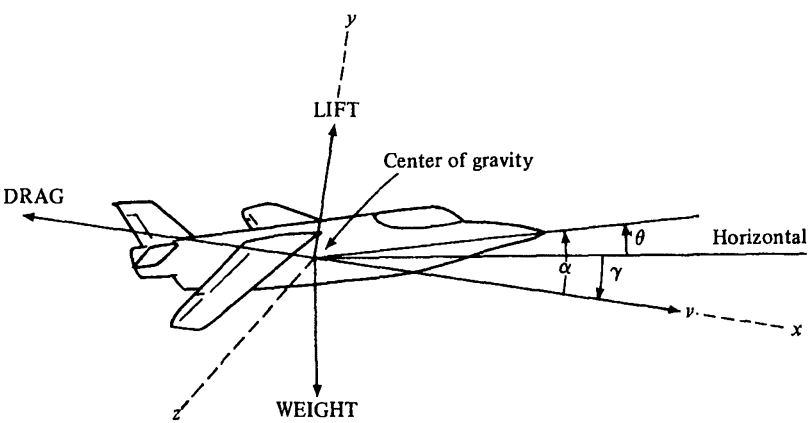


Figure 2-8 Aircraft coordinates and angles

We shall make the following simplifying assumptions:

1. Lateral motion is ignored; only motion in the x-y plane is considered.
2. Random disturbances, such as wind gusts and carrier deck motion, are neglected.
3. The nominal glide path angle γ is small, so that $\cos\gamma \approx 1$ and $\sin\gamma \approx \gamma$ in radians (it will be shown that the nominal γ is -0.0636 rad).
4. The velocity of the aircraft with respect to the nominal landing point is maintained at a constant value of 160 mph (235 ft/sec) by an automatic throttle control device.

5. The longitudinal motion of the aircraft is controlled entirely by the elevator deflection angle [$\delta_e(t)$, shown in Fig. 2-9], which has been trimmed to a nominal setting of 0° at the start of the automatic landing phase.

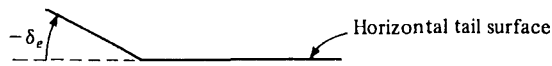


Figure 2-9 Elevator deflection angle

6. The aircraft dynamics are described by a set of differential equations that have been linearized about the equilibrium flight condition.

Since we desire to have readily available information concerning the system states to generate the required control, the altitude above the flight deck h, altitude rate $\dot{h}$, pitch angle θ, and pitch rate $\dot{\theta}$ are selected as the state variables. h is measured by a radar altimeter, $\dot{h}$ by a barometric rate meter, θ and $\dot{\theta}$ by gyros. If we define $x_1 \triangleq h$, $x_2 \triangleq \dot{h}$, $x_3 \triangleq \theta$, $x_4 \triangleq \dot{\theta}$, and $u \triangleq \delta_e$, the state equations that result from the linearization of the aircraft motion about the equilibrium flight condition are†

$$\begin{aligned}
\dot{x}_1(t) &= x_2(t) \\
\dot{x}_2(t) &= a_{22}x_2(t) + a_{23}x_3(t) \\
\dot{x}_3(t) &= x_4(t) \\
\dot{x}_4(t) &= a_{42}x_2(t) + a_{43}x_3(t) + a_{44}x_4(t) + b_4u(t),
\end{aligned} \tag{2.3-1}$$

where the a's and b_4 are known constants. In matrix form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t); \tag{2.3-2}$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & a_{42} & a_{43} & a_{44} \end{bmatrix}; \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ b_4 \end{bmatrix}.$$

Next, the desired behavior for the aircraft must be defined. The nominal flight path is selected as a straight line which begins at an altitude of 450 ft and at a range of 7,050 ft measured from the landing point on the carrier deck. This results in 30 seconds' being the nominal time required for the terminal phase of the landing. The desired altitude trajectory h_d is shown in Fig. 2-10. This selection for h_d implies that the desired altitude rate

† See [M-1].

$\dot{h}_d$ is as shown in Fig. 2-11. It is desired to maintain the attitude of the aircraft at 5°. This is most important at touchdown because it is required that the main landing gear touch before the nose or tail gear. Since $\theta_d(t) = 5°$ for $t \in [0, 30]$, $\dot{\theta}_d(t) = 0$ during this time interval, and the desired attitude and attitude rate profiles are shown in Figs. 2-12 and 2-13. The desired atti-

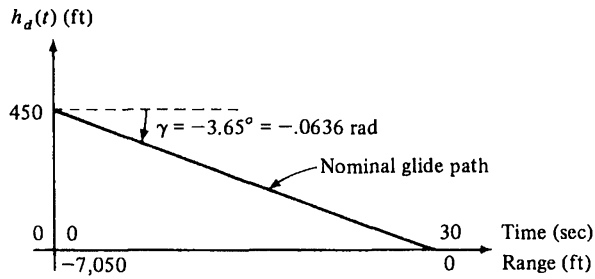


Figure 2-10 Desired altitude history

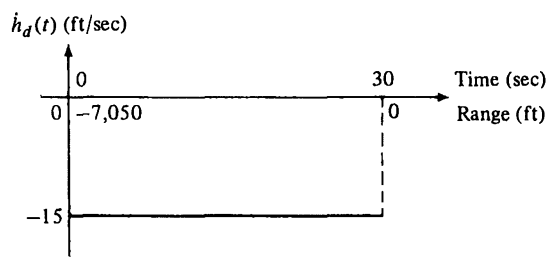


Figure 2-11 Desired rate of ascent history

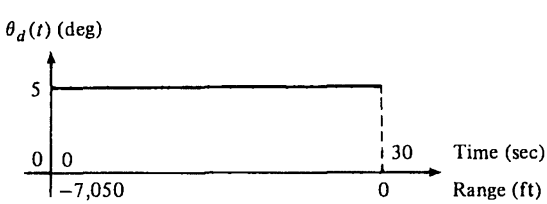


Figure 2-12 Desired attitude profile

Figure 2-13 Desired attitude rate profile

tude and glide path angle profiles imply that the desired angle of attack α_d is 8.65° during the entire 30 sec interval.

It is assumed that large deviations of δ_e from the nominal 0° setting are indicative of a suboptimal landing and should be avoided; therefore, the desired value of δ_e is 0° throughout the terminal phase of landing.

The assumption is also made that there are limits on allowable departure from nominal values during descent. If any of these limits are exceeded, a wave-off is given, and the pilot takes control.

The translation of the performance requirements into a quantitative measure is the next task. The performance measure is selected as the integral of a sum of quadratic terms in the state and control variables and some additional terms to account for quantities which are crucial at touchdown. The index selected is

$$\begin{aligned} J = {} & k_h[h(30) - h_d(30)]^2 + k_{\dot{h}}[\dot{h}(30) - \dot{h}_d(30)]^2 + k_\theta[\theta(30) - \theta_d(30)]^2 \\ & + \int_0^{30} \{q_h(\tau)[h(\tau) - h_d(\tau)]^2 + q_{\dot{h}}(\tau)[\dot{h}(\tau) - \dot{h}_d(\tau)]^2 \\ & + q_\theta(\tau)[\theta(\tau) - \theta_d(\tau)]^2 + q_{\dot{\theta}}(\tau)[\dot{\theta}(\tau) - \dot{\theta}_d(\tau)]^2 \\ & + r_{\delta_e}(\tau)[\delta_e(\tau) - \delta_{e_d}(\tau)]^2\}\, d\tau, \end{aligned} \tag{2.3-3}$$

where τ is a dummy variable of integration. The k's, q's, and r_{δ_e} are weighting factors that are specified to assign relative importance to each of the terms in the performance measure and to account for differences in the relative size of numerical values encountered. The q's and r_{δ_e} are written as time-varying functions because deviations of some of the variables from nominal values may be more critical during certain periods of time than others. For example, rate of ascent errors are more critical over the flight deck than at earlier instants, so $q_h(t)$ should increase as t approaches 30 sec. The terms outside of the integral are there to help ensure that the attitude, rate of ascent, and altitude are close to nominal at $t = 30$ sec. Notice that the term containing $h(30)$ penalizes a landing that occurs too soon or too late.

There is no term in the measure containing the angle of attack α explicitly; however, if the values for θ and $\dot{\theta}$ are maintained "close" to their desired values, then it is reasonable to expect that α will be satisfactory. Certainly a term could be added containing the deviation of angle of attack from its nominal value, but this would necessitate the selection of an additional weighting factor, and it is generally desirable to keep the problem simple for the initial solution. The desired, or nominal, aircraft trajectory is specified by Figs. 2-10 through 2-13. Figure 2-10 gives $h_d(t) = 450 - 15t$ ft as the desired altitude history, and the desired altitude at $t = 30$ (the nominal time touchdown occurs) is $h_d(30) = 0$ ft. From Fig. 2-11 the desired altitude rate history is -15 ft/sec throughout the interval [0, 30]; thus, $\dot{h}_d(t) = -15$ ft/sec and $\dot{h}_d(30) = -15$ ft/sec. The desired aircraft attitude is $+5°$ in the

entire landing interval; therefore, $\theta_d(t) = 0.0873$ rad, and $\theta_d(30) = 0.0873$ rad. From Fig. 2-13 we have $\dot{\theta}_d(t) = 0$ rad/sec as the nominal attitude rate, and $\dot{\theta}_d(30) = 0$ rad/sec. Substituting the desired values in (2.3-3) gives

$$\begin{aligned} J = {} & k_h[h(30)]^2 + k_{\dot{h}}[\dot{h}(30) + 15]^2 + k_\theta[\theta(30) - 0.0873]^2 \\ & + \int_0^{30} \{q_h(\tau)[h(\tau) - 450 + 15\tau]^2 + q_{\dot{h}}(\tau)[\dot{h}(\tau) + 15]^2 \\ & + q_\theta(\tau)[\theta(\tau) - 0.0873]^2 + q_{\dot{\theta}}(\tau)[\dot{\theta}(\tau)]^2 \\ & + r_{\delta_e}(\tau)[\delta_e(\tau)]^2\}\, d\tau, \end{aligned} \tag{2.3-4}$$

where θ is in radians, $\dot{\theta}$ in radians per second, h in feet, and $\dot{h}$ in feet per second. In matrix form

$$\begin{aligned} J = {} & [\mathbf{x}(30) - \mathbf{r}(30)]^T \mathbf{H}[\mathbf{x}(30) - \mathbf{r}(30)] \\ & + \int_0^{30} \{[\mathbf{x}(\tau) - \mathbf{r}(\tau)]^T \mathbf{Q}(\tau)[\mathbf{x}(\tau) - \mathbf{r}(\tau)] + r_{\delta_e}(\tau)u^2(\tau)\}\, d\tau, \end{aligned} \tag{2.3-5}$$

where $\mathbf{x}(t)$ is the state at time t, $\mathbf{r}(t)$ is the desired or nominal value of the state at time t, $u(t)$ is the control, r_{δ_e} is a positive function of time,

$$\mathbf{H} \triangleq \begin{bmatrix} k_h & 0 & 0 & 0 \\ 0 & k_{\dot{h}} & 0 & 0 \\ 0 & 0 & k_\theta & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and

$$\mathbf{Q}(t) \triangleq \begin{bmatrix} q_h(t) & 0 & 0 & 0 \\ 0 & q_{\dot{h}}(t) & 0 & 0 \\ 0 & 0 & q_\theta(t) & 0 \\ 0 & 0 & 0 & q_{\dot{\theta}}(t) \end{bmatrix}.$$

The designer must select functional relationships for q_h, $q_{\dot{h}}$, q_θ, $q_{\dot{\theta}}$, and r_{δ_e}, and numerical values for k_h, $k_{\dot{h}}$, and k_θ. In this example the deviation from the desired trajectory is to be minimized; therefore, the q's and k's would assume only nonnegative values, and r_{δ_e} would be positive for all $t \in [0, 30]$. This performance measure allows sufficient flexibility to satisfy system requirements, and also leads to an optimal control law that is relatively easy to implement. Reference [M-1] discusses implementation in more detail and also shows trajectories that illustrate the effects of varying weighting parameters in a performance measure.

REFERENCES

M-1 Merriam, C. W., III, and F. J. Ellert, "Synthesis of Feedback Controls Using Optimization Theory—An Example," *IEEE Trans. Automatic Control* (1963), 89–103.

S-2 Schultz, D. G., and J. L. Melsa, *State Functions and Linear Control Systems.* New York: McGraw-Hill, Inc., 1968.

PROBLEMS

2-1. Refer to the chemical mixing process of Problem 1-6. The amount of dye in tank 2, $v_2(t)$, is to be maintained as closely as possible to M ft^3 during a one-day interval.

(a) What would you suggest as a performance measure to be minimized?

(b) Determine a set of physically realistic state and control constraints.

2-2. Repeat Problem 2-1 if the objective is to maximize the amount of dye in tank 2 at the end of one day. It is specified that the total volume of dye that enters tank 1 in the one-day period cannot be more than N ft^3.

2-3. An unmanned roving vehicle has been proposed as part of the Mariner Mars exploration series of space missions. The roving vehicle is designed to navigate on the Martian surface and transmit television pictures and other scientific data to earth. Suppose that the rover is to be driven by a dc motor supplied from rechargeable storage batteries; a simplified model is shown in Fig. 2-P3.

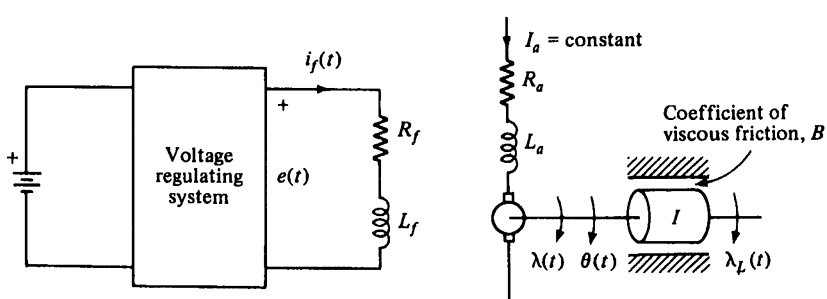


Figure 2-P3

The output of the voltage regulating system is the control signal $e(t)$. The developed torque is $\lambda(t) = K_t i_f(t)$, where K_t is a known constant; $\lambda_L(t)$ is the load torque caused by hills on the Martian surface. The vehicle's speed is to

deviate as little as possible from 5 mph without requiring excessive energy output from the voltage regulating system (to prolong the life of the batteries). Let $i_f(t)$ and $\dot{\theta}(t)$ be state variables.

(a) Write state equations for the motor-load combination.

(b) Determine a physically reasonable set of state and control constraints.

(c) Suggest a performance measure if:

(i) $L_f = 0$.

(ii) $L_f \neq 0$.

2-4. Refer to the simplified spacecraft model used in Example 2.2-1. Suppose that the objective is to change the spacecraft attitude from an arbitrary initial value to an angle of $+15° \pm 0.1°$ with respect to the reference axis shown in Fig. 2-3. This maneuver is to be accomplished in 30 sec with minimum fuel expenditure.

(a) Determine the state and control constraints.

(b) Suggest an appropriate performance measure.

2-5. Repeat Problem 2-4 if the maneuver is to be accomplished in minimum time.

2-6. Figure 2-P6 shows a rocket that is to be approximated by a particle of instantaneous mass $m(t)$. The instantaneous velocity is $v(t)$, $T(t)$ is the thrust, and

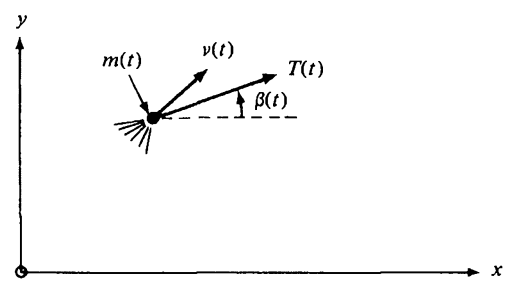


Figure 2-P6

$\beta(t)$ is the thrust angle. If we assume no aerodynamic or gravitational forces, and if we select $x_1 \triangleq x$, $x_2 \triangleq \dot{x}$, $x_3 \triangleq y$, $x_4 \triangleq \dot{y}$, $x_5 \triangleq m$, $u_1 \triangleq T$, $u_2 \triangleq \beta$, the state equations are

$$\dot{x}_1(t) = x_2(t)$$

$$\dot{x}_2(t) = \frac{[u_1(t) \cos u_2(t)]}{x_5(t)}$$

$$\dot{x}_3(t) = x_4(t)$$

$$\dot{x}_4(t) = \frac{[u_1(t) \sin u_2(t)]}{x_5(t)}$$

$$\dot{x}_5(t) = -\frac{1}{c} u_1(t),$$

where c is a constant of proportionality. The rocket starts from rest at the point $x = 0$, $y = 0$.

(a) Determine a set of physically reasonable state and control constraints.

(b) Suggest a performance measure, and any additional constraints imposed, if the objective is to make $y(t_f) = 3$ mi and maximize $x(t_f)$; t_f is specified.

(c) Suggest a performance measure, and any additional constraints imposed, if it is desired to reach the point $x = 500$ mi, $y = 3$ mi in 2.5 min with maximum possible vehicle mass.

II

Dynamic Programming

3

Dynamic Programming

Once the performance measure for a system has been chosen, the next task is to determine a control function that minimizes this criterion. Two methods of accomplishing the minimization are the minimum principle of Pontryagin [P-1], and the method of dynamic programming developed by R. E. Bellman [B-1, B-2, B-3]. The variational approach of Pontryagin (Chapter 5) leads to a nonlinear two-point boundary-value problem that must be solved (Chapter 6) to obtain an optimal control. In this chapter we shall consider the method of dynamic programming and see that it leads to a functional equation that is amenable to solution by use of a digital computer.

3.1 THE OPTIMAL CONTROL LAW

In Chapter 1 we defined an optimal control of the form

$$\mathbf{u}^*(t) = \mathbf{f}(\mathbf{x}(t), t) \tag{3.1-1}$$

as being a *closed-loop* or *feedback* optimal control. The functional relationship $\mathbf{f}$ is called the *optimal control law*, or the *optimal policy*. Notice that the optimal control law specifies how to generate the control value at time t from the state value at time t. The presence of t as an argument of $\mathbf{f}$ indicates that the optimal control law may be time-varying.

In the method of dynamic programming, an optimal policy is found by employing the intuitively appealing concept called the principle of optimality.

3.2 THE PRINCIPLE OF OPTIMALITY†

The *optimal* path for a multistage decision process is shown in Fig. 3-1(a). Suppose that the first decision (made at a) results in segment a-b with cost

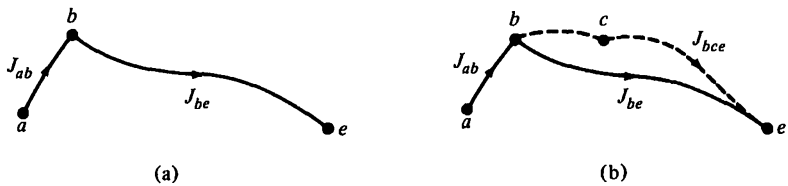


Figure 3-1 (a) Optimal path from a to e. (b) Two possible optimal paths from b to e

J_{ab} and that the remaining decisions yield segment b-e at a cost of J_{be}. The *minimum* cost J_{ae}^* from a to e is therefore

$$J_{ae}^* = J_{ab} + J_{be}. \tag{3.2-1}$$

ASSERTION: If a-b-e is the optimal path from a to e, then b-e is the optimal path from b to e.

Proof by contradiction: Suppose b-c-e in Fig. 3-1(b) is the optimal path from b to e; then

$$J_{bce} < J_{be}, \tag{3.2-2}$$

and

$$J_{ab} + J_{bce} < J_{ab} + J_{be} = J_{ae}^* \tag{3.2-3}$$

but (3.2-3) can be satisfied only by violating the condition that a-b-e is the optimal path from a to e. Thus the assertion is proved.

Bellman [B-1] has called the above property of an optimal policy the principle of optimality:

> An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

† Sections 3.2 through 3.6 follow the presentation given in [K-4].

3.3 APPLICATION OF THE PRINCIPLE OF OPTIMALITY TO DECISION-MAKING

The following example illustrates the procedure for making a single optimal decision with the aid of the principle of optimality.

Consider a process whose current state is b. The paths resulting from all allowable decisions at b are shown in Fig. 3-2(a). The optimal paths from c, d, and e to the terminal point f are shown in Fig. 3-2(b). The principle of

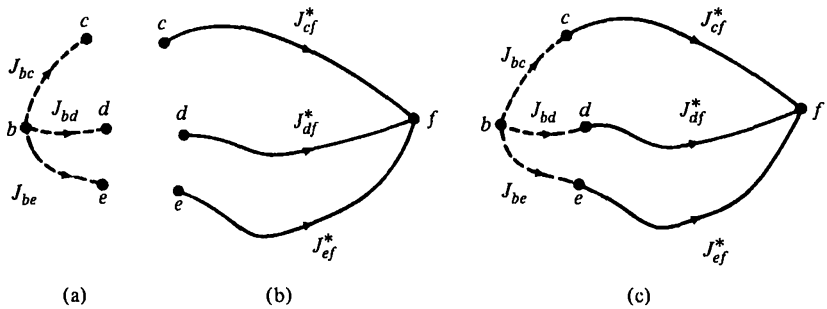


Figure 3-2 (a) Paths resulting from all allowable decisions at b. (b) Optimal paths from c, d, e to f. (c) Candidates for optimal paths from b to f

optimality implies that if b-c is the initial segment of the optimal path from b to f, then c-f is the terminal segment of this optimal path. The same reasoning applied to initial segments b-d and b-e indicates that the paths in Fig. 3-2(c) are the only candidates for the optimal trajectory from b to f. The optimal trajectory that starts at b is found by comparing

$$\begin{aligned} C^*_{bcf} &= J_{bc} + J^*_{cf} \\ C^*_{bdf} &= J_{bd} + J^*_{df} \\ C^*_{bef} &= J_{be} + J^*_{ef}. \end{aligned} \tag{3.3-1}$$

The minimum of these costs must be the one associated with the optimal decision at point b.

Dynamic programming is a computational technique which extends the above decision-making concept to *sequences* of decisions which together define an optimal policy and trajectory. The optimal routing problem in the next section illustrates the procedure.

3.4 DYNAMIC PROGRAMMING APPLIED TO A ROUTING PROBLEM

A motorist wishes to know how to minimize the cost of reaching some destination h from his current location. He can only travel (one-way as indicated) on the streets shown on his map (Fig. 3-3), and at the intersection-to-intersection costs given.

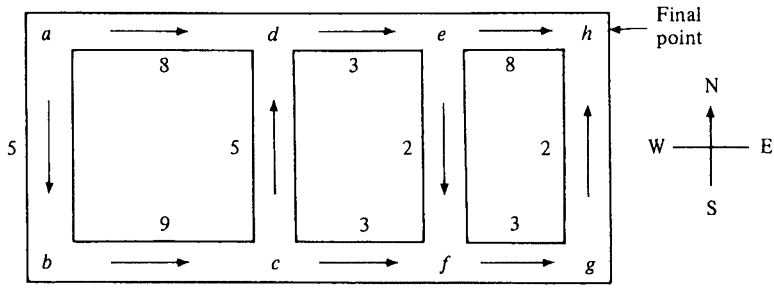


Figure 3-3 The road map

Instead of trying all allowable paths leading from each intersection to h and selecting the one with lowest cost (an exhaustive search), consider the application of the principle of optimality. In this problem, "state" refers to the intersection and a "decision" is the choice of heading (control) elected by the driver when he leaves an intersection.

Suppose the motorist is at c; from there he can go only to d or f, and then on to h. Let J_{cd} denote the cost of moving from c to d and J_{cf} the cost from c to f. Assume that the motorist already knows the minimum costs, J^*_{dh} and J^*_{fh}, to reach the final destination h from d and f. (In this example, $J^*_{dh} = 10$ and $J^*_{fh} = 5$.) Then the minimum cost J^*_{ch} to reach h from c is the smaller of

$$C^*_{cdh} = J_{cd} + J^*_{dh} = \text{minimum cost to reach } h \text{ from } c \text{ via } d \tag{3.4-1}$$

and

$$C^*_{cfh} = J_{cf} + J^*_{fh} = \text{minimum cost to reach } h \text{ from } c \text{ via } f. \tag{3.4-2}$$

Thus,

$$\begin{aligned} J^*_{ch} &= \min\{C^*_{cdh}, C^*_{cfh}\} \\ &= \min\{15, 8\} \\ &= 8 \end{aligned} \tag{3.4-3}$$

and the optimal decision at c is to go to f.

How does the motorist know the values for J^*_{dh} and J^*_{fh}? These quantities must have been calculated previously by working backward from h. For example, $J^*_{gh} = 2$—there is only one path from g to h. J^*_{gh} is then used to find J^*_{fh} from

$$\begin{aligned} J^*_{fh} &= J_{fg} + J^*_{gh} \\ &= 3 + 2 \\ &= 5. \end{aligned} \tag{3.4-4}$$

Then

$$J^*_{eh} = \min \{J_{eh}, [J_{ef} + J^*_{fh}]\} \tag{3.4-5}$$

and so on. The general approach should now be evident. It remains to formalize the computational algorithm. In this connection it will be convenient to introduce the following notation:

- α is the current state (intersection).
- u_i is an allowable decision (control) elected at the state α. In this example i can assume one or more of the values 1, 2, 3, 4, corresponding to the headings N, E, S, W.
- x_i is the state (intersection) adjacent to α which is reached by application of u_i at α.
- h is the final state.
- $J_{\alpha x_i}$ is the cost to move from α to x_i.
- $J^*_{x_i h}$ is the *minimum cost* to reach the final state h from x_i.
- $C^*_{\alpha x_i h}$ is the minimum cost to go from α to h *via* x_i.
- $J^*_{\alpha h}$ is the minimum cost to go from α to h (by any allowable path).
- $u^*(\alpha)$ is the optimal decision (control) at α.

When this notation is used, the principle of optimality implies that

$$C^*_{\alpha x_i h} = J_{\alpha x_i} + J^*_{x_i h} \tag{3.4-6}$$

and, as before, the optimal decision at α, $u^*(\alpha)$, is the decision that leads to

$$J^*_{\alpha h} = \min \{C^*_{\alpha x_1 h}, C^*_{\alpha x_2 h}, \ldots, C^*_{\alpha x_i h}, \ldots\}. \tag{3.4-7}$$

These two equations define the algorithm called dynamic programming. To illustrate the procedure, the automobile routing problem has been "solved" in Table 3-1, where only the consequences of lawful decisions are included. Notice particularly that the intersections nearest the destination h are considered first, and that the optimal trajectories (routes) are built up from h *backwards* toward the earlier states (intersections). This is necessary in order that $J^*_{x_i h}$ be known *prior* to the calculation of $C^*_{\alpha x_i h}$ $(= J_{\alpha x_i} + J^*_{x_i h})$.

Once the table has been completed, the optimal path from *any* intersection to h can be obtained by entering the table at the appropriate intersection and reading off the optimal heading at each successive intersection along the trajectory. For example, if the motorist starts at b, the table tells him to head east. Heading east, he arrives at c, where the table indicates he should again move east. Continuing the process, we find the optimal path from b to h to be *b-c-f-g-h* and the minimum cost to be 17.

The information in the table also allows the motorist to adjust his route if he is forced to deviate from the optimal path. Suppose he started at b and reached c only to find the road to f closed for repairs; he is forced to move to d. After doing so, he enters the table and finds that from d the optimal path to h is *d-e-f-g-h*.

Notice that a motorist at intersection a who heads south instead of east is being misled by the prospect of a short-term gain. His overall cost will be higher, even if he thereafter follows the optimal route.

Table 3-1 CALCULATION OF OPTIMAL HEADINGS BY DYNAMIC PROGRAMMING

Current intersection α	*Heading* u_l	*Next intersection* x_l	*Minimum cost from* α *to* h *via* x_l $J_{\alpha x_l} + J^*_{x_l h} = C^*_{\alpha x_l h}$	*Minimum cost to reach* h *from* α $J^*_{\alpha h}$	*Optimal heading at* α $u^*(\alpha)$
g	N	h	$2 + 0 = 2$	2	N
f	E	g	$3 + 2 = 5$	5	E
e	E	h	$8 + 0 = 8$		
	S	f	$2 + 5 = 7$	7	S
d	E	e	$3 + 7 = 10$	10	E
c	N	d	$5 + 10 = 15$		
	E	f	$3 + 5 = 8$	8	E
b	E	c	$9 + 8 = 17$	17	E
a	E	d	$8 + 10 = 18$	18	E
	S	b	$5 + 17 = 22$		

3.5 AN OPTIMAL CONTROL SYSTEM

Consider a system described by the first-order differential equation

$$\frac{d}{dt}[x(t)] = ax(t) + bu(t), \tag{3.5-1}$$

where $x(t)$ and $u(t)$ are the state and control variables, respectively, and a and b are constants. The admissible values of the state and control variables are constrained by

$$0.0 \leq x(t) \leq 1.5$$

and

$$-1.0 \leq u(t) \leq 1.0, \tag{3.5-2}$$

and the performance measure (cost) to be minimized is

$$J = x^2(T) + \lambda \int_0^T u^2(t)\, dt, \tag{3.5-3}$$

where T is the specified final time, and λ is a weighting factor included to permit adjustment of the relative importance of the two terms in J. $x(T)$ and $u(t)$ are squared because positive and negative values of these quantities are of equal importance. This performance measure reflects the desire to drive the final state $x(T)$ close to zero without excessive expenditure of control effort.

Before the numerical procedure of dynamic programming can be applied, the system differential equation must be approximated by a difference equation, and the integral in the performance measure must be approximated by a summation. This can be done most conveniently by dividing the time interval $0 \leq t \leq T$ into N equal increments, Δt. Then from (3.5-1)

$$\frac{x(t + \Delta t) - x(t)}{\Delta t} \approx ax(t) + bu(t), \tag{3.5-4}$$

or

$$x(t + \Delta t) = [1 + a\,\Delta t]x(t) + b\,\Delta t\, u(t). \tag{3.5-5}$$

It will be assumed that Δt is small enough so that the control signal can be approximated by a piecewise-constant function that changes only at the instants $t = 0, \Delta t, \ldots, (N-1)\,\Delta t$; thus, for $t = k\,\Delta t$,

$$x([k+1]\,\Delta t) = [1 + a\,\Delta t]x(k\,\Delta t) + b\,\Delta t\, u(k\,\Delta t); \qquad k = 0, 1, \ldots, N-1. \tag{3.5-6}$$

$x(k\,\Delta t)$ is referred to as the kth value of x and is denoted by $x(k)$. The system difference equation can then be written

$$x(k+1) = [1 + a\,\Delta t]x(k) + b\,\Delta t\, u(k). \tag{3.5-7}$$

In a similar way the performance measure becomes

$$J = x^2(N\,\Delta t) + \lambda\left[\int_0^{\Delta t} u^2(0)\,dt + \int_{\Delta t}^{2\,\Delta t} u^2(\Delta t)\,dt + \cdots + \int_{(N-1)\,\Delta t}^{N\,\Delta t} u^2([N-1]\,\Delta t)\,dt\right], \tag{3.5-8}$$

or,

$$\begin{aligned} J &= x^2(N) + \lambda\,\Delta t[u^2(0) + u^2(1) + \cdots + u^2(N-1)] \\ &= x^2(N) + \lambda\,\Delta t \sum_{k=0}^{N-1} u^2(k). \end{aligned} \tag{3.5-9}$$

Now the method of dynamic programming can be applied as in the automobile routing problem. For numerical simplicity let $a = 0$, $b = 1$, $\lambda = 2$, $T = 2$, $\Delta t = 1$, in which case $N = 2$; i.e., this is a two-stage process described by the difference equation

$$x(k+1) = x(k) + u(k); \qquad k = 0, 1 \tag{3.5-10}$$

where $u(0)$ and $u(1)$ are to be selected to minimize the performance measure (cost)

$$J = x^2(2) + 2u^2(0) + 2u^2(1) \tag{3.5-11}$$

subject to the constraints

$$0.0 \leq x(k) \leq 1.5; \qquad k = 0, 1, 2$$

and (3.5-12)

$$-1.0 \leq u(k) \leq 1.0; \qquad k = 0, 1.$$

The first step in the computational procedure is to find the optimal policy for the last stage of operation. This is essentially a matter of trying *all* of the allowable control values at *each* of the allowable state values. The optimal control for each state value is the one which yields the trajectory having the minimum cost. To limit the required number of calculations, and thereby make the computational procedure feasible, the allowable state and control values must be quantized. In this problem it will be assumed that the quantized values are $x(k) = 0.0, 0.5, 1.0, 1.5$, and $u(k) = -1.0, -0.5, 0.0, 0.5, 1.0$.

Using these values, we find that the computational procedure for determining the optimal policy over the last stage is

1. Put $k = 1$, select one of the quantized values of $x(1)$, try all quantized values of $u(1)$, and calculate the resulting trajectories. The optimal

control for this state value is the one which yields the minimum cost.

2. Repeat the procedure in step 1 for the remaining quantized levels of $x(1)$.

The resulting calculations are shown in Table 3-2, where calculations leading to a violation of the constraints have been omitted. Notice that the cost J_{12} of going from state $x(1)$ to state $x(2)$ is dependent on the *value* of the state $x(1)$ and on the *value* of the control applied, $u(1)$; hence the notation $J_{12}(x(1), u(1))$. Similarly, the minimum cost $J_{12}^*(x(1))$ and the optimal control $u^*(x(1), 1)$ applied at $k = 1$ are dependent on the value of the state $x(1)$.†

Now consider the next-to-last stage of the process by putting $k = 0$. At each quantized value of the state $x(0)$ all quantized values of the control $u(0)$ are tried. The trajectory from $x(0)$ to $x(1)$ is computed for each trial, together with the cost J_{01}. Then, knowing the value of $x(1)$ at the end of each such trajectory, we may follow the optimal trajectory over the last stage with the aid of the data available in Table 3-2. In mathematical terms this means that

$$C_{02}^*(x(0), u(0)) = J_{01}(x(0), u(0)) + J_{12}^*(x(1)), \tag{3.5-13}$$

and thus the cost of the optimal trajectory is

$$J_{02}^*(x(0)) = \min_{u(0)} [J_{01}(x(0), u(0)) + J_{12}^*(x(1))], \tag{3.5-14}$$

where

$C_{02}^*(x(0), u(0))$ is the minimum cost of operation over the last two stages for one quantized value of $x(0)$ *given a particular trial quantized value of* $u(0)$.

$J_{01}(x(0), u(0))$ is the cost of operation in the interval $k = 0$ to $k = 1$ for specified quantized values of $x(0)$ and $u(0)$.

$J_{12}^*(x(1))$ is the cost of the optimal last-stage trajectory which is a function of the state $x(1)$.

$J_{02}^*(x(0))$ is the minimum cost of operation over the last two stages for a specified quantized value of $x(0)$.

Notice that (3.5-13) and (3.5-14) are analogous to (3.4-6) and (3.4-7) in the automobile routing problem.

Finally, (3.5-13) and (3.5-14) are mechanized in Table 3-3 to complete the dynamic programming algorithm.

The information contained in Tables 3-2 and 3-3 may now be used to determine the optimal trajectory from any allowable quantized value of $x(0)$

† Rather than adhere to the form of (3.1-1) for the optimal control law, $\mathbf{u}^*(k) = \mathbf{f}(\mathbf{x}(k), k)$, we will shorten the notation by writing simply $\mathbf{u}^*(\mathbf{x}(k), k)$.

Table 3-2 COSTS OF OPERATION OVER THE LAST STAGE

Current state $x(1)$	*Control* $u(1)$	*Next state* $x(2) = x(1) + u(1)$	*Cost* $x^2(2) + 2u^2(1) = J_{12}(x(1), u(1))$	*Minimum cost* $J^*_{12}(x(1))$	*Optimal control applied at* $k = 1$ $u^*(x(1), 1)$
1.5	0.0	1.5	$(1.5)^2 + 2(0.0)^2 = 2.25$		
	−0.5	1.0	$(1.0)^2 + 2(-0.5)^2 = 1.50$	$J^*_{12}(1.5) = 1.50$	$u^*(1.5, 1) = -0.5$
	−1.0	0.5	$(0.5)^2 + 2(-1.0)^2 = 2.25$		
1.0	0.5	1.5	$(1.5)^2 + 2(0.5)^2 = 2.75$		
	0.0	1.0	$(1.0)^2 + 2(0.0)^2 = 1.00$		
	−0.5	0.5	$(0.5)^2 + 2(-0.5)^2 = 0.75$	$J^*_{12}(1.0) = 0.75$	$u^*(1.0, 1) = -0.5$
	−1.0	0.0	$(0.0)^2 + 2(-1.0)^2 = 2.00$		
0.5	1.0	1.5	$(1.5)^2 + 2(1.0)^2 = 4.25$		
	0.5	1.0	$(1.0)^2 + 2(0.5)^2 = 1.50$		
	0.0	0.5	$(0.5)^2 + 2(0.0)^2 = 0.25$	$J^*_{12}(0.5) = 0.25$	$u^*(0.5, 1) = 0.0$
	−0.5	0.0	$(0.0)^2 + 2(-0.5)^2 = 0.50$		
0.0	1.0	1.0	$(1.0)^2 + 2(1.0)^2 = 3.00$		
	0.5	0.5	$(0.5)^2 + 2(0.5)^2 = 0.75$		
	0.0	0.0	$(0.0)^2 + 2(0.0)^2 = 0.00$	$J^*_{12}(0.0) = 0.00$	$u^*(0.0, 1) = 0.0$

Table 3-3 COSTS OF OPERATION OVER THE LAST TWO STAGES

Current state $x(0)$	*Control* $u(0)$	*Next state* $x(1) = x(0) + u(0)$	*Minimum cost over last two stages for trial value* $u(0)$: $J_{01}(x(0), u(0)) + J_{12}^*(x(1)) = 2u^2(0) + J_{12}^*(x(1)) = C_{02}^*(x(0), u(0))$	*Minimum cost over last two stages* $J_{02}^*(x(0))$	*Optimal control applied at* $k = 0$ $u^*(x(0), 0)$
1.5	0.0	1.5	$2(0.0)^2 + 1.50 = 1.50$		
	−0.5	1.0	$2(-0.5)^2 + 0.75 = 1.25$	$J_{02}^*(1.5) = 1.25$	$u^*(1.5, 0) = -0.5$
	−1.0	0.5	$2(-1.0)^2 + 0.25 = 2.25$		
1.0	0.5	1.5	$2(0.5)^2 + 1.50 = 2.00$		
	0.0	1.0	$2(0.0)^2 + 0.75 = 0.75$	$J_{02}^*(1.0) = \begin{Bmatrix} 0.75 \\ 0.75 \end{Bmatrix}$	$u^*(1.0, 0) = \begin{Bmatrix} 0.0 \\ -0.5 \end{Bmatrix}$
	−0.5	0.5	$2(-0.5)^2 + 0.25 = 0.75$		
	−1.0	0.0	$2(-1.0)^2 + 0.00 = 2.00$		
0.5	1.0	1.5	$2(1.0)^2 + 1.50 = 3.50$		
	0.5	1.0	$2(0.5)^2 + 0.75 = 1.25$		
	0.0	0.5	$2(0.0)^2 + 0.25 = 0.25$	$J_{02}^*(0.5) = 0.25$	$u^*(0.5, 0) = 0.0$
	−0.5	0.0	$2(-0.5)^2 + 0.00 = 0.50$		
0.0	1.0	1.0	$2(1.0)^2 + 0.75 = 2.75$		
	0.5	0.5	$2(0.5)^2 + 0.25 = 0.75$		
	0.0	0.0	$2(0.0)^2 + 0.00 = 0.00$	$J_{02}^*(0.0) = 0.00$	$u^*(0.0, 0) = 0.0$

to the final state $x(2)$. For example, if $x(0) = 1.5$, Table 3-3 indicates that $u^*(1.5, 0) = -0.5$ and $J_{02}^*(1.5) = 1.25$. Application of $u^*(1.5, 0)$ at $x(0) = 1.5$ makes $x(1) = 1.0$, and Table 3-2 gives the optimal control applied at $k = 1$ as $u^*(1.0, 1) = -0.5$. Thus, for $x(0) = 1.5$ the optimal control sequence is $\{-0.5, -0.5\}$, and the minimum cost is 1.25.

In a similar way, the optimal policies and trajectories can be determined from the tables for the other values of $x(0)$. Observe that if $x(0) = 1.0$ the optimal policy is nonunique—the control sequences $\{0, -0.5\}$ and $\{-0.5, 0\}$ are both optimal. Notice also that in this problem there is no requirement that all the trajectories end at the same value of $x(2)$. A problem in which $x(T)$ is specified is included in the problems at the end of the chapter (Problem 3-3).

If a problem is segmented into more than two stages, the procedure must simply be extended by repeating the calculations of Table 3-3 for each preceding stage. In general, to determine the optimal control applied at $t = k\,\Delta t$ in an N-stage process the appropriate forms for (3.5-13) and (3.5-14) are

$$C_{kN}^*(x(k), u(k)) = J_{k,k+1}(x(k), u(k)) + J_{k+1,N}^*(x(k+1)), \tag{3.5-13a}$$

$$J_{kN}^*(x(k)) = \min_{u(k)} [C_{kN}^*(x(k), u(k))]. \tag{3.5-14a}$$

Taken together, equations (3.5-13a) and (3.5-14a) form the *functional equation of dynamic programming;* we shall have more to say about this in Section 3.7.

In more practical problems a digital computer would normally be needed, and it often becomes important to minimize the amount of storage required for the retention of intermediate results. The calculations in Table 3-3 and the determination of the optimal policy and trajectory for any allowable value of $x(0)$ require only the data in the last two columns of Tables 3-2 and 3-3; therefore, only these data need be stored.

3.6 INTERPOLATION

In the preceding control example all of the trial control values drive the state of the system either to a computational "grid" point or to a value outside of the allowable range. Had the numerical values not been carefully selected, this happy situation would not have been obtained and interpolation would have been required. For example, suppose that the trial values for $u(k)$ had been $-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1$. The values of $J_{12}^*(x(1))$ and $u^*(x(1), 1)$ shown next to the state points in Fig. 3-4(a) are

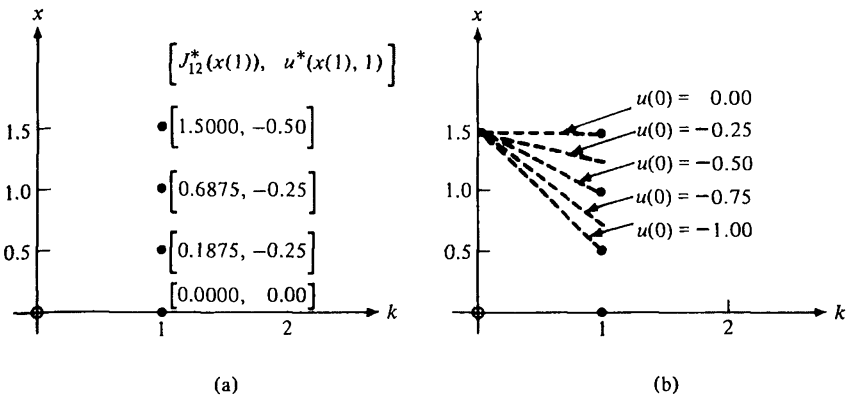


Figure 3-4 (a) Minimum costs and optimal controls for quantized values of $x(1)$. (b) Paths resulting from the application of quantized control values at $x(0) = 1.5$

the results of repeating the calculations in Table 3-2 with the new trial values for $u(1)$.

Next, suppose that all of the quantized values of the control are applied for a state value of $x(0) = 1.5$. The resulting values of $x(1)$ are shown in Fig. 3-4(b), where it can be seen that two of the end points do not coincide with the grid points of Fig. 3-4(a). But, by linear interpolation,

$$\begin{aligned} J_{12}^*(1.25) &= 0.68750 + \tfrac{1}{2}[1.50000 - 0.68750] \\ &= 1.09375 \end{aligned} \tag{3.6-1}$$

and

$$\begin{aligned} J_{12}^*(0.75) &= 0.18750 + \tfrac{1}{2}[0.68750 - 0.18750] \\ &= 0.43750 \end{aligned} \tag{3.6-2}$$

Finally, the result of repeating the calculations in Table 3-3 [for $x(0) = 1.5$ only], the interpolated values of $J_{12}^*(x(1))$ being used where required, is shown in Table 3-4.

Interpolation may also be required when one is using stored data to calculate an optimal control sequence. For example, if the optimal control applied at some value of $x(0)$ drives the system to a state value $x(1)$ that is halfway between two points where the optimal controls are -1 and -0.5, then by linear interpolation the optimal control is -0.75.

In summary, although a finite grid of state and control values must be employed in the numerical procedure, interpolation makes available approximate information about intermediate points. Naturally, the degree of approxi-

Table 3-4 COSTS OF OPERATION OVER THE LAST TWO STAGES FOR $x(0) = 1.50$

Current state $x(0)$	*Control* $u(0)$	*Next state* $x(1) = x(0) + u(0)$	*Minimum cost over last two stages for trial value* $u(0)$: $J_{01}(x(0), u(0)) + J_{12}^*(x(1)) =$ $2u^2(0) + J_{12}^*(x(1)) = C_{02}^*(x(0), u(0))$	*Minimum cost over last two stages* $J_{02}^*(x(0))$	*Optimal control applied at* $k = 0$ $u^*(x(0), 0)$
1.50	0.00	1.50	$2(0.00)^2 + 1.50000 = 1.50000$		
	−0.25	1.25	$2(-0.25)^2 + 1.09375 = 1.21875$		
	−0.50	1.00	$2(-0.50)^2 + 0.68750 = 1.18750$	$J_{02}^*(1.5) = 1.18750$	$u^*(1.5, 0) = -0.50$
	−0.75	0.75	$2(-0.75)^2 + 0.43750 = 1.56250$		
	−1.00	0.50	$2(-1.00)^2 + 0.18750 = 2.18750$		

mation depends on the separation of the grid points, the interpolation scheme used, and the system dynamics and performance measure. A finer grid generally means greater accuracy, but also increased storage requirements and computation time. The effects of these factors are illustrated in some of the exercises at the end of the chapter (Problems 3-14 through 3-18).

3.7 A RECURRENCE RELATION OF DYNAMIC PROGRAMMING

In this section we shall begin to formalize some of the ideas introduced intuitively in preceding sections. In particular, we wish to generalize the procedure in Section 3.5 which led to equations (3.5-13a) and (3.5-14a). Since our attention is focused on control systems, a recurrence relation will be derived by applying dynamic programming to a control process.

An nth-order time-invariant system† is described by the state equation

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t)). \tag{3.7-1}$$

It is desired to determine the control law which minimizes the performance measure

$$J = h(\mathbf{x}(t_f)) + \int_0^{t_f} g(\mathbf{x}(t), \mathbf{u}(t))\, dt, \tag{3.7-2}$$

where t_f is assumed fixed. The admissible controls are constrained to lie in a set U; i.e., $\mathbf{u} \in U$. As before, we first approximate the continuously operating system of Eq. (3.7-1) by a discrete system; this is accomplished by considering N equally spaced time increments in the interval $0 \leq t \leq t_f$. From (3.7-1)

$$\frac{\mathbf{x}(t + \Delta t) - \mathbf{x}(t)}{\Delta t} \approx \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t)) \tag{3.7-3}$$

or

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \Delta t\, \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t)). \tag{3.7-4}$$

Using the shorthand notation developed earlier for $\mathbf{x}(k\,\Delta t)$ gives

$$\mathbf{x}(k + 1) = \mathbf{x}(k) + \Delta t\, \mathbf{a}(\mathbf{x}(k), \mathbf{u}(k)), \tag{3.7-5}$$

which we will denote by

$$\mathbf{x}(k + 1) \triangleq \mathbf{a}_D(\mathbf{x}(k), \mathbf{u}(k)). \tag{3.7-6}$$

† The following derivation can be applied to time-varying systems as well; time-invariance is assumed only to simplify the notation.

Operating on the performance measure in a similar manner, we obtain

$$J = h(\mathbf{x}(N\,\Delta t)) + \int_0^{\Delta t} g\,dt + \int_{\Delta t}^{2\,\Delta t} g\,dt + \cdots + \int_{(N-1)\,\Delta t}^{N\,\Delta t} g\,dt, \quad (3.7\text{-}7)$$

which becomes for small Δt,

$$J \approx h(\mathbf{x}(N)) + \Delta t \sum_{k=0}^{N-1} g(\mathbf{x}(k), \mathbf{u}(k)), \quad (3.7\text{-}8)$$

which we shall denote by

$$J = h(\mathbf{x}(N)) + \sum_{k=0}^{N-1} g_D(\mathbf{x}(k), \mathbf{u}(k)). \quad (3.7\text{-}8a)$$

By making the problem discrete as we have done, it is now required that the optimal control law $\mathbf{u}^*(\mathbf{x}(0), 0), \mathbf{u}^*(\mathbf{x}(1), 1), \ldots, \mathbf{u}^*(\mathbf{x}(N-1), N-1)$ be determined for the system given by Eq. (3.7-6) which has the performance measure given by (3.7-8a). We are now ready to derive the recurrence equation.

Begin by defining

$$J_{NN}(\mathbf{x}(N)) \triangleq h(\mathbf{x}(N)); \quad (3.7\text{-}9)$$

J_{NN} is the cost of reaching the final state value $\mathbf{x}(N)$. Next, define

$$\begin{aligned} J_{N-1,\,N}(\mathbf{x}(N-1), \mathbf{u}(N-1)) &\triangleq g_D(\mathbf{x}(N-1), \mathbf{u}(N-1)) + h(\mathbf{x}(N)) \\ &= g_D(\mathbf{x}(N-1), \mathbf{u}(N-1)) + J_{NN}(\mathbf{x}(N)), \end{aligned} \quad (3.7\text{-}10)$$

which is the cost of operation during the interval $(N-1)\,\Delta t \leq t \leq N\,\Delta t$. Observe that $J_{N-1,\,N}$ is also the cost of a *one-stage process* with initial state $\mathbf{x}(N-1)$. The value of $J_{N-1,\,N}$ is dependent only on $\mathbf{x}(N-1)$ and $\mathbf{u}(N-1)$, since $\mathbf{x}(N)$ is related to $\mathbf{x}(N-1)$ and $\mathbf{u}(N-1)$ through the state equation (3.7-6), so we write

$$\begin{aligned} J_{N-1,\,N}(\mathbf{x}(N-1), \mathbf{u}(N-1)) &= g_D(\mathbf{x}(N-1), \mathbf{u}(N-1)) \\ &\quad + J_{NN}(\mathbf{a}_D(\mathbf{x}(N-1), \mathbf{u}(N-1))). \end{aligned} \quad (3.7\text{-}11)$$

The optimal cost is then

$$\begin{aligned} J^*_{N-1,\,N}(\mathbf{x}(N-1)) \triangleq \min_{\mathbf{u}(N-1)} \{ & g_D(\mathbf{x}(N-1), \mathbf{u}(N-1)) + J_{NN}(\mathbf{a}_D(\mathbf{x}(N-1), \\ & \mathbf{u}(N-1)))\}\dagger \end{aligned} \quad (3.7\text{-}12)$$

† Notice that the minimization is performed with only admissible control values being used.

We know that the optimal choice of $\mathbf{u}(N-1)$ will depend on $\mathbf{x}(N-1)$, so we denote the minimizing control by $\mathbf{u}^*(\mathbf{x}(N-1), N-1)$.

The cost of operation over the last two intervals is given by

$$\begin{aligned} &J_{N-2,\,N}(\mathbf{x}(N-2), \mathbf{u}(N-2), \mathbf{u}(N-1)) \\ &\quad = g_D(\mathbf{x}(N-2), \mathbf{u}(N-2)) + g_D(\mathbf{x}(N-1), \mathbf{u}(N-1)) + h(\mathbf{x}(N)) \\ &\quad = g_D(\mathbf{x}(N-2), \mathbf{u}(N-2)) + J_{N-1,\,N}(\mathbf{x}(N-1), \mathbf{u}(N-1)), \end{aligned} \tag{3.7-13}$$

where again we have used the dependence of $\mathbf{x}(N)$ on $\mathbf{x}(N-1)$ and $\mathbf{u}(N-1)$. As before, observe that $J_{N-2,\,N}$ is the cost of a *two-stage process* with initial state $\mathbf{x}(N-2)$. The optimal policy during the last two intervals is found from

$$\begin{aligned} &J^*_{N-2,\,N}(\mathbf{x}(N-2)) \\ &\quad \triangleq \min_{\mathbf{u}(N-2),\,\mathbf{u}(N-1)} \{g_D(\mathbf{x}(N-2), \mathbf{u}(N-2)) + J_{N-1,\,N}(\mathbf{x}(N-1), \mathbf{u}(N-1))\} \end{aligned} \tag{3.7-14}$$

The principle of optimality states that for this two-stage process, whatever the initial state $\mathbf{x}(N-2)$ and initial decision $\mathbf{u}(N-2)$, the remaining decision $\mathbf{u}(N-1)$ must be optimal with respect to the value of $\mathbf{x}(N-1)$ that results from application of $\mathbf{u}(N-2)$; therefore,

$$J^*_{N-2,\,N}(\mathbf{x}(N-2)) = \min_{\mathbf{u}(N-2)} \{g_D(\mathbf{x}(N-2), \mathbf{u}(N-2)) + J^*_{N-1,\,N}(\mathbf{x}(N-1))\}. \tag{3.7-15}$$

Since $\mathbf{x}(N-1)$ is related to $\mathbf{x}(N-2)$ and $\mathbf{u}(N-2)$ by the state equation, $J^*_{N-2,\,N}$ depends only on $\mathbf{x}(N-2)$; thus

$$\begin{aligned} &J^*_{N-2,\,N}(\mathbf{x}(N-2)) \\ &\quad = \min_{\mathbf{u}(N-2)} \{g_D(\mathbf{x}(N-2), \mathbf{u}(N-2)) + J^*_{N-1,\,N}(\mathbf{a}_D(\mathbf{x}(N-2), \mathbf{u}(N-2)))\} \end{aligned} \tag{3.7-15a}$$

By considering the cost of operation over the final *three* stages—a *three-stage process* with initial state $\mathbf{x}(N-3)$—we can follow exactly the same reasoning which led to Eqs. (3.7-13) through (3.7-15a) to obtain

$$\begin{aligned} &J^*_{N-3,\,N}(\mathbf{x}(N-3)) \\ &\quad = \min_{\mathbf{u}(N-3)} \{g_D(\mathbf{x}(N-3), \mathbf{u}(N-3)) + J^*_{N-2,\,N}(\mathbf{a}_D(\mathbf{x}(N-3), \mathbf{u}(N-3)))\} \end{aligned} \tag{3.7-16}$$

Continuing backward in this manner, we obtain for a K-stage process the result

$$\begin{aligned} &J^*_{N-K,\,N}(\mathbf{x}(N-K)) \\ &\quad = \min_{\mathbf{u}(N-K),\,\mathbf{u}(N-K+1),\,\ldots,\,\mathbf{u}(N-1)} \left\{h(\mathbf{x}(N)) + \sum_{k=N-K}^{N-1} g_D(\mathbf{x}(k), \mathbf{u}(k))\right\}, \end{aligned} \tag{3.7-17}$$

which by applying the principle of optimality becomes

$$J^*_{N-K,\,N}(\mathbf{x}(N-K)) = \min_{\mathbf{u}(N-K)} \{g_D(\mathbf{x}(N-K), \mathbf{u}(N-K)) + J^*_{N-(K-1),\,N}(\mathbf{a}_D(\mathbf{x}(N-K), \mathbf{u}(N-K)))\}.\dagger \tag{3.7-18}$$

Equation (3.7-18) is the recurrence relation that we set out to obtain. By knowing $J^*_{N-(K-1),\,N}$, the optimal cost for a $(K-1)$-stage policy, we can generate $J^*_{N-K,\,N}$, the optimal cost for a K-stage policy. To begin the process we simply start with a zero-stage process and generate $J^*_{NN} \triangleq J_{NN}$ (the * is just a notational convenience here; no choice of a control is implied). Next, the optimal cost can be found for a one-stage process by using J^*_{NN} and (3.7-18), and so on. Notice that beginning with a zero-stage process corresponds to starting at the terminal state h in the routing problem of Section 3.4 and starting at the final time $t = 2\,\Delta t$ in the control example of Section 3.5.

This derivation of the recurrence equation has also revealed another important concept—the *imbedding principle*. $J^*_{N-K,\,N}(\mathbf{x}(N-K))$ is the minimum cost possible for the final K stages of an N-stage process with state value $\mathbf{x}(N-K)$ at the beginning of the $(N-K)$th stage; however, $J^*_{N-K,\,N}(\mathbf{x}(N-K))$ is also the minimum cost possible for a K-stage process with initial state numerically equal to the value $\mathbf{x}(N-K)$. This means that the optimal policy and minimum costs for a K-stage process are contained (or imbedded) in the results for an N-stage process, provided that $N \geq K$.

Our discussion has been concerned primarily with the solution of optimal control problems; however, dynamic programming can also be applied to other types of optimization problems. For a more general treatment of dynamic programming and its applications, see references [B-2] and [N-1].

3.8 COMPUTATIONAL PROCEDURE FOR SOLVING OPTIMAL CONTROL PROBLEMS

Let us now summarize the dynamic programming computational procedure for determining optimal policies.

† An alternative notation often used is

$$J^*_K(\mathbf{x}(N-K)) = \min_{\mathbf{u}(N-K)} \{g_D(\mathbf{x}(N-K), \mathbf{u}(N-K)) + J^*_{K-1}(\mathbf{a}_D(\mathbf{x}(N-K), \mathbf{u}(N-K)))\},$$

where the subscripts of J^* indicate the number of stages. We shall use the notation of Eq. (3.7-18) because it more clearly indicates the computational procedure to be followed.

A system is described by the state difference equation†

$$\mathbf{x}(k+1) = \mathbf{a}_D(\mathbf{x}(k), \mathbf{u}(k)); \qquad k = 0, 1, \ldots, N-1. \tag{3.8-1}$$

It is desired to determine the control law that minimizes the criterion

$$J = h(\mathbf{x}(N)) + \sum_{k=0}^{N-1} g_D(\mathbf{x}(k), \mathbf{u}(k)).‡ \tag{3.8-2}$$

As shown in Section 3.7, the application of dynamic programming to this problem leads to the recurrence equation

$$J^*_{N-K,\,N}(\mathbf{x}(N-K)) = \min_{\mathbf{u}(N-K)} \{g_D(\mathbf{x}(N-K), \mathbf{u}(N-K)) + J^*_{N-(K-1),\,N}(\mathbf{a}_D(\mathbf{x}(N-K), \mathbf{u}(N-K)))\}; \qquad K = 1, 2, \ldots, N \tag{3.8-3}$$

with initial value

$$J^*_{NN}(\mathbf{x}(N)) = h(\mathbf{x}(N)). \tag{3.8-4}$$

It should be re-emphasized that Eq. (3.8-3) is simply a formalization of the computational procedure followed in solving the control problem in Section 3.5.

The solution of this recurrence equation is an optimal control law or optimal policy, $\mathbf{u}^*(\mathbf{x}(N-K), N-K)$, $K = 1, 2, \ldots, N$, which is obtained by trying *all* admissible control values at *each* admissible state value. To make the computational procedure feasible it is necessary to quantize the admissible state and control values into a finite number of levels. For example, if the system is second order, the grid of state values would appear as shown in Fig. 3-5. The heavily dotted points are the state values at which each of the quantized control values is to be tried. In this second-order example, the total number of state grid points for each time, $k\,\Delta t$, is $s_1 s_2$, where s_1 is the number of points in the x_1 coordinate direction and s_2 is the number of points in the x_2 coordinate direction. s_1 and s_2 are determined by the relationship

$$s_r = \frac{x_{r\max} - x_{r\min}}{\Delta x_r} + 1; \qquad r = 1, 2, \tag{3.8-5}$$

where it is assumed that Δx_r is selected so that the interval $x_{r\max} - x_{r\min}$

† This difference equation and the performance measure may be a discrete approximation to a continuous system, or they may represent a system that is actually discrete.

‡ To simplify the notation, it is assumed that the state equations and performance measure do not contain k explicitly. The algorithm is easily modified if this is not the case.

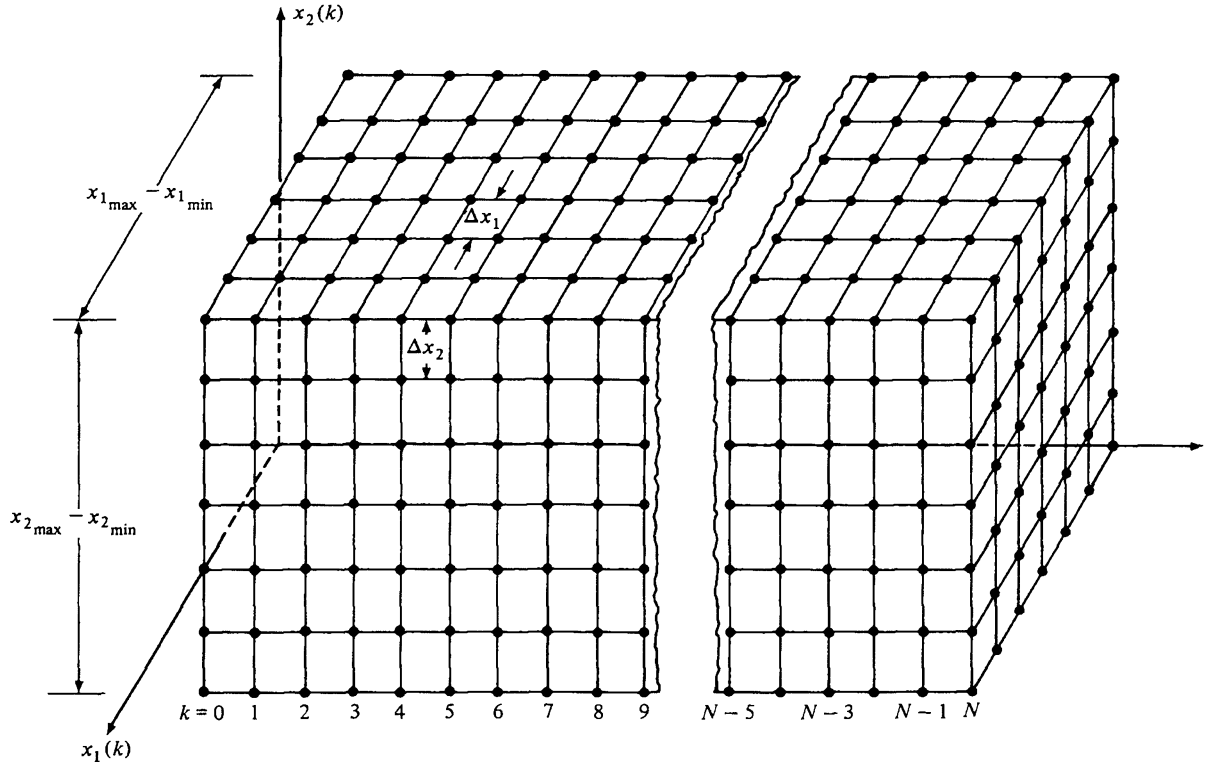


Figure 3-5 Grid of admissible state values

contains an integer number of points. For an nth-order system the number of state grid points for each time, $t = k\,\Delta t$, is

$$S = s_1 \cdot s_2 \cdot \cdots \cdot s_n \tag{3.8-6}$$

where

$$s_r = \frac{x_{r_{\max}} - x_{r_{\min}}}{\Delta x_r} + 1; \qquad r = 1, 2, \ldots, n, \tag{3.8-7}$$

if it is assumed that the ratio $[x_{r_{\max}} - x_{r_{\min}}]/\Delta x_r$ is an integer. The admissible range of control values is quantized in exactly the same way; if C is the total number of quantized values of $\mathbf{u}(k)$, then

$$C = c_1 \cdot c_2 \cdot \cdots \cdot c_m \tag{3.8-8}$$

where

$$c_q = \frac{u_{q_{\max}} - u_{q_{\min}}}{\Delta u_q} + 1; \qquad q = 1, 2, \ldots, m. \tag{3.8-9}$$

In the following development $\mathbf{x}^{(i)}(k)(i = 1, 2, \ldots, S)$ and $\mathbf{u}^{(j)}(k)$ $(j = 1, 2, \ldots, C)$ denote the admissible quantized state and control values at time $t = k\,\Delta t$.

The first step in the computational procedure is to calculate the values of $J^*_{NN}(\mathbf{x}^{(i)}(N))$ $(i = 1, 2, \ldots, S)$ which are used to begin solution of the recurrence equation.

Next, we set $K = 1$, and select the first trial state point by putting $i = 1$ in the subroutine which generates the points $\mathbf{x}^{(i)}(N - K)$. Each control value, $\mathbf{u}^{(j)}(N - K)$ $(j = 1, 2, \ldots, C)$, is then tried at the state value $\mathbf{x}^{(i)}(N - K)$ to determine the next state value, $\mathbf{x}^{(i,j)}(N - K + 1)$, which is used to look up the appropriate value of $J^*_{N-(K-1),N}(\mathbf{x}^{(i,j)}(N - K + 1))$ in computer memory—interpolation will be required if $\mathbf{x}^{(i,j)}(N - K + 1)$ does not fall exactly on a grid value. Using this value of $J^*_{N-(K-1),N}(\mathbf{x}^{(i,j)}(N - K + 1))$ we evaluate

$$C^*_{N-K,N}(\mathbf{x}^{(i)}(N - K), \mathbf{u}^{(j)}(N - K)) = g_D(\mathbf{x}^{(i)}(N - K), \mathbf{u}^{(j)}(N - K)) + J^*_{N-(K-1),N}(\mathbf{x}^{(i,j)}(N - K + 1)), \tag{3.8-10}$$

which is the minimum cost of operation over the final K stages of an N-stage process *assuming that* the control value $\mathbf{u}^{(j)}(N - K)$ is applied at the state value $\mathbf{x}^{(i)}(N - K)$. The idea is to find the value of $\mathbf{u}^{(j)}(N - K)$ that yields $J^*_{N-K,N}(\mathbf{x}^{(i)}(N - K))$, the minimum of $C^*_{N-K,N}(\mathbf{x}^{(i)}(N - K), \mathbf{u}^{(j)}(N - K))$. Only the smallest value of $C^*_{N-K,N}(\mathbf{x}^{(i)}(N - K), \mathbf{u}^{(j)}(N - K))$ and the associated control need to be retained in storage; thus, as each control value is applied at $\mathbf{x}^{(i)}(N - K)$ the $C^*_{N-K,N}(\mathbf{x}^{(i)}(N - K), \mathbf{u}^{(j)}(N - K))$ that results is compared with the variable named COSMIN—the COSt which is the MINi-

mum of those which have been previously calculated. If $C^*_{N-K,N}(\mathbf{x}^{(i)}(N-K), \mathbf{u}^{(j)}(N-K)) < \text{COSMIN}$, then the current value of COSMIN is replaced by this new smaller value. The control that corresponds to the value of COSMIN is also retained—as the variable named UMIN. Naturally, when COSMIN is changed, so is UMIN.

After all control values have been tried at the state value $\mathbf{x}^{(i)}(N-K)$, the numbers stored in COSMIN and UMIN are transferred to storage in arrays named COST$(N-K, I)$ and UOPT$(N-K, I)$, respectively. The arguments $(N-K)$ and I indicate that these values correspond to the state value $\mathbf{x}^{(i)}(N-K)$.

The above procedure is carried out for each quantized state value; then K is increased by one and the procedure is repeated until $K = N$, at which point the COST$(N-K, I)$ and UOPT$(N-K, I)$ arrays are printed out for $K = 1, 2, \ldots, N$ and $I = 1, 2, \ldots, S$. A flow chart of the computational procedure is shown in Fig. 3-6.

The result of the computational procedure is a number for the optimal

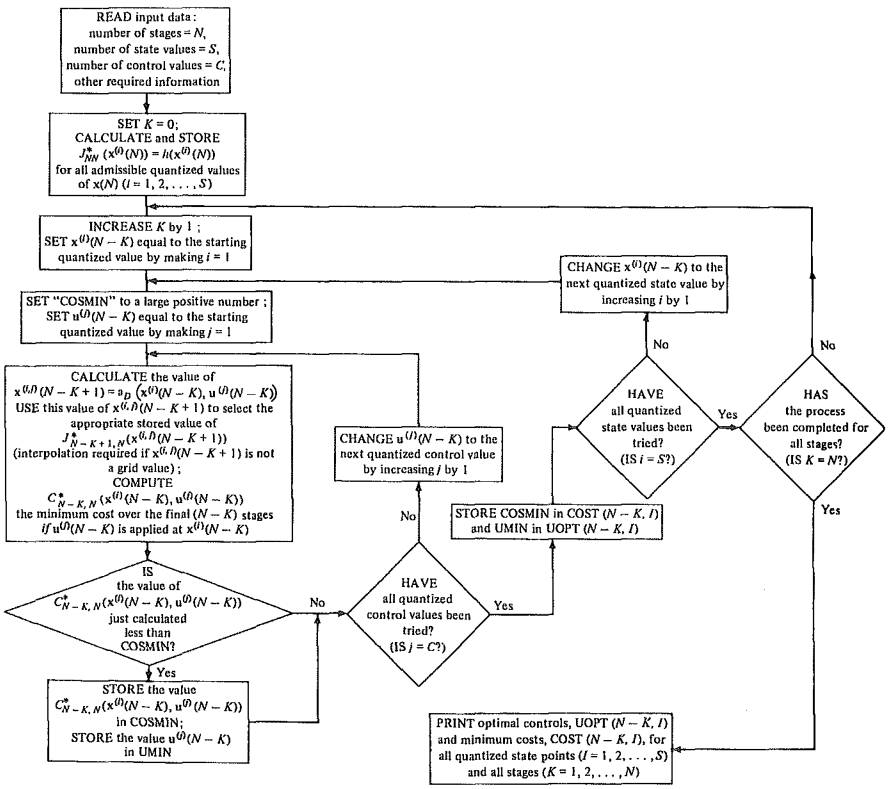


Figure 3-6 Flow chart of the computational procedure

control and the minimum cost at every point on the $(n + 1)$-dimensional state-time grid. To calculate the optimal control sequence for a given initial condition, we enter the storage location corresponding to the specified initial condition and extract the control value $\mathbf{u}^*(0)$ and the minimum cost. Next, by solving the state equation we determine the state of the system at $k = 1$ which results from applying $\mathbf{u}^*(0)$ at $k = 0$. The resulting value of $\mathbf{x}(1)$ is then used to reenter the table and extract $\mathbf{u}^*(1)$, and so on. We see that the optimal controller is physically realized by a table look-up device and a generator of piecewise-constant signals.

3.9 CHARACTERISTICS OF DYNAMIC PROGRAMMING SOLUTION

In Section 3.8 we formalized the algorithm for computing the optimal control law from the functional equation

$$J^*_{N-K,N}(\mathbf{x}(N-K)) = \min_{\mathbf{u}(N-K)} \{g_D(\mathbf{x}(N-K), \mathbf{u}(N-K)) + J^*_{N-(K-1),N}(\mathbf{a}_D(\mathbf{x}(N-K), \mathbf{u}(N-K)))\}. \tag{3.8-3}$$

Let us now summarize the important characteristics of the computational procedure and the solution it provides.

Absolute Minimum

Since a direct search is used to solve the functional recurrence equation (3.8-3), the solution obtained is the absolute (or global) minimum. Dynamic programming makes the direct search feasible because instead of searching among the set of *all* admissible controls that cause admissible trajectories, we consider only those controls that satisfy an additional necessary condition—the principle of optimality. This concept is illustrated in Fig. 3-7. S_1 is the set of all controls; S_2 is the set of admissible controls; S_3 is the set of controls that yield admissible state trajectories; S_4 is the set of controls that satisfy the principle of optimality. Without the principle of optimality we would search in the intersection of sets S_2 and S_3.† The dynamic programming algorithm, however, searches only in the shaded region—the intersection of S_2, S_3, and S_4 $(S_2 \cap S_3 \cap S_4)$.

† The set that is the intersection of S_2 and S_3, denoted by $S_2 \cap S_3$, is composed of the elements that belong to *both* S_2 and S_3.

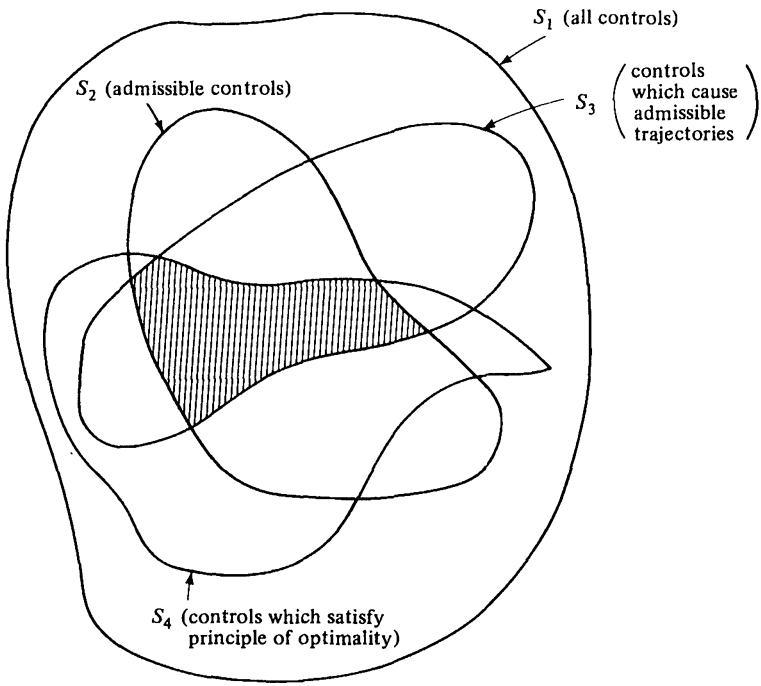


Figure 3-7 Subsets of the control space

Presence of Constraints

As shown in Fig. 3-7, the presence of constraining relations on admissible state and/or control values simplifies the numerical procedure. For example, if the control is a scalar and is constrained by the relationship

$$-1.0 \leq u(t) \leq 1.0, \tag{3.9-1}$$

then in the direct search procedure we need to try only values of u in the allowed interval instead of values of u throughout the interval

$$-\infty < u(t) < \infty. \tag{3.9-2}$$

Form of the Optimal Control

Dynamic programming yields the optimal control in closed-loop or feedback form—for every state value in the admissible region we know what the optimal control is. However, although $\mathbf{u}^*$ is obtained in the form

$$\mathbf{u}^*(t) = \mathbf{f}(\mathbf{x}(t), t), \tag{3.9-3}$$

unfortunately the computational procedure does not yield a nice analytical expression for **f**. It may be possible to approximate **f** in some fashion, but if this cannot be done, the optimal control law must be implemented by extracting the control values from a storage device that contains the solution of Eq. (3.8-3) in tabular form.

A Comparison of Dynamic Programming and Direct Enumeration

Dynamic programming uses the principle of optimality to reduce dramatically the number of calculations required to determine the optimal control law. In order to appreciate more fully the importance of the principle of optimality, let us compare the dynamic programming algorithm with direct enumeration of all possible control sequences.

Consider a first-order control process with one control input. Assume that the admissible state values are quantized into 10 levels, and the admissible control values into four levels. In direct enumeration we try all of the four control values at each of the 10 initial state values for one time increment Δt. In general, this will allow $x(\Delta t)$ to assume any of 40 admissible state values. Assuming that all of these state values are admissible, we apply all four control values at each of the 40 state values and determine the resulting values of $x(2\,\Delta t)$. This procedure continues for the appropriate number of

Table 3-5 AN EXAMPLE COMPARISON OF DYNAMIC PROGRAMMING AND DIRECT ENUMERATION

Number of stages in the process N	*Number of calculations required by dynamic programming*	*Number of calculations required by direct enumeration*	*Number of calculations required by direct enumeration (assuming 50% of state values admissible and distinct)*
1	40	40	40
2	80	200	120
3	120	840	280
4	160	3,400	600
5	200	13,640	1,240
6	240	54,600	2,520
L	$40L$	$\sum_{k=1}^{L} [10 \cdot 4^k]$	$\sum_{k=1}^{L} [20 \cdot 2^k]$

stages. In dynamic programming, at every stage we try four control values at each of 10 state values. Table 3-5 shows a comparison of the number of calculations required by the two methods. The table also includes the number of calculations required for direct enumeration if it is assumed that at the end of each stage only half of the state values are distinct and admissible. The important point is that the number of calculations required by direct enumeration increases exponentially with the number of stages, while the computational requirements of dynamic programming increase linearly.

The Curse of Dimensionality

From the preceding discussion it may seem that perhaps dynamic programming is the answer to all of our problems; unfortunately, there is one serious drawback: for high-dimensional systems the number of high-speed storage locations becomes prohibitive. Bellman calls this difficulty the "curse of dimensionality." To appreciate the nature of the problem, recall that to evaluate $J^*_{N-K,N}$ we need access to the values of $J^*_{N-(K-1),N}$ which have been previously computed. For a third-order system with 100 quantization levels in each state coordinate direction, this means that $10^2 \times 10^2 \times 10^2 = 10^6$ storage locations are required; this number approaches the limit of rapid-access storage available with current computers. There is nothing to prevent us from using low-speed storage; however, this will drastically increase computation time. Of the techniques that have been developed to alleviate the curse of dimensionality, Larson's "state increment dynamic programming" [L-1] seems to be the most promising. There are other methods, however, several of which are explained in [N-1]. [L-2] contains an excellent survey of computational procedures used in dynamic programming.

3.10 ANALYTICAL RESULTS—DISCRETE LINEAR REGULATOR PROBLEMS

In this section we consider the discrete system described by the state equation

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k). \tag{3.10-1}$$

The states and controls are not constrained by any boundaries. The problem is to find an optimal policy $\mathbf{u}^*(\mathbf{x}(k), k)$ that minimizes the performance measure

$$J = \tfrac{1}{2}\mathbf{x}^T(N)\mathbf{H}\mathbf{x}(N) + \tfrac{1}{2}\sum_{k=0}^{N-1}\left[\mathbf{x}^T(k)\mathbf{Q}(k)\mathbf{x}(k) + \mathbf{u}^T(k)\mathbf{R}(k)\mathbf{u}(k)\right], \tag{3.10-2}$$

where

$\mathbf{H}$ and $\mathbf{Q}(k)$ are real symmetric positive semi-definite $n \times n$ matrices.
$\mathbf{R}(k)$ is a real symmetric positive definite $m \times m$ matrix.
N is a fixed integer greater than 0.

The above problem is the discrete counterpart of the continuous linear regulator problem considered in Sections 3.12 and 5.2.† To simplify the notation in the derivation that follows, let us make the assumption that $\mathbf{A}$, $\mathbf{B}$, $\mathbf{R}$, and $\mathbf{Q}$ are constant matrices. The approach we will take is to solve the functional equation (3.7-18). We begin by defining

$$J_{NN}(\mathbf{x}(N)) = \tfrac{1}{2}\mathbf{x}^T(N)\mathbf{H}\mathbf{x}(N) = J^*_{NN}(\mathbf{x}(N)) \triangleq \tfrac{1}{2}\mathbf{x}^T(N)\mathbf{P}(0)\mathbf{x}(N) \tag{3.10-3}$$

where $\mathbf{P}(0) \triangleq \mathbf{H}$. The cost over the final interval is given by

$$\begin{aligned} J_{N-1,N}(\mathbf{x}(N-1), \mathbf{u}(N-1)) &= \tfrac{1}{2}\mathbf{x}^T(N-1)\mathbf{Q}\mathbf{x}(N-1) \\ &+ \tfrac{1}{2}\mathbf{u}^T(N-1)\mathbf{R}\mathbf{u}(N-1) + \tfrac{1}{2}\mathbf{x}^T(N)\mathbf{P}(0)\mathbf{x}(N), \end{aligned} \tag{3.10-4}$$

and the minimum cost is

$$J^*_{N-1,N}(\mathbf{x}(N-1)) \triangleq \min_{\mathbf{u}(N-1)} \{J_{N-1,N}(\mathbf{x}(N-1), \mathbf{u}(N-1))\}. \tag{3.10-5}$$

Now $\mathbf{x}(N)$ is related to $\mathbf{u}(N-1)$ by the state equation, so

$$\begin{aligned} J^*_{N-1,N}(\mathbf{x}(N-1)) = \min_{\mathbf{u}(N-1)} \{&\tfrac{1}{2}\mathbf{x}^T(N-1)\mathbf{Q}\mathbf{x}(N-1) + \tfrac{1}{2}\mathbf{u}^T(N-1)\mathbf{R}\mathbf{u}(N-1) \\ &+ \tfrac{1}{2}[\mathbf{A}\mathbf{x}(N-1) + \mathbf{B}\mathbf{u}(N-1)]^T\mathbf{P}(0)[\mathbf{A}\mathbf{x}(N-1) + \mathbf{B}\mathbf{u}(N-1)]\}. \end{aligned} \tag{3.10-6}$$

It is assumed that the admissible controls are not bounded; therefore, to minimize $J_{N-1,N}$ with respect to $\mathbf{u}(N-1)$ we need to consider only those control values for which

$$\begin{bmatrix} \dfrac{\partial J_{N-1,N}}{\partial u_1(N-1)} \\ \dfrac{\partial J_{N-1,N}}{\partial u_2(N-1)} \\ \vdots \\ \dfrac{\partial J_{N-1,N}}{\partial u_m(N-1)} \end{bmatrix} \triangleq \frac{\partial J_{N-1,N}}{\partial \mathbf{u}(N-1)} = \mathbf{0}. \tag{3.10-7}$$

† Equations (3.10-1) and (3.10-2) may be the result of a discrete approximation to a continuous problem, or the formulation for a linear, sampled-data system (see Appendix 2).

Evaluating the indicated partial derivatives gives

$$\mathbf{Ru}(N-1) + \mathbf{B}^T\mathbf{P}(0)[\mathbf{Ax}(N-1) + \mathbf{Bu}(N-1)] = \mathbf{0}.\dagger \quad (3.10\text{-}8)$$

The control values that satisfy this equation may yield a minimum of $J_{N-1,N}$, a maximum, or neither. To investigate further, we form the matrix of second partials given by

$$\begin{bmatrix} \frac{\partial^2 J_{N-1,N}}{\partial u_1^2(N-1)} & \frac{\partial^2 J_{N-1,N}}{\partial u_1(N-1)\partial u_2(N-1)} & \cdots & \frac{\partial^2 J_{N-1,N}}{\partial u_1(N-1)\partial u_m(N-1)} \\ \frac{\partial^2 J_{N-1,N}}{\partial u_2(N-1)\partial u_1(N-1)} & \frac{\partial^2 J_{N-1,N}}{\partial u_2^2(N-1)} & \cdots & \frac{\partial^2 J_{N-1,N}}{\partial u_2(N-1)\partial u_m(N-1)} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 J_{N-1,N}}{\partial u_m(N-1)\partial u_1(N-1)} & \frac{\partial^2 J_{N-1,N}}{\partial u_m(N-1)\partial u_2(N-1)} & \cdots & \frac{\partial^2 J_{N-1,N}}{\partial u_m^2(N-1)} \end{bmatrix}$$

$$\triangleq \frac{\partial^2 J_{N-1,N}}{\partial \mathbf{u}^2(N-1)} = \mathbf{R} + \mathbf{B}^T\mathbf{P}(0)\mathbf{B}. \quad (3.10\text{-}9)$$

By assumption $\mathbf{H}$ [and hence $\mathbf{P}(0)$] is a positive semi-definite matrix, and $\mathbf{R}$ is a positive definite matrix. It can be shown that since $\mathbf{P}(0)$ is positive semi-definite, so is $\mathbf{B}^T\mathbf{P}(0)\mathbf{B}$. This means that $\mathbf{R} + \mathbf{B}^T\mathbf{P}(0)\mathbf{B}$ is the sum of a positive definite matrix and a positive semi-definite matrix, and this implies that $\mathbf{R} + \mathbf{B}^T\mathbf{P}(0)\mathbf{B}$ is positive definite.‡ Since $J_{N-1,N}$ is a quadratic function of $\mathbf{u}(N-1)$ and the matrix $\partial^2 J_{N-1,N}/\partial \mathbf{u}^2(N-1)$ is positive definite, the control that satisfies Eq. (3.10-8) yields the absolute, or global, minimum of $J_{N-1,N}$.

Solving (3.10-8) for the optimal control gives

$$\mathbf{u}^*(N-1) = -[\mathbf{R} + \mathbf{B}^T\mathbf{P}(0)\mathbf{B}]^{-1}\mathbf{B}^T\mathbf{P}(0)\mathbf{Ax}(N-1) \triangleq \mathbf{F}(N-1)\mathbf{x}(N-1). \quad (3.10\text{-}10)$$

Since $\mathbf{R} + \mathbf{B}^T\mathbf{P}(0)\mathbf{B}$ is positive definite, the indicated inverse is guaranteed to exist. Substituting the expression for $\mathbf{u}^*(N-1)$ into the equation for $J_{N-1,N}$ gives $J^*_{N-1,N}$, which after terms have been collected becomes

$$J^*_{N-1,N}(\mathbf{x}(N-1)) = \tfrac{1}{2}\mathbf{x}^T(N-1)\{[\mathbf{A} + \mathbf{BF}(N-1)]^T\mathbf{P}(0)[\mathbf{A} + \mathbf{BF}(N-1)] + \mathbf{F}^T(N-1)\mathbf{RF}(N-1) + \mathbf{Q}\}\mathbf{x}(N-1) \triangleq \tfrac{1}{2}\mathbf{x}^T(N-1)\mathbf{P}(1)\mathbf{x}(N-1). \quad (3.10\text{-}11)$$

† The symmetry of $\mathbf{R}$ and $\mathbf{P}(0)$ have also been used here. The reader will find the matrix calculus relationships given in Appendix 1 helpful in following the steps of this derivation.

‡ See Appendix 1.

The definition for $\mathbf{P}(1)$ is clear, by inspection of (3.10-11). The important point is that $J^*_{N-1,N}$ *is of exactly the same form as* $J^*_{N,N}$, which means that when we continue the process one stage further back, the results will have exactly the same form; i.e.,

$$\begin{aligned}\mathbf{u}^*(N-2) &= -[\mathbf{R}+\mathbf{B}^T\mathbf{P}(1)\mathbf{B}]^{-1}\mathbf{B}^T\mathbf{P}(1)\mathbf{A}\mathbf{x}(N-2)\\ &\triangleq \mathbf{F}(N-2)\mathbf{x}(N-2),\end{aligned} \tag{3.10-12}$$

and

$$\begin{aligned}J^*_{N-2,N}(\mathbf{x}(N-2)) &= \tfrac{1}{2}\mathbf{x}^T(N-2)\{[\mathbf{A}+\mathbf{B}\mathbf{F}(N-2)]^T\mathbf{P}(1)[\mathbf{A}+\mathbf{B}\mathbf{F}(N-2)]\\ &\quad +\mathbf{F}^T(N-2)\mathbf{R}\mathbf{F}(N-2)+\mathbf{Q}\}\mathbf{x}(N-2)\\ &\triangleq \tfrac{1}{2}\mathbf{x}^T(N-2)\mathbf{P}(2)\mathbf{x}(N-2).\end{aligned} \tag{3.10-13}$$

If you do not believe this, try it and see.

By induction, for the Kth stage

$$\begin{aligned}\mathbf{u}^*(N-K) &= -[\mathbf{R}+\mathbf{B}^T\mathbf{P}(K-1)\mathbf{B}]^{-1}\mathbf{B}^T\mathbf{P}(K-1)\mathbf{A}\mathbf{x}(N-K)\\ &\triangleq \mathbf{F}(N-K)\mathbf{x}(N-K)\end{aligned} \tag{3.10-14}$$

and

$$\begin{aligned}J^*_{N-K,N}(\mathbf{x}(N-K)) &= \tfrac{1}{2}\mathbf{x}^T(N-K)\{[\mathbf{A}+\mathbf{B}\mathbf{F}(N-K)]^T\mathbf{P}(K-1)[\mathbf{A}+\mathbf{B}\mathbf{F}(N-K)]\\ &\quad +\mathbf{F}^T(N-K)\mathbf{R}\mathbf{F}(N-K)+\mathbf{Q}\}\mathbf{x}(N-K)\\ &\triangleq \tfrac{1}{2}\mathbf{x}^T(N-K)\mathbf{P}(K)\mathbf{x}(N-K).\end{aligned} \tag{3.10-15}$$

In the general time-varying case the same derivation gives

$$\begin{aligned}\mathbf{u}^*(N-K) &= -[\mathbf{R}(N-K)+\mathbf{B}^T(N-K)\mathbf{P}(K-1)\mathbf{B}(N-K)]^{-1}\\ &\quad \times\mathbf{B}^T(N-K)\mathbf{P}(K-1)\mathbf{A}(N-K)\mathbf{x}(N-K)\\ &\triangleq \mathbf{F}(N-K)\mathbf{x}(N-K)\end{aligned} \tag{3.10-16}$$

$$\begin{aligned}J^*_{N-K,N}(\mathbf{x}(N-K)) &= \tfrac{1}{2}\mathbf{x}^T(N-K)\{[\mathbf{A}(N-K)\\ &\quad +\mathbf{B}(N-K)\mathbf{F}(N-K)]^T\\ &\quad \times\mathbf{P}(K-1)[\mathbf{A}(N-K)+\mathbf{B}(N-K)\mathbf{F}(N-K)]\\ &\quad +\mathbf{F}^T(N-K)\mathbf{R}(N-K)\mathbf{F}(N-K)\\ &\quad +\mathbf{Q}(N-K)\}\mathbf{x}(N-K)\\ &\triangleq \tfrac{1}{2}\mathbf{x}^T(N-K)\mathbf{P}(K)\mathbf{x}(N-K).\end{aligned} \tag{3.10-17}$$

What are the implications of these results? First, and most important, observe that *the optimal control at each stage is a linear combination of the states;* therefore, the optimal policy is linear state-variable feedback. Notice that the feedback is time-varying, even if **A**, **B**, **R**, and **Q** are *all* constant matrices—this means that the controller for the optimal policy can be implemented by the m time-varying amplifier-summers each with n inputs shown in Fig. 3-8. At the conclusion of Section 3.8 we remarked, "... the optimal controller is physically realized by a table look-up device and a generator of piecewise-constant signals"; when the system is linear and the performance measure quadratic in the states and controls, the only table

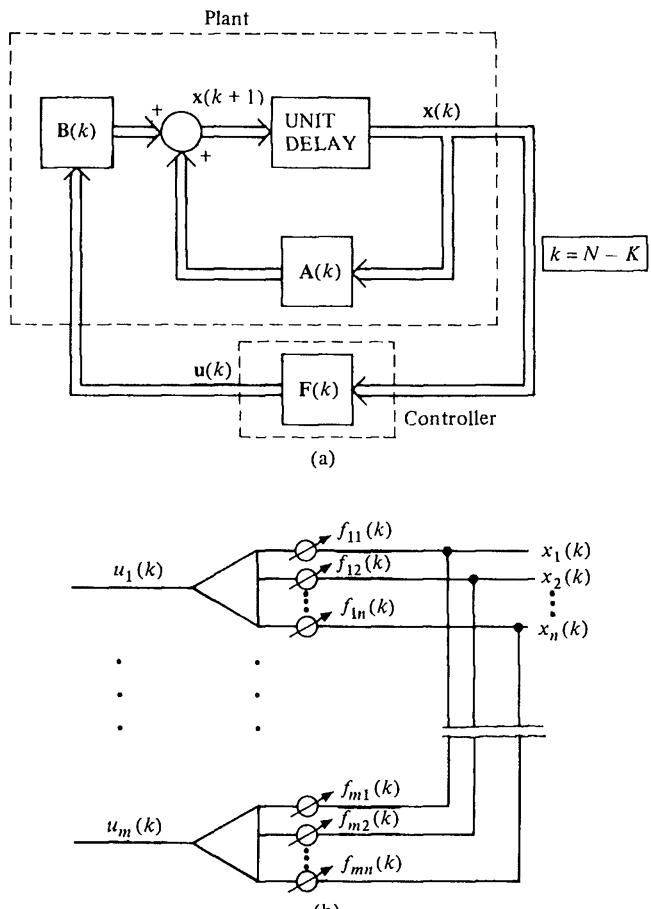


Figure 3-8 (a) Plant and linear time-varying feedback controller (b) Controller configuration

look-up involved in the controller is to determine the appropriate gain settings from stage to stage.

Another important result of the derivation is that the minimum cost for an N-stage process with initial state $\mathbf{x}_0$ is given by

$$\boxed{J^*_{0,N}(\mathbf{x}_0) = \tfrac{1}{2}\mathbf{x}_0^T\mathbf{P}(N)\mathbf{x}_0,} \tag{3.10-18}$$

which follows directly from the definition of $\mathbf{P}(N-K)$. This means that storage of the $\mathbf{P}(N-K)$ matrices for $K = 1, 2, \ldots, N$ provides us with a means of determining the minimum costs for processes of from 1 to N stages.

The computational implications of these results are also important. In order to evaluate the feedback gains and the minimum cost for any initial state, it is necessary only to solve the equations

$$\begin{aligned}\mathbf{F}(N-K) = -&[\mathbf{R}(N-K) + \mathbf{B}^T(N-K)\mathbf{P}(K-1)\mathbf{B}(N-K)]^{-1} \\ &\times \mathbf{B}^T(N-K)\mathbf{P}(K-1)\mathbf{A}(N-K)\end{aligned} \tag{3.10-19}$$

and

$$\begin{aligned}\mathbf{P}(K) = &[\mathbf{A}(N-K) + \mathbf{B}(N-K)\mathbf{F}(N-K)]^T\mathbf{P}(K-1) \\ &\times [\mathbf{A}(N-K) + \mathbf{B}(N-K)\mathbf{F}(N-K)] \\ &+ \mathbf{F}^T(N-K)\mathbf{R}(N-K)\mathbf{F}(N-K) + \mathbf{Q}(N-K)\end{aligned} \tag{3.10-20}$$

with $\mathbf{P}(0) = \mathbf{H}$. We obtain the solution by evaluating $\mathbf{F}(N-1)$ using $\mathbf{P}(0) = \mathbf{H}$, and then substituting $\mathbf{F}(N-1)$ in (3.10-20) to determine $\mathbf{P}(1)$. This constitutes one cycle of the procedure, which we then continue by calculating $\mathbf{F}(N-2)$, $\mathbf{P}(2)$, and so on. The solution is best done by a digital computer; for a reduction in the number of arithmetic operations, it is helpful to define

$$\mathbf{V}(N-K) \triangleq \mathbf{A}(N-K) + \mathbf{B}(N-K)\mathbf{F}(N-K) \tag{3.10-21}$$

so that the procedure is to solve (3.10-19), then (3.10-21), and finally the equation

$$\begin{aligned}\mathbf{P}(K) = &\mathbf{V}^T(N-K)\mathbf{P}(K-1)\mathbf{V}(N-K) \\ &+ \mathbf{F}^T(N-K)\mathbf{R}(N-K)\mathbf{F}(N-K) + \mathbf{Q}(N-K).\end{aligned} \tag{3.10-20a}$$

The $\mathbf{F}$ and $\mathbf{P}$ matrices are printed for use in synthesizing optimal controls and determining minimum costs.

It is important to realize that the solution of these equations is equivalent to the computational procedure outlined in Section 3.8; however, because

of the *linear plant dynamics and quadratic performance measure* we obtain the closed-form results given in Eqs. (3.10-16) through (3.10-20a).

The reader may have noticed that the control problem of Section 3.5 is of the linear regulator type. Why then are not the optimal controls in the rightmost columns of Tables 3-2 and 3-3 linear functions of the state values? The answer is that the quantized grid of points is very coarse, causing numerical inaccuracies. When the quantization increments are made much smaller, the linear relationship between the optimal control and state values is apparent; this effect is illustrated in Problems 3-14 through 3-17 at the end of the chapter.

Another important characteristic of the linear regulator problem is that if the system (3.10-1) is completely controllable† and time-invariant, $\mathbf{H} = \mathbf{0}$, and $\mathbf{R}$ and $\mathbf{Q}$ are constant matrices, then the optimal control law is time-invariant for an infinite-stage process; that is

$$\mathbf{F}(N - K) \longrightarrow \mathbf{F} \text{ (a constant matrix)} \quad \text{as} \quad N \longrightarrow \infty.$$

From a physical point of view this means that if a process is to be controlled for a large number of stages the optimal control can be implemented by feedback of the states through a configuration of amplifier-summers as shown in Fig. 3-8(b), but with *fixed* gain factors. One way of determining the constant $\mathbf{F}$ matrix is to solve the recurrence relations for as many stages as required for $\mathbf{F}(N - K)$ to converge to a constant matrix.

Let us now conclude our consideration of the discrete linear regulator problem with the following example.

Example 3.10-1. The linear discrete system

$$\mathbf{x}(k+1) = \begin{bmatrix} 0.9974 & 0.0539 \\ -0.1078 & 1.1591 \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} 0.0013 \\ 0.0539 \end{bmatrix} u(k) \tag{3.10-22}$$

is to be controlled to minimize the performance measure

$$J = \tfrac{1}{2} \sum_{k=0}^{N-1} [0.25x_1^2(k) + 0.05x_2^2(k) + 0.05u^2(k)]. \tag{3.10-23}$$

Determine the optimal control law.

Equations (3.10-19), (3.10-21), and (3.10-20a) are most easily solved by using a digital computer with $\mathbf{A}$ and $\mathbf{B}$ as specified in Eq. (3.10-22),

† The discrete system of Eq. (3.10-1) with $\mathbf{A}$ and $\mathbf{B}$ constant matrices is completely controllable if and only if the $n \times mn$ matrix

$$[\mathbf{B} \vdots \mathbf{AB} \vdots \cdots \vdots \mathbf{A}^{n-1}\mathbf{B}]$$

is of rank n. For a proof of this theorem, see [P-2].

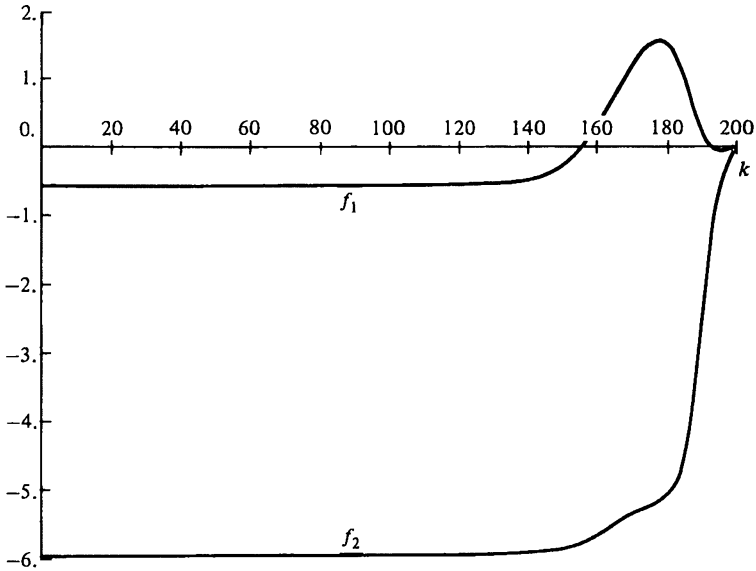


Figure 3-9(a) Feedback gain coefficients for optimal control of a second-order discrete linear regulator

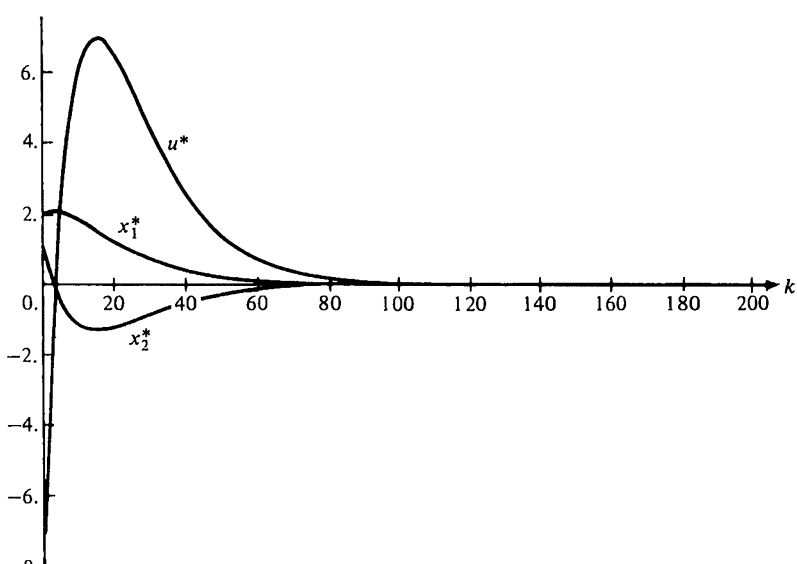


Figure 3-9(b) Optimal control and trajectory for a second-order discrete linear regulator

$$\mathbf{H} = \mathbf{0}, \qquad \mathbf{Q} = \begin{bmatrix} 0.25 & 0.00 \\ 0.00 & 0.05 \end{bmatrix}, \quad \text{and} \quad R = 0.05.$$

The optimal feedback gain matrix $\mathbf{F}(k)$ is shown in Fig. 3-9(a) for $N = 200$. Looking backward from $k = 199$, we observe that at $k \approx 130$ the $\mathbf{F}(k)$ matrix has reached the steady-state value

$$\mathbf{F}(k) = [-0.5522 \quad -5.9668], \qquad 0 \lesssim k \lesssim 130. \tag{3.10-24}$$

The optimal control history and the optimal trajectory for $\mathbf{x}(0) = [2 \quad 1]^T$ are shown in Fig. 3-9(b). Notice that the optimal trajectory has essentially reached $\mathbf{0}$ at $k = 100$. Thus, we would expect that insignificant performance degradation would be caused by simply using the steady-state value of $\mathbf{F}$ given in (3.10-24) rather than $\mathbf{F}(k)$ as specified in Fig. 3-9(a).

3.11 THE HAMILTON-JACOBI-BELLMAN EQUATION

In our initial exposure to dynamic programming, we approximated continuously operating systems by discrete systems. This approach leads to a recurrence relation that is ideally suited for digital computer solution. In this section we shall consider an alternative approach which leads to a nonlinear *partial* differential equation—the Hamilton-Jacobi-Bellman (H-J-B) equation. The derivation that will be given in this section parallels the development of the functional recurrence equation (3.7-18) in Section 3.7.

The process described by the state equation

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{3.11-1}$$

is to be controlled to minimize the performance measure

$$J = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau)\, d\tau, \tag{3.11-2}$$

where h and g are specified functions, t_0 and t_f are fixed, and τ is a dummy variable of integration. Let us now use the *imbedding principle* to include this problem in a larger class of problems by considering the performance measure

$$J(\mathbf{x}(t), t, \underset{t \le \tau \le t_f}{\mathbf{u}(\tau)}) = h(\mathbf{x}(t_f), t_f) + \int_{t}^{t_f} g(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau)\, d\tau, \tag{3.11-3}$$

where t can be any value less than or equal to t_f, and $\mathbf{x}(t)$ can be any admissible state value. Notice that the performance measure will depend on the

numerical values for $\mathbf{x}(t)$ and t, and on the optimal control history in the interval $[t, t_f]$.

Let us now attempt to determine the controls that minimize (3.11-3) for all admissible $\mathbf{x}(t)$, and for all $t \leq t_f$. The minimum cost function is then

$$J^*(\mathbf{x}(t), t) = \min_{\substack{\mathbf{u}(\tau) \\ t \leq \tau \leq t_f}} \left\{ \int_t^{t_f} g(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau)\, d\tau + h(\mathbf{x}(t_f), t_f) \right\}. \quad (3.11\text{-}4)$$

By subdividing the interval, we obtain

$$J^*(\mathbf{x}(t), t) = \min_{\substack{\mathbf{u}(\tau) \\ t \leq \tau \leq t_f}} \left\{ \int_t^{t+\Delta t} g\, d\tau + \int_{t+\Delta t}^{t_f} g\, d\tau + h(\mathbf{x}(t_f), t_f) \right\}. \quad (3.11\text{-}5)$$

The principle of optimality requires that

$$J^*(\mathbf{x}(t), t) = \min_{\substack{\mathbf{u}(\tau) \\ t \leq \tau \leq t+\Delta t}} \left\{ \int_t^{t+\Delta t} g\, d\tau + J^*(\mathbf{x}(t + \Delta t), t + \Delta t) \right\}, \quad (3.11\text{-}6)$$

where $J^*(\mathbf{x}(t + \Delta t), t + \Delta t)$ is the minimum cost of the process for the time interval $t + \Delta t \leq \tau \leq t_f$ with "initial" state $\mathbf{x}(t + \Delta t)$.

Assuming that the second partial derivatives of J^* exist and are bounded, we can expand $J^*(\mathbf{x}(t + \Delta t), t + \Delta t)$ in a Taylor series about the point $(\mathbf{x}(t), t)$ to obtain

$$\begin{aligned} J^*(\mathbf{x}(t), t) = \min_{\substack{\mathbf{u}(\tau) \\ t \leq \tau \leq t+\Delta t}} \Bigg\{ & \int_t^{t+\Delta t} g\, d\tau + J^*(\mathbf{x}(t), t) + \left[\frac{\partial J^*}{\partial t}(\mathbf{x}(t), t)\right] \Delta t \\ & + \left[\frac{\partial J^*}{\partial \mathbf{x}}(\mathbf{x}(t), t)\right]^T \left[\mathbf{x}(t + \Delta t) - \mathbf{x}(t)\right] \\ & + \text{terms of higher order} \Bigg\}. \end{aligned} \quad (3.11\text{-}7)$$

Now for small Δt

$$\begin{aligned} J^*(\mathbf{x}(t), t) = \min_{\mathbf{u}(t)} \{ & g(\mathbf{x}(t), \mathbf{u}(t), t)\, \Delta t + J^*(\mathbf{x}(t), t) \\ & + J_t^*(\mathbf{x}(t), t)\, \Delta t + J_{\mathbf{x}}^{*T}(\mathbf{x}(t), t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)]\, \Delta t \\ & + o(\Delta t) \}, \dagger \end{aligned} \quad (3.11\text{-}8)$$

where $o(\Delta t)$ denotes the terms containing $[\Delta t]^2$ and higher orders of Δt that arise from the approximation of the integral and the truncation of the Taylor series expansion. Next, removing the terms involving $J^*(\mathbf{x}(t), t)$ and $J_t^*(\mathbf{x}(t), t)$

† $J_{\mathbf{x}}^* \triangleq \dfrac{\partial J^*}{\partial \mathbf{x}} = \left[\dfrac{\partial J^*}{\partial x_1} \quad \dfrac{\partial J^*}{\partial x_2} \quad \cdots \quad \dfrac{\partial J^*}{\partial x_n}\right]^T$ and $J_t^* \triangleq \dfrac{\partial J^*}{\partial t}$.

from the minimization [since they do not depend on $\mathbf{u}(t)$], we obtain

$$0 = J_t^*(\mathbf{x}(t), t)\,\Delta t + \min_{\mathbf{u}(t)} \{g(\mathbf{x}(t), \mathbf{u}(t), t)\,\Delta t + J_{\mathbf{x}}^{*T}(\mathbf{x}(t), t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)]\,\Delta t + o(\Delta t)\}. \tag{3.11-9}$$

Dividing by Δt and taking the limit as $\Delta t \to 0$ gives†

$$\boxed{0 = J_t^*(\mathbf{x}(t), t) + \min_{\mathbf{u}(t)} \{g(\mathbf{x}(t), \mathbf{u}(t), t) + J_{\mathbf{x}}^{*T}(\mathbf{x}(t), t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)]\}.} \tag{3.11-10}$$

To find the boundary value for this partial differential equation, set $t = t_f$; from Eq. (3.11-4) it is apparent that

$$\boxed{J^*(\mathbf{x}(t_f), t_f) = h(\mathbf{x}(t_f), t_f).} \tag{3.11-11}$$

We define the Hamiltonian $\mathscr{H}$ as

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), J_{\mathbf{x}}^*, t) \triangleq g(\mathbf{x}(t), \mathbf{u}(t), t) + J_{\mathbf{x}}^{*T}(\mathbf{x}(t), t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)] \tag{3.11-12}$$

and

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}^*(\mathbf{x}(t), J_{\mathbf{x}}^*, t), J_{\mathbf{x}}^*, t) = \min_{\mathbf{u}(t)} \mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), J_{\mathbf{x}}^*, t), \tag{3.11-13}$$

since the minimizing control will depend on $\mathbf{x}$, $J_{\mathbf{x}}^*$, and t. Using these definitions, we have obtained the Hamilton-Jacobi equation

$$\boxed{0 = J_t^*(\mathbf{x}(t), t) + \mathscr{H}(\mathbf{x}(t), \mathbf{u}^*(\mathbf{x}(t), J_{\mathbf{x}}^*, t), J_{\mathbf{x}}^*, t).} \tag{3.11-10a}$$

This equation is the continuous-time analog of Bellman's recurrence equation (3.7-18); therefore, we shall refer to (3.11-10a) as the "Hamilton-Jacobi-Bellman equation."

Example 3.11-1. A first-order system is described by the differential equation

$$\dot{x}(t) = x(t) + u(t); \tag{3.11-14}$$

† $\lim_{\Delta t \to 0} \left| \frac{o(\Delta t)}{\Delta t} \right| = 0.$

it is desired to find the control law that minimizes the performance measure

$$J = \tfrac{1}{4}x^2(T) + \int_0^T \tfrac{1}{4}u^2(t)\,dt. \tag{3.11-15}$$

The final time T is specified, and the admissible state and control values are not constrained by any boundaries.

Substituting $g = \frac{1}{4}u^2(t)$ and $a = x(t) + u(t)$ into Eq. (3.11-12), we find that the Hamiltonian is (omitting the arguments of J_x^*)

$$\mathscr{H}(x(t), u(t), J_x^*, t) = \tfrac{1}{4}u^2(t) + J_x^*[x(t) + u(t)], \tag{3.11-16}$$

and since the control is unconstrained, a necessary condition that the optimal control must satisfy is

$$\frac{\partial \mathscr{H}}{\partial u} = \frac{1}{2}u(t) + J_x^*(x(t), t) = 0. \tag{3.11-17}$$

Observe that

$$\frac{\partial^2 \mathscr{H}}{\partial u^2} = \frac{1}{2} > 0; \tag{3.11-18}$$

thus, the control that satisfies Eq. (3.11-17) does minimize $\mathscr{H}$. From (3.11-17)

$$u^*(t) = -2J_x^*(x(t), t), \tag{3.11-19}$$

which when substituted in the Hamilton-Jacobi-Bellman equation gives

$$\begin{aligned} 0 &= J_t^* + \tfrac{1}{4}[-2J_x^*]^2 + [J_x^*]x(t) - 2[J_x^*]^2 \\ &= J_t^* - [J_x^*]^2 + [J_x^*]x(t). \end{aligned} \tag{3.11-20}$$

The boundary value is, from (3.11-15),

$$J^*(x(T), T) = \tfrac{1}{4}x^2(T). \tag{3.11-21}$$

One way to solve the Hamilton-Jacobi-Bellman equation is to guess a form for the solution and see if it can be made to satisfy the differential equation and the boundary conditions. Let us assume a solution of the form

$$J^*(x(t), t) = \tfrac{1}{2}K(t)x^2(t), \tag{3.11-22}$$

where $K(t)$ represents an unknown scalar function of t that is to be determined. Notice that

$$J_x^*(x(t), t) = K(t)x(t), \tag{3.11-23}$$

which, together with Eq. (3.11-19), implies that

$$u^*(t) = -2K(t)x(t). \tag{3.11-24}$$

Thus, if a function $K(t)$ can be found such that (3.11-20) and (3.11-21) are satisfied, the optimal control is *linear* feedback of the state—indeed, this was the motivation for selecting the form (3.11-22).

By making $K(T) = \frac{1}{2}$, the assumed solution matches the boundary condition specified by Eq. (3.11-21).

Substituting (3.11-23) for J_x^* and

$$J_t^*(x(t), t) = \tfrac{1}{2}\dot{K}(t)x^2(t)$$

into Eq. (3.11-20) gives

$$0 = \tfrac{1}{2}\dot{K}(t)x^2(t) - K^2(t)x^2(t) + K(t)x^2(t). \tag{3.11-25}$$

Since this equation must be satisfied for all $x(t)$,

$$\tfrac{1}{2}\dot{K}(t) - K^2(t) + K(t) = 0. \tag{3.11-26}$$

$K(t)$ is a scalar function of t; therefore, the solution can be obtained by separation of variables with the result

$$K(t) = \frac{\epsilon^{(T-t)}}{\epsilon^{(T-t)} + \epsilon^{-(T-t)}}. \tag{3.11-27}$$

The optimal control law is then

$$\begin{aligned} u^*(t) &= -2J_x^*(x(t), t) \\ &= -2K(t)x(t). \end{aligned} \tag{3.11-28}$$

Notice that as $T \longrightarrow \infty$, the linear time-varying feedback approaches constant feedback ($K(t) \longrightarrow 1$), and that the controlled system

$$\begin{aligned} \dot{x}(t) &= x(t) - 2x(t) \\ &= -x(t) \end{aligned} \tag{3.11-29}$$

is stable. If this were not the case, the performance measure would be infinite.

3.12 CONTINUOUS LINEAR REGULATOR PROBLEMS

Problems like Example 3.11-1 with linear plant dynamics and quadratic performance criteria are referred to as linear regulator problems. In this section we investigate the use of the Hamilton-Jacobi-Bellman equation as

a means of solving the general form of the continuous linear regulator problem.†

The process to be controlled is described by the state equations

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \tag{3.12-1}$$

and the performance measure to be minimized is

$$J = \tfrac{1}{2}\mathbf{x}^T(t_f)\mathbf{H}\mathbf{x}(t_f) + \int_{t_0}^{t_f} \tfrac{1}{2}\left[\mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t)\right] dt. \tag{3.12-2}$$

$\mathbf{H}$ and $\mathbf{Q}$ are real symmetric positive semi-definite matrices, $\mathbf{R}$ is a real, symmetric positive definite matrix, the initial time t_0 and the final time t_f are specified, and $\mathbf{u}(t)$ and $\mathbf{x}(t)$ are not constrained by any boundaries.

To use the Hamilton-Jacobi-Bellman equation, we first form the Hamiltonian:

$$\begin{aligned}\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), J_{\mathbf{x}}^*, t) = {} & \tfrac{1}{2}\mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \tfrac{1}{2}\mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t) + J_{\mathbf{x}}^{*T}(\mathbf{x}(t), t) \\ & \cdot\left[\mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t)\right].\end{aligned} \tag{3.12-3}$$

A necessary condition for $\mathbf{u}(t)$ to minimize $\mathscr{H}$ is that $\partial\mathscr{H}/\partial\mathbf{u} = \mathbf{0}$; thus

$$\frac{\partial\mathscr{H}}{\partial\mathbf{u}}(\mathbf{x}(t), \mathbf{u}(t), J_{\mathbf{x}}^*, t) = \mathbf{R}(t)\mathbf{u}(t) + \mathbf{B}^T(t)J_{\mathbf{x}}^*(\mathbf{x}(t), t) = \mathbf{0}. \tag{3.12-4}$$

Since the matrix

$$\frac{\partial^2\mathscr{H}}{\partial\mathbf{u}^2} = \mathbf{R}(t) \tag{3.12-5}$$

is positive definite and $\mathscr{H}$ is a quadratic form in $\mathbf{u}$, the control that satisfies Eq. (3.12-4) does minimize $\mathscr{H}$ (globally). Solving Eq. (3.12-4) for $\mathbf{u}^*(t)$ gives

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)J_{\mathbf{x}}^*(\mathbf{x}(t), t), \tag{3.12-6}$$

which when substituted in (3.12-3) yields

$$\begin{aligned}\mathscr{H}(\mathbf{x}(t), \mathbf{u}^*(t), J_{\mathbf{x}}^*, t) &= \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \tfrac{1}{2}J_{\mathbf{x}}^{*T}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T J_{\mathbf{x}}^* \\ &\quad + J_{\mathbf{x}}^{*T}\mathbf{A}\mathbf{x} - J_{\mathbf{x}}^{*T}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T J_{\mathbf{x}}^* \\ &= \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \tfrac{1}{2}J_{\mathbf{x}}^{*T}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T J_{\mathbf{x}}^* + J_{\mathbf{x}}^{*T}\mathbf{A}\mathbf{x}.\ddagger\end{aligned} \tag{3.12-7}$$

The Hamilton-Jacobi-Bellman equation is

$$0 = J_t^* + \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \tfrac{1}{2}J_{\mathbf{x}}^{*T}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T J_{\mathbf{x}}^* + J_{\mathbf{x}}^{*T}\mathbf{A}\mathbf{x}. \tag{3.12-8}$$

† Refer also to Section 5.2, where this same problem is considered and the variational approach is used.

‡ Where no ambiguity exists, the arguments will be omitted.

From Eq. (3.12-2) the boundary condition is

$$J^*(\mathbf{x}(t_f), t_f) = \tfrac{1}{2}\mathbf{x}^T(t_f)\mathbf{H}\mathbf{x}(t_f). \tag{3.12-9}$$

Since we found in Section 3.10 that the minimum cost for the discrete linear regulator problem is a quadratic function of the state, it seems reasonable to guess as a solution the form

$$J^*(\mathbf{x}(t), t) = \tfrac{1}{2}\mathbf{x}^T(t)\mathbf{K}(t)\mathbf{x}(t), \tag{3.12-10}$$

where $\mathbf{K}(t)$ is a real symmetric positive-definite matrix that is to be determined. Substituting this assumed solution in Eq. (3.12-8) yields the result

$$\begin{aligned} 0 = {} & \tfrac{1}{2}\mathbf{x}^T\dot{\mathbf{K}}\mathbf{x} + \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \tfrac{1}{2}\mathbf{x}^T\mathbf{K}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{K}\mathbf{x} \\ & + \mathbf{x}^T\mathbf{K}\mathbf{A}\mathbf{x}. \end{aligned} \tag{3.12-11}$$

The matrix product $\mathbf{KA}$ appearing in the last term can be written as the sum of a symmetric part and an unsymmetric part,

$$\mathbf{KA} = \tfrac{1}{2}[\mathbf{KA} + (\mathbf{KA})^T] + \tfrac{1}{2}[\mathbf{KA} - (\mathbf{KA})^T]. \tag{3.12-12}$$

Using the matrix property $(\mathbf{CD})^T = \mathbf{D}^T\mathbf{C}^T$ and the knowledge that the transpose of a scalar equals itself, we can show that only the symmetric part of $\mathbf{KA}$ contributes anything to (3.12-11). Thus Eq. (3.12-11) can be written

$$\begin{aligned} 0 = {} & \tfrac{1}{2}\mathbf{x}^T\dot{\mathbf{K}}\mathbf{x} + \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \tfrac{1}{2}\mathbf{x}^T\mathbf{K}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{K}\mathbf{x} \\ & + \tfrac{1}{2}\mathbf{x}^T\mathbf{K}\mathbf{A}\mathbf{x} + \tfrac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{K}\mathbf{x}. \end{aligned} \tag{3.12-13}$$

This equation must hold for all $\mathbf{x}(t)$, so

$$\boxed{\begin{aligned} \mathbf{0} = {} & \dot{\mathbf{K}}(t) + \mathbf{Q}(t) - \mathbf{K}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t) \\ & + \mathbf{K}(t)\mathbf{A}(t) + \mathbf{A}^T(t)\mathbf{K}(t), \end{aligned}} \tag{3.12-14}$$

and the boundary condition is [from (3.12-9) and (3.12-10)]

$$\boxed{\mathbf{K}(t_f) = \mathbf{H}.} \tag{3.12.15}$$

Let us consider the implications of this result: first, the H-J-B partial differential equation reduces to a set of ordinary nonlinear differential equations. Second, the $\mathbf{K}(t)$ matrix can be determined by numerical integration

of Eq. (3.12-14) from $t = t_f$ to $t = t_0$ by using the boundary condition $\mathbf{K}(t_f) = \mathbf{H}$. Actually, since the $n \times n$ $\mathbf{K}(t)$ matrix is symmetric, we need to integrate only $n(n + 1)/2$ differential equations.

Once $\mathbf{K}(t)$ has been determined, the optimal control law is given by

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t)\mathbf{x}(t). \tag{3.12-16}$$

Thus, by assuming a solution of the form (3.12-10) the optimal control law is linear, time-varying state feedback. It should be pointed out, however, that other forms are possible as solutions of the Hamilton-Jacobi-Bellman equation. Reference [J-1] gives an alternative approach which, under certain conditions, leads to a nonlinear but time-invariant form for the optimal control law.

Our approach in this section leads to Eq. (3.12-14), which is a differential equation of the Riccati type, and thus is referred to as "the Riccati equation"; in Section 5.2 this same equation is developed by variational methods—in linear regulator problems all routes lead to the same destination.

3.13 THE HAMILTON-JACOBI-BELLMAN EQUATION—SOME OBSERVATIONS

We have derived the Hamilton-Jacobi-Bellman equation and used it to solve two examples of the linear regulator type. Let us now make some observations concerning the H-J-B functional equation.

Boundary Conditions

In our derivation we have assumed that t_f is fixed; however, the results still apply if t_f is free. For example, if S represents some hypersurface in the state space and t_f is defined as the first time the system's trajectory intersects S, then the boundary condition is

$$J^*(\mathbf{x}(t_f), t_f) = h(\mathbf{x}(t_f), t_f). \tag{3.13-1}$$

A Necessary Condition

The results we have obtained represent a necessary condition for optimality; that is, the minimum cost function $J^*(\mathbf{x}(t), t)$ must satisfy the Hamilton-Jacobi-Bellman equation.

A Sufficient Condition

Although we have not derived it here, it is also true that if there is a cost function $J'(\mathbf{x}(t), t)$ that satisfies the Hamilton-Jacobi-Bellman equation, then J' is the minimum cost function; i.e.,

$$J'(\mathbf{x}(t), t) = J^*(\mathbf{x}(t), t). \tag{3.13-2}$$

Rigorous proofs of the necessary and sufficient conditions embodied in the H-J-B equation are given in [K-5] and also in [A-2], which contains several examples.

Solution of the Hamilton-Jacobi-Bellman Equation

In both of the examples that we considered, a solution was obtained by guessing a form for the minimum cost function. Unfortunately, we are normally unable to find a solution so easily. In general, the H-J-B equation must be solved by numerical techniques—see [F-1], for example. Actually, a numerical solution involves some sort of a discrete approximation to the exact optimization relationship [Eq. (3.11-10)]; alternatively, by solving the recurrence relation [Eq. (3.7-18)] we obtain the exact solution to a discrete approximation of the Hamilton-Jacobi-Bellman functional equation.

Applications of the Hamilton-Jacobi-Bellman Equation

Two examples of the use of the H-J-B equation to find a solution to optimal control problems have been given; in these examples we used the necessary condition.

Alternatively, if we have in our possession a proposed solution to an optimal control problem, the sufficiency condition can be used to verify the optimality. Several examples of this type are given in [A-2]. It should be pointed out that the derivation of the sufficient condition requires that trajectories remain in certain regions in state-time space. Unfortunately, these regions are not specified in advance—they must be determined in order to use the Hamilton-Jacobi-Bellman equation.

In Chapter 7 we shall see that the Hamilton-Jacobi-Bellman equation provides us with a bridge from the dynamic programming approach to variational methods.

3.14 SUMMARY

The central theme in this chapter has been the development of dynamic programming as it applies to a class of control problems. The principle of

optimality is the cornerstone upon which the computational algorithm is built. We have seen that dynamic programming leads to a functional recurrence relation [Eq. (3.7-18)] when a continuous process is approximated by a discrete system. Alternatively, when we deal with a continuous process, the H-J-B partial differential equation results. In either case, a digital computer solution is generally required, and the curse of dimensionality rears its ugly head. In solving the recurrence equation (3.7-18) we obtain an *exact solution to a discrete approximation of the optimization equation*, whereas in performing a numerical solution to the H-J-B equation we obtain an *approximate solution to the exact optimization equation*. Both approaches lead to an optimal control law (closed-loop optimal control). In linear regulator problems we are able to obtain the optimal control law in closed form.

REFERENCES

A-2 Athans, M., and P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications*. New York: McGraw-Hill, Inc., 1966.

B-1 Bellman, R. E., and S. E. Dreyfus, *Applied Dynamic Programming*. Princeton, N.J.: Princeton University Press, 1962.

B-2 Bellman, R. E., and R. E. Kalaba, *Dynamic Programming and Modern Control Theory*. New York: Academic Press, 1965.

B-3 Bellman, R. E., *Dynamic Programming*. Princeton, N.J.: Princeton University Press, 1957.

F-1 Fox, L., *Numerical Solution of Ordinary and Partial Differential Equations*. Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1962.

J-1 Johnson, C. D., and J. E. Gibson, "Optimal Control with Quadratic Performance Index and Fixed Terminal Time," *IEEE Trans. Automatic Control* (1964), 355–360.

K-4 Kirk, D. E., "An Introduction to Dynamic Programming," *IEEE Trans. Education* (1967), 212–219.

K-5 Kalman, R. E., "The Theory of Optimal Control and the Calculus of Variations," *Mathematical Optimization Techniques*, R. E. Bellman, ed. Santa Monica, Cal.: The RAND Corporation, 1963.

L-1 Larson, R. E., "Dynamic Programming with Reduced Computational Requirements," *IEEE Trans. Automatic Control* (1965), 135–143.

L-2 Larson, R. E., "A Survey of Dynamic Programming Computational Procedures," *IEEE Trans. Automatic Control* (1967), 767–774.

N-1 Nemhauser, G. L., *Introduction to Dynamic Programming*. New York: John Wiley & Sons, Inc., 1966.

P-1 Pontryagin, L. S., V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko,

The Mathematical Theory of Optimal Processes. New York: Interscience Publishers, Inc., 1962.

P-2 Perkins, W.R., and J.B. Cruz, Jr., *Engineering of Dynamic Systems.* New York: John Wiley & Sons, Inc., 1969.

PROBLEMS

3-1. To apply (discrete) dynamic programming to a continuously operating system it is necessary to use discrete approximations to the state differential equations and the performance measure.

(a) Determine the discrete approximations to use for the system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -x_1(t) + [1 - x_1^2(t)]x_2(t) + u(t),$$

which is to be controlled to minimize the performance measure

$$J = [x_1(T) - 5]^2 + \int_0^T \{x_2^2(t) + 20[x_1(t) - 5]^2 + u^2(t)\}\, dt.$$

The final time T is 10.0; use an interval Δt equal to 0.01.

(b) What adjustments are required to apply the dynamic programming algorithm to this system because of the nonlinearity of the differential equations?

3-2. A first-order discrete system is described by the difference equation

$$x(k + 1) = -0.5x(k) + u(k).$$

The performance measure to be minimized is

$$J = \sum_{k=0}^{2} |x(k)|,$$

and the admissible states and controls are constrained by

$$-0.2 \leq x(k) \leq 0.2, \qquad k = 0, 1, 2$$
$$-0.1 \leq u(k) \leq 0.1, \qquad k = 0, 1.$$

(a) Carry out by hand the computational steps required to determine the optimal control law by using dynamic programming. Quantize both $u(k)$ and $x(k)$ in steps of 0.1 about zero, and use linear interpolation.

(b) What is the optimal control sequence for an initial state value of 0.2?

3-3. The first-order discrete system

$$x(k + 1) = 0.5x(k) + u(k)$$

is to be transferred to the origin in two stages ($x(2) = 0$) while the performance measure

$$J = \sum_{k=0}^{1} [|x(k)| + 5|u(k)|]$$

is minimized.

(a) Use the method of dynamic programming to determine the optimal control law for each of the heavily dotted points in Fig. 3-P3. Assume that the admissible control values are quantized into the levels 1, 0.5, 0, −0.5, −1.

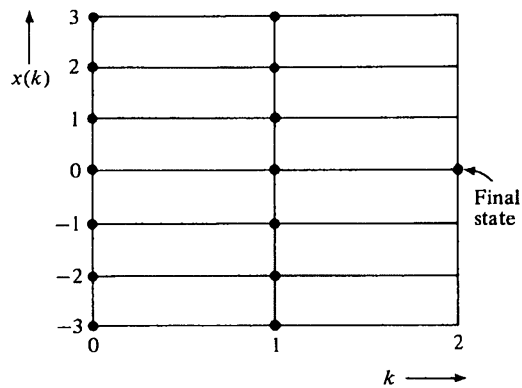


Figure 3-P3

(b) Find the optimal control sequence $\{u^*(0), u^*(1)\}$ that corresponds to the initial state $x(0) = -2$.

3-4. The discrete approximation to a nonlinear continuously operating system is given by

$$x(k+1) = x(k) - 0.4x^2(k) + u(k).$$

The state and control values are constrained by

$$0.0 \leq x(k) \leq 1.0$$
$$-0.4 \leq u(k) \leq 0.4.$$

Quantize the state into the levels 0, 0.5, 1, and the control into the levels −0.4, −0.2, 0, 0.2, 0.4. The performance measure to be minimized is

$$J = 4|x(2)| + \sum_{k=0}^{1} |u(k)|.$$

(a) Use dynamic programming with linear interpolation to complete the tables shown below.

$x(0)$	$J^*_{0,2}(x(0))$	$u^*(x(0), 0)$
0.0		
0.5		
1.0		

$x(1)$	$J^*_{1,2}(x(1))$	$u^*(x(1), 1)$
0.0		
0.5		
1.0		

(b) From the results of part (a) find the optimal control sequence $\{u^*(0), u^*(1)\}$ and the minimum cost if the initial state is 1.0.

3-5. The approximating difference equation representation for a continuously operating system is

$$x(k + 1) = 0.75x(k) + u(k).$$

It is desired to bring the system state to the target set S defined by

$$0.0 \leq x(2) \leq 2.0$$

with minimum expenditure of control effort; i.e., minimize

$$J = u^2(0) + u^2(1).$$

The allowable state and control values are constrained by

$$0.0 \leq x(k) \leq 6.0$$
$$-1.0 \leq u(k) \leq 1.0.$$

Quantize the state values into the levels $x(k) = 0, 2.0, 4.0, 6.0$ for $k = 0, 1, 2$ and the control values into the levels $u(k) = -1.0, -0.5, 0.0, 0.5, 1.0$ for $k = 0, 1$.

(a) Find the optimal control value(s) and the minimum cost for each point on the state grid. Use linear interpolation.

(b) What is the optimal control sequence $\{u^*(0), u^*(1)\}$ if $x(0) = 6.0$?

3-6. A discrete system described by the difference equation

$$x(k + 1) = x(k) + u(k)$$

is to be controlled to minimize the performance measure

$$J = \sum_{k=1}^{2} [2\,|x(k) - 0.1k^2| + |u(k - 1)|].$$

The state and control values must satisfy the constraints

$$0.0 \leq x(k) \leq 0.4, \qquad k = 0, 1, 2$$
$$-0.2 \leq u(k) \leq 0.2, \qquad k = 0, 1.$$

(a) Use the dynamic programming algorithm to determine the optimal control law $u^*(x(k), k)$. Quantize the state into the values $x(k) = 0, 0.1, 0.2, 0.3, 0.4$ $(k = 0, 1, 2)$ and the control into the values $u(k) = -0.2, -0.1, 0, 0.1, 0.2$ $(k = 0, 1)$.

(b) Determine the optimal control sequence $\{u^*(0), u^*(1)\}$ if the initial state value is $x(0) = 0.2$.

3-7. The first step in using the Hamilton-Jacobi-Bellman equation

$$0 = J_t^*(\mathbf{x}(t), t) + \min_{\mathbf{u}(t)} \{g(\mathbf{x}(t), \mathbf{u}(t), t) + J_{\mathbf{x}}^{*T}(\mathbf{x}(t), t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)]\}$$

is to determine the admissible control $\mathbf{u}^*(t)$ [in terms of $\mathbf{x}(t)$, t, and $J_{\mathbf{x}}^*$] that minimizes $\{ \cdot \}$. Find $\mathbf{u}^*(t)$—expressed as a function of $\mathbf{x}(t)$, t, and $J_{\mathbf{x}}^*$—for the system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -x_1(t) + x_2(t) + u(t),$$

and the performance measure

$$J = \int_0^T \tfrac{1}{2}[q_1 x_1^2(t) + q_2 x_2^2(t) + u^2(t)]\, dt, \qquad q_1, q_2 > 0.$$

The admissible controls are constrained by

$$|u(t)| \leq 1.$$

3-8. The first-order linear system

$$\dot{x}(t) = -10x(t) + u(t)$$

is to be controlled to minimize the performance measure

$$J = \tfrac{1}{2}x^2(0.04) + \int_0^{0.04} [\tfrac{1}{4}x^2(t) + \tfrac{1}{2}u^2(t)]\, dt.$$

The admissible state and control values are not constrained by any boundaries. Find the optimal control law by using the Hamilton-Jacobi-Bellman equation.

3-9. Assume that **A**, **B**, **R**, and **Q** may be dependent on k and derive the recurrence relations (3.10-19) and (3.10-20) for the discrete nth-order linear regulator problem with m control inputs. Appendix 1 contains some useful matrix relationships.

3-10. (a) Follow the steps in the derivation given in Section 3.10 to determine the optimal control law for the first-order system

$$x(k + 1) = Ax(k) + Bu(k).$$

To minimize the performance measure

$$J = \tfrac{1}{2}x(N)Hx(N) + \sum_{k=0}^{N-1} [[x(k) - r]Q[x(k) - r] + u(k)Ru(k)],$$

r is a specified constant, and R, $Q > 0$ are scalar weighting factors.

(b) Repeat part (a) for $N = 3$ with

$$r = r(k) \text{ (a known function of } k).$$

3-11. Consider the system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), u(t), t)$$

which is to be controlled to minimize some performance measure J. The admissible state and control values are bounded, and, in addition, the control must satisfy the total energy constraint

$$\int_{t_0}^{t_f} u^2(t)\, dt \leq M;$$

M is a specified positive number. Can this problem be solved by applying dynamic programming? Explain.

3-12. Figure 3-P12 illustrates a routing problem that is to be solved by using dynamic programming. Table 3-P12 gives the costs (elapsed time, consumed fuel, etc.) of moving between any two nodes. For example, entry ij in the matrix is the cost of going from node i directly to node j. It is desired to find the minimum-cost route between any two nodes. One way to solve this problem is to determine the cost matrices $\mathbf{C}^{(k)}$ ($k = 1, 2, 3$), where $c_{ij}^{(k)}$ denotes the minimum cost to go from node i to node j *via at most* k intermediate nodes. Table 3-P12 gives the matrix $\mathbf{C}^{(0)}$. This technique is called "approximation in policy space"† because at each stage the original problem

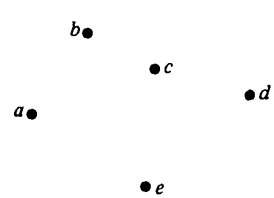


Figure 3-P12

	a	*b*	*c*	*d*	*e*
a	0	1	5	10	2
b	1	0	6	3	9
c	5	6	0	2	15
d	10	3	2	0	4
e	2	9	15	4	0

Table 3-P12

†See [B-1] and [B-2].

is solved subject to a simplifying approximation involving the allowable decisions.

(a) Determine the appropriate functional recurrence equation to find $\mathbf{C}^{(k+1)}$ from $\mathbf{C}^{(k)}$.
(b) Find the cost matrices $\mathbf{C}^{(k)}$ $(i, j = a, b, c, d, e;\ k = 1, 2, 3)$.
(c) Why is it unnecessary to determine $\mathbf{C}^{(k)}$ for $k > 3$?
(d) What are the properties of the cost matrices? In particular, compare $\mathbf{C}^{(0)}$, $\mathbf{C}^{(1)}$, $\mathbf{C}^{(2)}$, $\mathbf{C}^{(3)}$.
(e) What changes in the computational procedure are required if the costs are not independent of direction, i.e., $c_{ij}^{(0)} \neq c_{ji}^{(0)}$?

3-13. In addition to routing and control problems, dynamic programming can be advantageously applied to allocation problems. For example, suppose that a truck of capacity 11,000 lb is to transport automobiles, refrigerators, and kitchen sinks between points X and Y. The items to be transported have the weights and values shown in Table 3-P13B. The problem is to determine the number of each item that should be transported to maximize the total value of a shipment. If noninteger quantities of the items could be taken, the solution would be to carry as many as possible of the item having the highest dollar-to-weight ratio. In this case, we would take 2.75 automobiles with a shipment value of \$8250. Since this is not a feasible solution, the optimal strategy is to take

Table 3-P13a

	Value	*Weight*
2 automobiles	\$6000	8000 lb
7 refrigerators	\$1960	2800 lb
2 kitchen sinks	\$ 100	200 lb
	\$8060	11,000 lb

How was this strategy determined?

Table 3-P13b

Description	*Wt/unit*	*Value/unit*	*Value/wt*
Automobile	4000 lb	\$3000	\$0.75
Refrigerator	400 lb	\$ 280	\$0.70
Kitchen sink	100 lb	\$ 50	\$0.50

Let us now generalize this problem. Let W represent the total available resources (W = load-carrying capacity in the preceding problem). The problem is to ascertain the portion of the available resources to allocate to each of N activities in order to maximize the total return. Let

w_i = the quantity of available resources allocated to activity i.
v_i = the per-unit value of carrying out activity i.

In the shipping problem $w_1 = 8000$ lb, $w_2 = 2800$ lb, $w_3 = 200$ lb, $v_1 = \$0.75$, $v_2 = \$0.70$, $v_3 = \$0.50$. Let $J_N^*(W)$ be the maximum return that can be obtained by allocating resources of amount W among N activities. Clearly,

$$J_N^*(W) = \max_{\substack{w_1, w_2, \ldots, w_N \geq 0 \\ w_1 + w_2 + \cdots + w_N \leq W}} \left\{ \sum_{i=1}^{N} w_i v_i \right\} \qquad \text{(I)}$$

(a) Show that Eq. (I) leads to the functional recurrence equation

$$J_N^*(W) = \max_{\substack{w_N \\ 0 \le w_N \le W}} \{w_N v_N + J_{N-1}^*(W - w_N)\}. \tag{II}$$

Hint: Start with a one-activity process, then consider a two-activity process, and so on.

(b) Use Eq. (II) to verify the solution of the shipping problem given above.

(c) Suppose that in the shipping problem the value of the second car is \$2500, and each refrigerator after the tenth is valued at only \$250. The kitchen sinks remain at \$50 apiece. Use dynamic programming to determine the optimal loading schedule.

Use a digital computer to solve the following problems

3-14. A system is described by the first-order difference equation

$$x(k+1) = [1 + a\,\Delta t]x(k) + b\,\Delta t\,u(k), \qquad k = 0, 1, \ldots, N-1,$$

and the performance measure

$$J = x^2(N) + \lambda\,\Delta t \sum_{k=0}^{N-1} u^2(k)$$

is to be minimized subject to the constraints

$$-1.0 \le u(k) \le 1.0, \qquad k = 0, 1, \ldots, N-1$$

and

$$0.0 \le x(k) \le 1.5, \qquad k = 0, 1, \ldots, N.$$

(a) Use the dynamic programming algorithm to find the optimal control value(s) and minimum cost for each state value on the grid $x(k) = 0.0, 0.02, \ldots, 1.5$. Assume quantization levels for the control of $u(k) = -1.00, -0.98, \ldots, 0.98, 1.00$; and assume $a = 0.0$; $\Delta t = 1.0$; $b = 1.0$; $\lambda = 2.0$; $N = 2$. (This is the same problem as considered in Section 3.5, but with a finer grid structure.)

(b) Repeat part (a) with $N = 3$, i.e.,

$$J = x^2(3) + 2\sum_{k=0}^{2} u^2(k).$$

(c) Repeat part (a) with $\lambda = 4.0$.

(d) Repeat part (a) with $\lambda = 0.5$.

3-15. Repeat Problem 3-14 with the state constraints

$$0.0 \le x(k) \le 3.0.$$

Use the same quantization increments as in Problem 3-14.

3-16. Put $a = -0.4$ and repeat Problem 3-14 with the other data unchanged.

3-17. Put $a = -0.4$ and repeat Problem 3-15 with the other data unchanged.

3-18. Find the control law to transfer the system described by the difference equation

$$x(k+1) = [1 + a\,\Delta t]x(k) + b\,\Delta t\,u(k)$$

precisely to the origin in one stage with no penalty for control expenditure. The numerical values are $a = 0.0$; $b = 1.0$; $\Delta t = 1.0$. The state values are to be quantized in steps of 0.02, and the control values in steps of 0.02. These values are constrained by

$$-1.0 \leq u(k) \leq 1.0, \qquad k = 0$$

and

$$0.0 \leq x(k) \leq 3.0, \qquad k = 0, 1.$$

Problems 3-19 through 3-22 pertain to the digital computer results obtained for Problems 3-14 through 3-17.

3-19. Use the results of Problems 3-14(a), (c), (d) with $x(0) = 1.5$ to explain qualitatively

(a) The effect of varying λ on the optimal control sequence.

(b) The effect of varying λ on the final state value $x(2)$

Show all *applicable* numerical results used.

3-20. (a) Use the results of Problems 3-14(a) and 3-16(a) with $x(0) = 1.5$ to explain qualitatively the effect of the system dynamics (state equation) on the optimal control sequence, final state value, and minimum cost. Show all work.

(b) What would you expect if in Problem 3-16(a) the parameter a had been $+0.4$ instead of -0.4?

3-21. For the plant and performance measure of Problem 3-15(b)

(a) Find the optimal control sequence $\{u^*(0), u^*(1), u^*(2)\}$, the minimum cost, and the final state value $x(3)$, if the initial state value is $x(0) = 2.5$.

(b) Suppose that there is an unpredictable disturbance such that at $k = 1$ the actual value of $x(k)$ is 0.1 larger than expected. Find the optimal control sequence and the final state value, if $x(0) = 2.5$.

3-22. The statement "This means that the optimal policy and minimum costs for a K-stage process are contained (or imbedded) in the results for an N-stage process, provided that $N \geq K$" appears on page 70 of the text.

(a) Demonstrate that this statement is true for Problems 3-14(a) and 3-14(b). A brief, but clear, explanation is sufficient.

(b) Is this statement valid for time-varying processes? Explain.

3-23. It is desired to determine the control law that causes the plant

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -x_1(t) - 2x_2(t) + u(t)$$

to minimize the performance measure

$$J = 10x_1^2(T) + \tfrac{1}{2}\int_0^T [x_1^2(t) + 2x_2^2(t) + u^2(t)]\,dt.$$

The final time T is 10, and the states and control are not constrained by any boundaries. Find the optimal control law by

(a) Integrating the Riccati equation (3.12-14) with an integration interval of 0.02.

(b) Solving the recurrence equations (3.10-19), (3.10-21), and (3.10-20a). Use $\Delta t = 0.02$ in approximating the state differential equations by a set of difference equations.

3-24. Repeat Problem 3-23 for the plant

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -x_1(t) + 2x_2(t) + u(t).$$

III

The Calculus of Variations and Pontryagin's Minimum Principle

4

The Calculus of Variations

A branch of mathematics that is extremely useful in solving optimization problems is the calculus of variations. Queen Dido of Carthage was apparently the first person to attack a problem that can readily be solved by using variational calculus.† Dido, having been promised all of the land she could enclose with a bull's hide, cleverly cut the hide into many lengths and tied the ends together. Having done this, her problem was to find the closed curve with a fixed perimeter that encloses the maximum area. We know that she should have chosen a circle. The calculus of variations enables us to prove this fact and, in addition, other results that are more useful, since real estate transactions are performed somewhat differently today.

Although the history of the calculus of variations dates back to the ancient Greeks, it was not until the seventeenth century in western Europe that substantial progress was made. Sir Isaac Newton used variational principles to determine the shape of a body moving in air that encounters the least resistance. Another problem of historical interest is the brachistochrone problem shown in Fig. 4-1, posed by Johann Bernoulli in 1696. Under the influence of gravity, the bead slides along a frictionless wire with fixed end points A and B. The problem is to find the shape of the wire that causes the bead to move from A to B in minimum time. The solution, a cycloid lying in the vertical plane, is credited to Johann and Jacob Bernoulli, Newton, and L'Hospital.

† See [M-2].

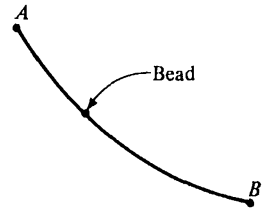


Figure 4-1 The brachistochrone problem

In Dido's problem, and in the brachistochrone problem, curves are sought which cause some criterion to assume extreme values. The connection with the optimal control problem, wherein we seek a control function that minimizes a performance measure, should be apparent.

4.1 FUNDAMENTAL CONCEPTS

In optimal control problems the objective is to determine a function that minimizes a specified *functional*—the performance measure. The analogous problem in calculus is to determine a point that yields the minimum value of a function. In this section we shall introduce some new concepts concerning functionals by appealing to some familiar results from the theory of functions.†

Functionals

To begin, let us review the definition of a function.

DEFINITION 4-1

A *function* f is a rule of correspondence that assigns to each element $\mathbf{q}$ in a certain set $\mathscr{D}$ a unique element in a set $\mathscr{R}$. $\mathscr{D}$ is called the *domain* of f and $\mathscr{R}$ is the *range*.

We shall be considering functions that assign a real number to each point (or vector) in n-dimensional Euclidean space.‡

Example 4.1-1. Suppose $q_1, q_2, \ldots, q_n$ are the coordinates of a point in n-dimensional Euclidean space and

† Appropriate references for functions of real variables are [B-4] and [O-2]. For additional reading on the calculus of variations see [G-1] and [E-1].

‡ It is assumed that the reader is familiar with the concept of a Euclidean space. See [O-2], pp. 293–301 for a detailed exposition.

$$f(\mathbf{q}) = \sqrt{q_1^2 + q_2^2 + \cdots + q_n^2}. \tag{4.1-1}$$

The real number assigned by f is the distance of the point $\mathbf{q}$ from the origin.

The definition of a functional parallels that of a function.

DEFINITION 4-2

A *functional* J is a rule of correspondence that assigns to each function $\mathbf{x}$ in a certain class Ω a unique real number. Ω is called the *domain* of the functional, and the set of real numbers associated with the functions in Ω is called the *range* of the functional.

Notice that the domain of a functional is a class of functions; intuitively, we might say that a functional is a "function of a function."

Example 4.1-2. Suppose that x is a continuous function of t defined in the interval $[t_0, t_f]$ and

$$J(x) = \int_{t_0}^{t_f} x(t)\, dt; \tag{4.1-2}$$

the real number assigned by the functional J is the area under the $x(t)$ curve.

Linearity of Functionals

Let us review the concept of linearity, which will be useful to us later, by considering a function f of $\mathbf{q}$, defined for $\mathbf{q} \in \mathscr{D}$.

DEFINITION 4-3

f is a *linear function* of $\mathbf{q}$ if and only if it satisfies the *principle of homogeneity*

$$f(\alpha\mathbf{q}) = \alpha f(\mathbf{q}) \tag{4.1-3}$$

for all $\mathbf{q} \in \mathscr{D}$ and for all real numbers α such that $\alpha\mathbf{q} \in \mathscr{D}$, *and* the *principle of additivity*

$$f(\mathbf{q}^{(1)} + \mathbf{q}^{(2)}) = f(\mathbf{q}^{(1)}) + f(\mathbf{q}^{(2)}) \tag{4.1-4}$$

for all $\mathbf{q}^{(1)}$, $\mathbf{q}^{(2)}$, and $\mathbf{q}^{(1)} + \mathbf{q}^{(2)}$ in $\mathscr{D}$.†

† In our applications we shall be concerned only with functions of real variables, so α and the components of $\mathbf{q}$ will be real numbers.

Example 4.1-3. If $f(t) = 5t$ for all t, then

$$f(\alpha t) = 5[\alpha t] \tag{4.1-5a}$$

and

$$\alpha f(t) = \alpha[5t]; \tag{4.1-5b}$$

therefore, since

$$5[\alpha t] = \alpha[5t] \tag{4.1-5c}$$

for all t, the principle of homogeneity is satisfied. Now, let us test to see if the property of additivity is satisfied.

$$f(t^{(1)} + t^{(2)}) = 5[t^{(1)} + t^{(2)}] \tag{4.1-6a}$$

and

$$f(t^{(1)}) + f(t^{(2)}) = 5t^{(1)} + 5t^{(2)}; \tag{4.1-6b}$$

thus, since

$$5[t^{(1)} + t^{(2)}] = 5t^{(1)} + 5t^{(2)} \tag{4.1-6c}$$

for all $t^{(1)}$, $t^{(2)}$, the principle of additivity is satisfied. Since the principle of homogeneity *and* the principle of additivity are both satisfied, f is a linear function.

Now consider the function g; with $g(t) = 2/t$ for all $t > 0$, then

$$g(\alpha t) = \frac{2}{\alpha t} \tag{4.1-7a}$$

and

$$\alpha g(t) = \alpha\left[\frac{2}{t}\right] \tag{4.1-7b}$$

Clearly,

$$\frac{2}{\alpha t} \neq \alpha\left[\frac{2}{t}\right] \tag{4.1-7c}$$

for all α; therefore, the principle of homogeneity is not satisfied, and g is a nonlinear function.

Next, we shall define a linear functional. Assume that $\mathbf{x}$ is a function which is a member of some class Ω, and J is a functional of $\mathbf{x}$; that is, to each $\mathbf{x}$ in Ω, J assigns a unique real number.

DEFINITION 4-4

J is a *linear functional* of $\mathbf{x}$ if and only if it satisfies the *principle of homogeneity*

$$J(\alpha\mathbf{x}) = \alpha J(\mathbf{x}) \tag{4.1-8a}$$

for all $\mathbf{x} \in \Omega$ and for all real numbers α such that $\alpha\mathbf{x} \in \Omega$, *and* the *principle of additivity*

$$J(\mathbf{x}^{(1)} + \mathbf{x}^{(2)}) = J(\mathbf{x}^{(1)}) + J(\mathbf{x}^{(2)}) \tag{4.1-8b}$$

for all $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, and $\mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ in Ω.

Example 4.1-4. Consider the functional

$$J(x) = \int_{t_0}^{t_f} x(t)\, dt, \tag{4.1-9}$$

where x is a continuous function of t. Let us see if this functional satisfies the principles of homogeneity and additivity.

Homogeneity:

$$\alpha J(x) = \alpha \int_{t_0}^{t_f} x(t)\, dt, \tag{4.1-10a}$$

$$J(\alpha x) = \int_{t_0}^{t_f} \alpha x(t)\, dt; \tag{4.1-10b}$$

therefore,

$$J(\alpha x) = \alpha J(x) \tag{4.1-10c}$$

for all real α and for all x and αx in Ω.

Additivity:

$$J(x^{(1)} + x^{(2)}) = \int_{t_0}^{t_f} [x^{(1)}(t) + x^{(2)}(t)]\, dt, \tag{4.1-11a}$$

$$J(x^{(1)}) = \int_{t_0}^{t_f} x^{(1)}(t)\, dt, \tag{4.1-11b}$$

$$J(x^{(2)}) = \int_{t_0}^{t_f} x^{(2)}(t)\, dt; \tag{4.1-11c}$$

therefore,

$$J(x^{(1)} + x^{(2)}) = J(x^{(1)}) + J(x^{(2)}) \tag{4.1-11d}$$

for all $x^{(1)}$, $x^{(2)}$, and $x^{(1)} + x^{(2)}$ in Ω.

Since additivity *and* homogeneity are both satisfied, the functional is linear.

Now consider the functional

$$J(x) = \int_{t_0}^{t_f} x^2(t)\, dt, \tag{4.1-12}$$

where x is a continuous function of t. Again let us ascertain whether homogeneity and additivity are satisfied.

Homogeneity:

$$\begin{aligned} J(\alpha x) &= \int_{t_0}^{t_f} [\alpha x(t)]^2\, dt \\ &= \alpha^2 \int_{t_0}^{t_f} x^2(t)\, dt, \end{aligned} \tag{4.1-13a}$$

$$\alpha J(x) = \alpha \int_{t_0}^{t_f} x^2(t)\, dt. \tag{4.1-13b}$$

Clearly,

$$J(\alpha x) \neq \alpha J(x) \tag{4.1-13c}$$

for all α, so the functional (4.1-12) is *nonlinear*.

Closeness of Functions

If two points are said to be close to one another, a geometric interpretation springs immediately to mind. But what do we mean when we say two *functions* are close to one another? To give a precise meaning to the term "close" we next introduce the concept of a norm.

DEFINITION 4-5

The *norm* in n-dimensional Euclidean space is a rule of correspondence that assigns to each point $\mathbf{q}$ a real number. The norm of $\mathbf{q}$, denoted by $\|\mathbf{q}\|$, satisfies the following properties:

1. $\|\mathbf{q}\| \geq 0$ and $\|\mathbf{q}\| = 0$ if and only if $\mathbf{q} = \mathbf{0}$. (4.1-14a)
2. $\|\alpha\mathbf{q}\| = |\alpha| \cdot \|\mathbf{q}\|$ for all real numbers α. (4.1-14b)
3. $\|\mathbf{q}^{(1)} + \mathbf{q}^{(2)}\| \leq \|\mathbf{q}^{(1)}\| + \|\mathbf{q}^{(2)}\|$. (4.1-14c)

When we say that two points $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$ are close together, we mean that

$$\|\mathbf{q}^{(1)} - \mathbf{q}^{(2)}\| \text{ is small.}$$

Example 4.1-5. What is a suitable norm for two-dimensional Euclidean space? It is easily verified that

$$\|\mathbf{q}\|_2 \triangleq \sqrt{q_1^2 + q_2^2}, \quad \text{or} \quad \|\mathbf{q}\|_1 \triangleq |q_1| + |q_2|$$

satisfies properties (4.1-14). Now suppose that a point $\mathbf{q}^{(1)}$ is specified and it is required that $\|\mathbf{q}^{(2)} - \mathbf{q}^{(1)}\| < \delta$. What are the acceptable locations for $\mathbf{q}^{(2)}$? If $\|\mathbf{q}\|_2$ is used as the norm, $\mathbf{q}^{(2)}$ must lie within the circle centered at $\mathbf{q}^{(1)}$ having radius δ as shown in Fig. 4-2(a). On the other hand, if $\|\mathbf{q}\|_1$ is used as the norm, the acceptable locations for $\mathbf{q}^{(2)}$ are as shown in Fig. 4-2(b).

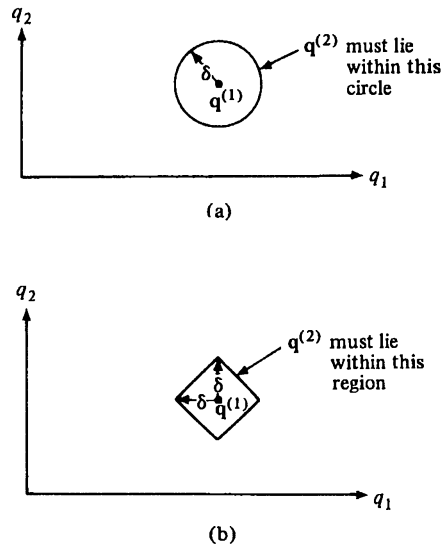


Fig. 4-2 (a) The set of points that satisfy $\|\mathbf{q}^{(2)} - \mathbf{q}^{(1)}\|_2 < \delta$ (b) The set of points that satisfy $\|\mathbf{q}^{(2)} - \mathbf{q}^{(1)}\|_1 < \delta$

Next, let us define the norm of a function.

DEFINITION 4-6

The *norm of a function* is a rule of correspondence that assigns to each function $\mathbf{x} \in \Omega$, defined for $t \in [t_0, t_f]$, a real number. The norm of $\mathbf{x}$, denoted by $\|\mathbf{x}\|$, satisfies the following properties:

1. $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x}(t) = \mathbf{0}$ for all $t \in [t_0, t_f]$. (4.1-15a)
2. $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ for all real numbers α. (4.1-15b)
3. $\|\mathbf{x}^{(1)} + \mathbf{x}^{(2)}\| \leq \|\mathbf{x}^{(1)}\| + \|\mathbf{x}^{(2)}\|$. (4.1-15c)

To compare the closeness of two functions $\mathbf{y}$ and $\mathbf{z}$ that are defined for $t \in [t_0, t_f]$, let $\mathbf{x}(t) = \mathbf{y}(t) - \mathbf{z}(t)$.

Intuitively speaking, the norm of the difference of two functions should

be zero if the functions are identical, small if the functions are "close," and large if the functions are "far apart."

Example 4.1-6. x is a continuous scalar function of t defined in the interval $[t_0, t_f]$. Define an acceptable norm for x.

$$\|x\| = \max_{t_0 \le t \le t_f} \{|x(t)|\} \tag{4.1-16}$$

is a suitable norm because it satisfies the three properties given in (4.1-15).

The Increment of a Functional

In order to consider extreme values of a function, we now define the concept of an increment.

DEFINITION 4-7

If $\mathbf{q}$ and $\mathbf{q} + \Delta\mathbf{q}$ are elements for which the function f is defined, then the *increment* of f, denoted by Δf, is

$$\Delta f \triangleq f(\mathbf{q} + \Delta\mathbf{q}) - f(\mathbf{q}). \tag{4.1-17}$$

Notice that Δf depends on both $\mathbf{q}$ and $\Delta\mathbf{q}$, in general, so to be more explicit we would write $\Delta f(\mathbf{q}, \Delta\mathbf{q})$.

Example 4.1-7. Consider the function

$$f(\mathbf{q}) = q_1^2 + 2q_1q_2 \quad \text{for all real } q_1, q_2. \tag{4.1-18}$$

The increment of f is

$$\begin{aligned} \Delta f = f(\mathbf{q} + \Delta\mathbf{q}) - f(\mathbf{q}) &= [q_1 + \Delta q_1]^2 \\ &\quad + 2[q_1 + \Delta q_1][q_2 + \Delta q_2] - [q_1^2 + 2q_1q_2] \\ &= 2q_1\,\Delta q_1 + [\Delta q_1]^2 + 2\,\Delta q_1 q_2 + 2\,\Delta q_2 q_1 + 2\,\Delta q_1\,\Delta q_2 \end{aligned} \tag{4.1-19}$$

In an analogous manner, we next define the increment of a functional.

DEFINITION 4-8

If $\mathbf{x}$ and $\mathbf{x} + \delta\mathbf{x}$ are functions for which the functional J is defined, then the *increment of* J, denoted by ΔJ, is

$$\Delta J \triangleq J(\mathbf{x} + \delta\mathbf{x}) - J(\mathbf{x}). \tag{4.1-20}$$

Again, to be more explicit, we would write $\Delta J(\mathbf{x}, \delta\mathbf{x})$ to emphasize that the increment depends on the functions $\mathbf{x}$ and $\delta\mathbf{x}$. $\delta\mathbf{x}$ is called the *variation* of the function $\mathbf{x}$.

Example 4.1-8. Find the increment of the functional

$$J(x) = \int_{t_0}^{t_f} x^2(t)\, dt, \tag{4.1-21}$$

where x is a continuous function of t.

The increment is

$$\begin{aligned} \Delta J &= J(x + \delta x) - J(x) \\ &= \int_{t_0}^{t_f} [x(t) + \delta x(t)]^2\, dt - \int_{t_0}^{t_f} x^2(t)\, dt \\ &= \int_{t_0}^{t_f} [2x(t)\delta x(t) + [\delta x(t)]^2]\, dt. \end{aligned} \tag{4.1-22}$$

The Variation of a Functional

The preceding definitions have laid the foundation for considering the variation of a functional. The variation plays the same role in determining extreme values of *functionals* as the differential does in finding maxima and minima of *functions*. As review, we next state the definition of the differential of a function.

DEFINITION 4-9

The increment of a function of n variables can be written as

$$\Delta f(\mathbf{q}, \Delta\mathbf{q}) = df(\mathbf{q}, \Delta\mathbf{q}) + g(\mathbf{q}, \Delta\mathbf{q}) \cdot \| \Delta\mathbf{q} \|, \tag{4.1-23}$$

where df is a linear function of $\Delta\mathbf{q}$. If

$$\lim_{\| \Delta\mathbf{q} \| \to 0} \{g(\mathbf{q}, \Delta\mathbf{q})\} = 0,$$

then f is said to be *differentiable* at $\mathbf{q}$, and df is *the differential of* f at the point $\mathbf{q}$.

If f is a differentiable function of *one* variable t, then the differential can be written

$$df(t, \Delta t) = f'(t)\, \Delta t; \tag{4.1-24}$$

$f'(t)$ is called the *derivative* of f at t. Figure 4-3 gives a geometric interpretation of the increment Δf, the differential df, and the derivative f': $f'(t_1)$ is the slope of the line that is tangent to f at the time t_1; $f'(t_1)\, \Delta t$ is a first-order (linear) approximation to Δf (the smaller Δt, the better the approximation).

Example 4.1-9. Find the differential of

$$f(\mathbf{q}) = q_1^2 + 2q_1q_2 \tag{4.1-25}$$

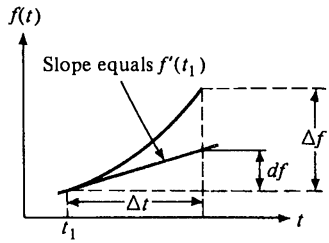


Figure 4-3 Geometric interpretation of Δf, df, f'

In Example 4.1-7 we found that the increment is

$$\Delta f(\mathbf{q}, \Delta\mathbf{q}) = [2q_1 + 2q_2]\,\Delta q_1 + [2q_1]\,\Delta q_2 + [\Delta q_1]^2 + 2\,\Delta q_1\,\Delta q_2. \tag{4.1-26}$$

The first two terms are linear in $\Delta\mathbf{q}$. Letting

$$\|\Delta\mathbf{q}\| \triangleq \sqrt{[\Delta q_1]^2 + [\Delta q_2]^2}, \tag{4.1-27}$$

we can write the last two terms as

$$\frac{[\Delta q_1]^2 + 2\,\Delta q_1\,\Delta q_2}{\sqrt{[\Delta q_1]^2 + [\Delta q_2]^2}} \cdot \sqrt{[\Delta q_1]^2 + [\Delta q_2]^2}, \tag{4.1-28}$$

which is of the form $g(\mathbf{q}, \Delta\mathbf{q}) \cdot \|\Delta\mathbf{q}\|$. To show that f is differentiable we must verify that

$$\lim_{\|\Delta\mathbf{q}\| \to 0} \left\{ \frac{[\Delta q_1]^2 + 2\,\Delta q_1\,\Delta q_2}{\sqrt{[\Delta q_1]^2 + [\Delta q_2]^2}} \right\} = 0. \tag{4.1-29}$$

It will be left as an exercise for the interested reader to verify that this limit exists and is zero; hence f is differentiable, and the differential is

$$df(\mathbf{q}, \Delta\mathbf{q}) = [2q_1 + 2q_2]\,\Delta q_1 + [2q_1]\,\Delta q_2. \tag{4.1-30}$$

Rather than go through all of these steps, we can use Definition 4-9 to develop a rule for finding the differential of a function. In particular, if f is a differentiable function of n variables, the differential df is given by

$$df = \frac{\partial f}{\partial q_1}\,\Delta q_1 + \frac{\partial f}{\partial q_2}\,\Delta q_2 + \cdots + \frac{\partial f}{\partial q_n}\,\Delta q_n. \tag{4.1-31}$$

We shall also find it convenient to develop a formal procedure for finding the variation of a functional rather than starting each time from the definition which follows.

DEFINITION 4-10

The increment of a functional can be written as

$$\Delta J(\mathbf{x}, \delta\mathbf{x}) = \delta J(\mathbf{x}, \delta\mathbf{x}) + g(\mathbf{x}, \delta\mathbf{x}) \cdot \|\delta\mathbf{x}\|, \tag{4.1-32}$$

where δJ is linear in $\delta\mathbf{x}$. If

$$\lim_{\|\delta\mathbf{x}\| \to 0} \{g(\mathbf{x}, \delta\mathbf{x})\} = 0,$$

then J is said to be *differentiable* on $\mathbf{x}$ and δJ is the *variation of* J evaluated for the function $\mathbf{x}$.

Example 4.10. Let x be a continuous scalar function defined for $t \in [0, 1]$. Find the variation of the functional

$$J(x) = \int_0^1 [x^2(t) + 2x(t)]\, dt. \tag{4.1-33}$$

First, find the increment of J,

$$\begin{aligned} \Delta J(x, \delta x) &= J(x + \delta x) - J(x) \\ &= \int_0^1 \{[x(t) + \delta x(t)]^2 + 2[x(t) + \delta x(t)]\}\, dt \\ &\quad - \int_0^1 [x^2(t) + 2x(t)]\, dt. \end{aligned} \tag{4.1-34}$$

Expanding, and combining these integrals, we obtain

$$\Delta J(x, \delta x) = \int_0^1 \{[2x(t) + 2]\, \delta x(t) + [\delta x(t)]^2\}\, dt. \tag{4.1-35}$$

Separating the terms which are linear in δx, we have

$$\Delta J(x, \delta x) = \int_0^1 \{[2x(t) + 2]\, \delta x(t)\}\, dt + \int_0^1 [\delta x(t)]^2\, dt. \tag{4.1-36}$$

Now let us verify that the second integral can be written

$$\int_0^1 [\delta x(t)]^2\, dt = g(x, \delta x) \cdot \|\delta x\| \tag{4.1-37}$$

and that

$$\lim_{\|\delta x\| \to 0} \{g(x, \delta x)\} = 0. \tag{4.1-38}$$

Since x is a continuous function, let

$$\|\delta x\| \triangleq \max_{0 \le t \le 1} \{|\delta x(t)|\}. \tag{4.1-39}$$

Multiplying the left side of (4.1-37) by $\|\delta x\|/\|\delta x\|$ gives

$$\frac{\|\delta x\|}{\|\delta x\|} \cdot \int_0^1 [\delta x(t)]^2 \, dt = \|\delta x\| \cdot \int_0^1 \frac{[\delta x(t)]^2}{\|\delta x\|} \, dt; \tag{4.1-40}$$

the right side of Eq. (4.1-40) follows because $\|\delta x\|$ does not depend on t. Comparing (4.1-40) with (4.1-37), we observe that

$$g(x, \delta x) = \int_0^1 \frac{[\delta x(t)]^2}{\|\delta x\|} \, dt. \tag{4.1-41}$$

Writing $[\delta x(t)]^2$ as $|\delta x(t)| \cdot |\delta x(t)|$ gives

$$\int_0^1 \frac{|\delta x(t)| \cdot |\delta x(t)|}{\|\delta x\|} \, dt \leq \int_0^1 |\delta x(t)| \, dt, \tag{4.1-42}$$

because of the definition of the norm of δx, which implies that $\|\delta x\| \geq |\delta x(t)|$ for all $t \in [0, 1]$. Clearly, if $\|\delta x\| \to 0$, $|\delta x(t)| \to 0$ for all $t \in [0, 1]$, and thus

$$\lim_{\|\delta x\| \to 0} \left\{ \int_0^1 |\delta x(t)| \, dt \right\} = 0. \tag{4.1-43}$$

We have succeeded in verifying that the increment can be written in the form of Eq. (4.1-32) and that $g(x, \delta x) \to 0$ as $\|\delta x\| \to 0$; therefore, the variation of J is

$$\delta J(x, \delta x) = \int_0^1 \{[2x(t) + 2] \, \delta x(t)\} \, dt. \tag{4.1-44}$$

This expression can also be obtained by formally expanding the integrand of ΔJ in a Taylor series about $x(t)$ and retaining only the terms of first order in $\delta x(t)$.

It is very important to keep in mind that δJ is the linear approximation to the difference in the functional J caused by two comparison curves. If the comparison curves are close ($\|\delta \mathbf{x}\|$ small), then the variation should be a good approximation to the increment; however, δJ may be a poor approximation to ΔJ if the comparison curves are far apart. The analogy in calculus is illustrated in Fig. 4-3, where it is seen that df is a good approximation to Δf for small Δt.

As with differentials, we would prefer to avoid using the definition each time the variation of a functional is to be determined; in Section 4.2 we shall develop a formal procedure for finding variations of functionals.

Maxima and Minima of Functionals

Let us now review the definition of an extreme value of a function.

DEFINITION 4-11

A function f with domain $\mathscr{D}$ has a *relative extremum* at the point $\mathbf{q}^*$ if there is an $\epsilon > 0$ such that for all points $\mathbf{q}$ in $\mathscr{D}$ that satisfy $\|\mathbf{q} - \mathbf{q}^*\| < \epsilon$ the increment of f has the same sign. If

$$\Delta f = f(\mathbf{q}) - f(\mathbf{q}^*) \geq 0, \tag{4.1-45}$$

$f(\mathbf{q}^*)$ is a *relative minimum;* if

$$\Delta f = f(\mathbf{q}) - f(\mathbf{q}^*) \leq 0, \tag{4.1-46}$$

$f(\mathbf{q}^*)$ is a *relative maximum.*

If (4.1-45) is satisfied for arbitrarily large ϵ, then $f(\mathbf{q}^*)$ is a *global,* or *absolute, minimum.* Similarly, if (4.1-46) holds for arbitrarily large ϵ, then $f(\mathbf{q}^*)$ is a *global,* or *absolute, maximum.*

Recall the procedure for locating extrema of functions. Generally, one attempts to find points where the differential vanishes—a necessary condition for an extremum at an interior point of $\mathscr{D}$. Assuming that there are such points and that they can be determined, then one can examine the behavior of the function in the vicinity of these points.

Example 4.1-11. Consider the function of one variable illustrated in Fig. 4-4. The function is defined for $t \in [t_0, t_f]$. Since the interval is bounded

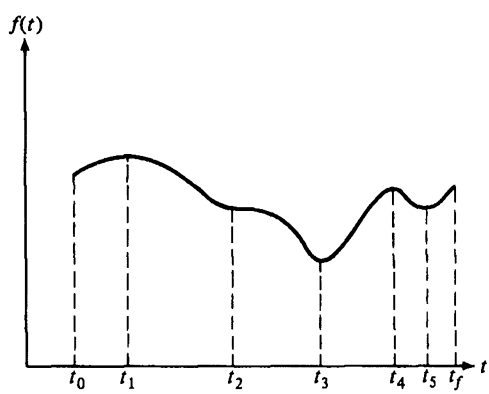


Figure 4-4 A function with several extrema

and closed, candidates for extrema are located at points where the differential vanishes and also at the end points. For this function, the differential vanishes at t_1, t_2, t_3, t_4, and t_5—these are called stationary points. t_2, however, is not an extreme point; it is a horizontal inflection

point. t_1 and t_4 are relative maxima, and t_3 and t_5 are relative minima. Examining the function at the end points, we see that t_0 is a relative minimum and t_f is a relative maximum. It is easily shown for a function of one variable that at the left end point

$$\frac{df}{dt} > 0 \quad \text{implies that } t_0 \text{ is a relative minimum,}$$

and

$$\frac{df}{dt} < 0 \quad \text{implies that } t_0 \text{ is a relative maximum.}$$

For the right-hand end point the sense of the inequalities is reversed. Finally, observe that t_1 is the absolute or global maximum point and t_3 is the global minimum.

Next, consider a functional J which is defined for all functions $\mathbf{x}$ in a class Ω.

DEFINITION 4-12

A functional J with domain Ω has a relative extremum at $\mathbf{x}^*$ if there is an $\epsilon > 0$ such that for all functions $\mathbf{x}$ in Ω which satisfy $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$ the increment of J has the same sign. If

$$\Delta J = J(\mathbf{x}) - J(\mathbf{x}^*) \geq 0, \tag{4.1-47}$$

$J(\mathbf{x}^*)$ is a *relative minimum*; if

$$\Delta J = J(\mathbf{x}) - J(\mathbf{x}^*) \leq 0, \tag{4.1-48}$$

$J(\mathbf{x}^*)$ is a *relative maximum*.

If (4.1-47) is satisfied for arbitrarily large ϵ, then $J(\mathbf{x}^*)$ is a *global*, or *absolute*, *minimum*. Similarly, if (4.1-48) holds for arbitrarily large ϵ, then $J(\mathbf{x}^*)$ is a *global*, or *absolute*, *maximum*. $\mathbf{x}^*$ is called an *extremal*, and $J(\mathbf{x}^*)$ is referred to as an *extremum*.

The Fundamental Theorem of the Calculus of Variations

The fundamental theorem used in finding extreme values of functions is the necessary condition that the differential vanish at an extreme point (except extrema at the boundaries of closed regions). In variational problems, the analogous theorem is that the variation must be zero on an extremal curve, provided that there are no bounds imposed on the curves. We next state this theorem and give the proof.

Let $\mathbf{x}$ be a vector function of t in the class Ω, and $J(\mathbf{x})$ be a differentiable

functional of $\mathbf{x}$. *Assume that the functions in* Ω *are not constrained by any boundaries.*

The fundamental theorem of the calculus of variations is

If $\mathbf{x}^*$ is an extremal, the variation of J must vanish on $\mathbf{x}^*$; that is,

$$\delta J(\mathbf{x}^*, \delta\mathbf{x}) = 0 \text{ for all admissible } \delta\mathbf{x}.\dagger \tag{4.1-49}$$

Proof by contradiction: Assume that $\mathbf{x}^*$ is an extremal and that $\delta J(\mathbf{x}^*, \delta\mathbf{x}) \neq 0$. Let us show that these assumptions imply that the increment ΔJ can be made to change sign in an arbitrarily small neighborhood of $\mathbf{x}^*$.

The increment is

$$\begin{aligned} \Delta J(\mathbf{x}^*, \delta\mathbf{x}) &= J(\mathbf{x}^* + \delta\mathbf{x}) - J(\mathbf{x}^*) \\ &= \delta J(\mathbf{x}^*, \delta\mathbf{x}) + g(\mathbf{x}^*, \delta\mathbf{x}) \cdot \|\delta\mathbf{x}\|, \end{aligned} \tag{4.1-50}$$

where $g(\mathbf{x}^*, \delta\mathbf{x}) \to 0$ as $\|\delta\mathbf{x}\| \to 0$; thus, there is a neighborhood, $\|\delta\mathbf{x}\| < \epsilon$, where $g(\mathbf{x}^*, \delta\mathbf{x}) \cdot \|\delta\mathbf{x}\|$ is small enough so that δJ dominates the expression for ΔJ.

Now let us select the variation

$$\delta\mathbf{x} = \alpha\delta\mathbf{x}^{(1)} \tag{4.1-51}$$

shown in Fig. 4-5 (for a scalar function), where $\alpha > 0$ and $\|\alpha\delta\mathbf{x}^{(1)}\| < \epsilon$. Suppose that

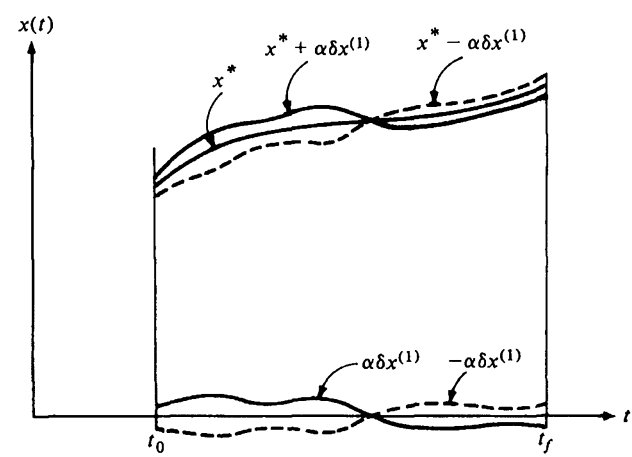


Figure 4-5 An extremal and two neighboring curves

† By admissible $\delta\mathbf{x}$ we mean that $\mathbf{x} + \delta\mathbf{x}$ must be a member of the class Ω; thus, if Ω is the class of continuous functions, $\mathbf{x}$ and $\delta\mathbf{x}$ are required to be continuous.

$$\delta J(\mathbf{x}^*, \alpha\delta\mathbf{x}^{(1)}) < 0. \tag{4.1-52}$$

Since δJ is a linear functional of $\delta\mathbf{x}$, the principle of homogeneity [see Eq. (4.1-8a)] gives

$$\delta J(\mathbf{x}^*, \alpha\delta\mathbf{x}^{(1)}) = \alpha\delta J(\mathbf{x}^*, \delta\mathbf{x}^{(1)}) < 0. \tag{4.1-53}$$

The signs of ΔJ and δJ are the same for $\|\delta\mathbf{x}\| < \epsilon$; thus,

$$\Delta J(\mathbf{x}^*, \alpha\delta\mathbf{x}^{(1)}) < 0. \tag{4.1-54}$$

Next, we consider the variation

$$\delta\mathbf{x} = -\alpha\delta\mathbf{x}^{(1)}$$

shown in Fig. 4-5. Clearly, $\|\alpha\delta\mathbf{x}^{(1)}\| < \epsilon$ implies that $\|-\alpha\delta\mathbf{x}^{(1)}\| < \epsilon$; therefore, the sign of $\Delta J(\mathbf{x}^*, -\alpha\delta\mathbf{x}^{(1)})$ is the same as the sign of $\delta J(\mathbf{x}^*, -\alpha\delta\mathbf{x}^{(1)})$. Again using the principle of homogeneity, we obtain

$$\delta J(\mathbf{x}^*, -\alpha\delta\mathbf{x}^{(1)}) = -\alpha\delta J(\mathbf{x}^*, \delta\mathbf{x}^{(1)}); \tag{4.1-55}$$

therefore, since $\delta J(\mathbf{x}^*, \alpha\delta\mathbf{x}^{(1)}) < 0$, $\delta J(\mathbf{x}^*, -\alpha\delta\mathbf{x}^{(1)}) > 0$, and this implies

$$\Delta J(\mathbf{x}^*, -\alpha\delta\mathbf{x}^{(1)}) > 0. \tag{4.1-56}$$

To recapitulate, we have shown that if $\delta J(\mathbf{x}^*, \delta\mathbf{x}) \not\equiv 0$, then in an arbitrarily small neighborhood of $\mathbf{x}^*$

$$\Delta J(\mathbf{x}^*, \alpha\delta\mathbf{x}^{(1)}) < 0 \tag{4.1-57}$$

and

$$\Delta J(\mathbf{x}^*, -\alpha\delta\mathbf{x}^{(1)}) > 0, \tag{4.1-58}$$

thus contradicting the assumption that $\mathbf{x}^*$ is an extremal (see Definition 4-12). Therefore, if $\mathbf{x}^*$ is an extremal it is necessary that

$$\delta J(\mathbf{x}^*, \delta\mathbf{x}) = 0 \quad \text{for arbitrary } \delta\mathbf{x}. \tag{4.1-59}$$

The assumption that the functions in Ω are not bounded guarantees that $\alpha\delta\mathbf{x}^{(1)}$ and $-\alpha\delta\mathbf{x}^{(1)}$ are both admissible variations.

Summary

In this section important definitions have been given and the fundamental theorem of the calculus of variations has been proved. The analogy between

certain concepts of calculus and the calculus of variations has been exploited. It is helpful to think in terms of the analogies that exist; by doing so, we can appeal to familiar geometric ideas from the calculus. At the same time, we must be careful not to extrapolate results from calculus to the calculus of variations merely by using "intuitive continuation." In the next section we shall apply the fundamental theorem to problems that become progressively more general; eventually, we shall be able to attack the optimal control problem.

4.2 FUNCTIONALS OF A SINGLE FUNCTION

In this section we shall use the fundamental theorem to determine extrema of functionals depending on a single function. To relate our discussion to "the optimal control problem" posed in Chapter 1 we shall think in terms of finding state trajectories that minimize performance measures. In control problems state trajectories are determined by control histories (and initial conditions); however, to simplify the discussion it will be assumed initially that there are no such constraints and that the states can be directly and independently varied. Subsequently, this assumption will be removed.

The Simplest Variational Problem

Problem 1: Let x be a scalar function in the class of functions with continuous first derivatives. It is desired to find the function x^* for which the functional

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\, dt \tag{4.2-1}$$

has a relative extremum. The notation $J(x)$ means that J is a functional of the function x; $g(x(t), \dot{x}(t), t)$, on the other hand, is a function—g assigns a real number to the point $(x(t), \dot{x}(t), t)$. It is assumed that the integrand g has continuous first and second partial derivatives with respect to all of its arguments; t_0 and t_f are fixed, and the end points of the curve are specified as x_0 and x_f.

Curves in the class Ω which also satisfy the end conditions are called admissible. Several admissible curves are shown in Fig. 4-6.

We wish to find the curves (if any exist) that extremize $J(x)$. The search begins by finding the curves that satisfy the fundamental theorem. Let x be any curve in Ω, and determine the variation $\delta J(x, \delta x)$ from the increment

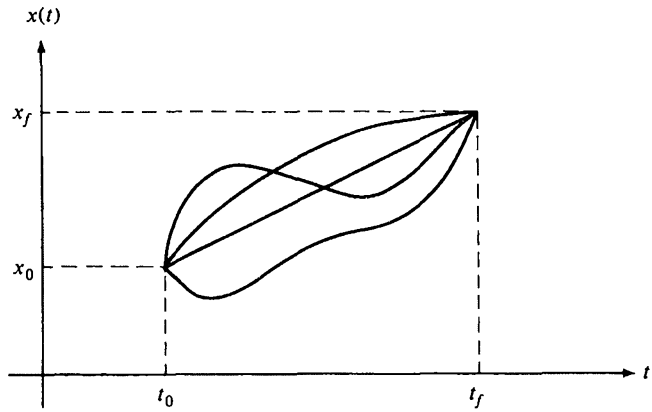


Figure 4-6 Admissible curves for *Problem 1*

$$\begin{aligned}\Delta J(x, \delta x) &= J(x + \delta x) - J(x) \\ &= \int_{t_0}^{t_f} g(x(t) + \delta x(t), \dot{x}(t) + \delta \dot{x}(t), t)\, dt \\ &\quad - \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\, dt.\end{aligned} \tag{4.2-2}$$

Combining the integrals gives

$$\Delta J(x, \delta x) = \int_{t_0}^{t_f} [g(x(t) + \delta x(t), \dot{x}(t) + \delta \dot{x}(t), t) - g(x(t), \dot{x}(t), t)]\, dt. \tag{4.2-3}$$

Notice that the dependence on $\dot{x}$ and $\delta \dot{x}$ is not indicated in the argument of ΔJ, because x and $\dot{x}$, δx and $\delta \dot{x}$ are not independent;

$$\dot{x}(t) = \frac{d}{dt}[x(t)], \qquad \delta \dot{x}(t) = \frac{d}{dt}[\delta x(t)].$$

Eventually, ΔJ will be expressed entirely in terms of x, $\dot{x}$ and δx.

Expanding the integrand of (4.2-3) in a Taylor series about the point $x(t)$, $\dot{x}(t)$ gives

$$\begin{aligned}\Delta J = \int_{t_0}^{t_f} \{&g(x(t), \dot{x}(t), t) + \left[\frac{\partial g}{\partial x}(x(t), \dot{x}(t), t)\right] \delta x(t) \\ &+ \left[\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)\right] \delta \dot{x}(t) \\ &+ \frac{1}{2}\left[\left[\frac{\partial^2 g}{\partial x^2}(x(t), \dot{x}(t), t)\right] [\delta x(t)]^2\right.\end{aligned} \tag{4.2-4}$$

$$+ 2\left[\frac{\partial^2 g}{\partial x \partial \dot{x}}(x(t), \dot{x}(t), t)\right] \delta x(t)\delta\dot{x}(t)$$
$$+ \left[\frac{\partial^2 g}{\partial \dot{x}^2}(x(t), \dot{x}(t), t)\right] [\delta\dot{x}(t)]^2\Big]$$
$$+ o([\delta x(t)]^2, [\delta\dot{x}(t)]^2) - g(x(t), \dot{x}(t), t)\} \, dt.$$

The notation $o([\delta x(t)]^2, [\delta\dot{x}(t)]^2)$ denotes terms in the expansion of order three and greater in $\delta x(t)$ and $\delta\dot{x}(t)$—these terms are smaller in magnitude than $[\delta x(t)]^2$ and $[\delta\dot{x}(t)]^2$ as $\delta x(t)$ and $\delta\dot{x}(t)$ approach zero. As indicated, the partial derivatives in Eq. (4.2-4) are evaluated on the trajectory x, $\dot{x}$.

Next, we extract the terms in ΔJ that are linear in $\delta x(t)$ and $\delta\dot{x}(t)$ to obtain the variation

$$\delta J(x, \delta x) = \int_{t_0}^{t_f} \left\{\left[\frac{\partial g}{\partial x}(x(t), \dot{x}(t), t)\right] \delta x(t) + \left[\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)\right] \delta\dot{x}(t)\right\} dt. \tag{4.2-5}$$

$\delta x(t)$ and $\delta\dot{x}(t)$ are related by

$$\delta x(t) = \int_{t_0}^{t} \delta\dot{x}(t) \, dt + \delta x(t_0); \tag{4.2-6}$$

thus, selecting δx uniquely determines $\delta\dot{x}$. We shall regard δx as being the function that is varied independently. To express (4.2-5) entirely in terms containing δx, we integrate by parts the term involving $\delta\dot{x}$ to obtain

$$\delta J(x, \delta x) = \left[\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)\right] \delta x(t)\Big|_{t_0}^{t_f} + \int_{t_0}^{t_f} \left\{\left[\frac{\partial g}{\partial x}(x(t), \dot{x}(t), t)\right] - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)\right]\right\} \delta x(t) \, dt. \tag{4.2-7}$$

Since $x(t_0)$ and $x(t_f)$ are specified, all admissible curves must pass through these points; therefore, $\delta x(t_0) = 0$, $\delta x(t_f) = 0$, and the terms outside the integral vanish.

If we now consider an extremal curve, applying the fundamental theorem yields

$$\delta J(x^*, \delta x) = 0 = \int_{t_0}^{t_f} \left\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]\right\} \delta x(t) \, dt. \tag{4.2-8}$$

Thus, the integral must be zero; does this tell us anything about the integrand?

To answer this question, consider the function δx; it has continuous derivatives, and must be zero at t_0 and t_f, but aside from these requirements it is completely arbitrary. The assumptions made regarding the function g guarantee that the term which multiplies $\delta x(t)$ in Eq. (4.2-8) is continuous. It can be shown that if a function h is continuous and

$$\int_{t_0}^{t_f} h(t)\,\delta x(t)\,dt = 0 \tag{4.2-9}$$

for every function δx that is continuous in the interval $[t_0, t_f]$, then h must be zero everywhere in the interval $[t_0, t_f]$.

This result, called the *fundamental lemma of the calculus of variations*, is proved in references [E-1] and [G-1]. The essence of the proof is as follows: Suppose that h is not zero everywhere in the interval; then, since h is continuous, there is a neighborhood in $[t_0, t_f]$ in which h has the same sign everywhere. Select δx, which is arbitrary, to be positive (or negative) throughout the neighborhood where h has the same sign, and zero elsewhere. By selecting δx in this manner the integral in Eq. (4.2-9) will be nonzero; thus, h must be identically zero for (4.2-9) to be satisfied.

Figure 4-7 shows a function h that is not identically zero in the interval

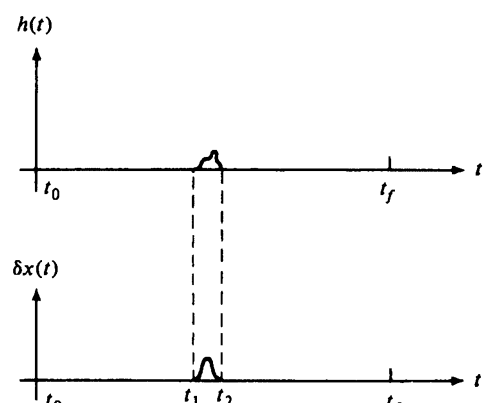


Figure 4-7 A nonzero h and an admissible δx

$[t_0, t_f]$. Selecting δx as shown makes the product $h(t)\,\delta x(t)$ greater than zero in the interval $[t_1, t_2]$, and zero elsewhere. By inspection, the integral of $h(t)\,\delta x(t)$ is certainly not zero. Notice that it does not matter what values h assumes outside of the interval $[t_1, t_2]$.

An intuitive way of looking at this lemma is the following: Given *any*

continuous function h that is not identically zero in the interval $[t_0, t_f]$, a function δx, with continuous derivatives, can be selected which makes the integral $\int_{t_0}^{t_f} h(t)\,\delta x(t)\,dt \neq 0$.

Applying the fundamental lemma to (4.2-8), we find that a necessary condition for x^* to be an extremal is

$$\boxed{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right] = 0} \qquad (4.2\text{-}10)$$

for all $t \in [t_0, t_f]$.

Let us now examine Eq. (4.2-10), called the *Euler equation*, in more detail. The presence of d/dt and/or $\dot{x}^*(t)$ means that this is a differential equation.

$$\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]$$

is, in general, a function of $x^*(t)$, $\dot{x}^*(t)$, and t; thus, when this function is differentiated with respect to t, $\ddot{x}^*(t)$ may be present. This means that the differential equation is generally of second order. There may also be terms involving products or powers of $\ddot{x}^*(t)$, $\dot{x}^*(t)$, and $x^*(t)$, in which case the differential equation is nonlinear, and the presence of t in the arguments indicates that the coefficients may be time-varying. Differential equations of this type are normally hard to solve analytically. There are, however, certain special cases (summarized in Appendix 3) in which the Euler equation can be reduced to a first-order differential equation, or solved by evaluating integrals.

In summary then, the Euler equation for *Problem 1* is generally a nonlinear, ordinary, time-varying, hard-to-solve, second-order differential equation.

Since the Euler equation usually cannot be solved analytically, one naturally thinks of using numerical integration. The characteristics of the Euler equation which make analytical solution difficult do not present serious difficulties numerically. Unfortunately, there is another factor that prevents us from simply solving the Euler equation by numerical integration—*the boundary conditions are split.* Instead of having $x(t_0)$ and $\dot{x}(t_0)$ specified [or $x(t_f)$, $\dot{x}(t_f)$], we know $x(t_0)$ and $x(t_f)$. To integrate numerically, we need values for all of the boundary conditions at one end. Thus, we see that to obtain the optimal trajectory x^*, a *nonlinear, two-point boundary-value problem* must be solved. The problem is difficult because of the combination of split boundary values and the nonlinearity of the differential equation. Separately,

either of these difficulties can be surmounted without tremendous effort, but together they present a formidable challenge. For the moment we shall consider only problems that can be solved analytically. In Chapter 6 we shall consider some numerical techniques for solving nonlinear, two-point boundary-value problems.

It should be emphasized that since the Euler equation is a necessary condition, further investigation is required to ascertain whether a solution x^* is a minimizing curve, a maximizing curve, or neither.

Example 4.2-1. Find an extremal for the functional

$$J(x) = \int_0^{\pi/2} [\dot{x}^2(t) - x^2(t)]\, dt \tag{4.2-11}$$

which satisfies the boundary conditions $x(0) = 0$ and $x(\pi/2) = 1$.

The Euler equation is

$$\begin{aligned} 0 &= \frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right] \\ &= -2x^*(t) - \frac{d}{dt}[2\dot{x}^*(t)], \end{aligned} \tag{4.2-12}$$

or

$$\ddot{x}^*(t) + x^*(t) = 0. \tag{4.2-13}$$

Since Eq. (4.2-13) is linear and has constant coefficients, it can be readily solved by using classical differential equation theory. Assuming a solution of the form $x^*(t) = k\epsilon^{st}$ and substituting this in (4.2-13), we obtain

$$ks^2\epsilon^{st} + k\epsilon^{st} = 0. \tag{4.2-14}$$

Since (4.2-14) must be satisfied for all t,

$$s^2 + 1 = 0. \tag{4.2-15}$$

The roots of this characteristic equation are $s = \pm j1$,† so the solution has the form

$$x^*(t) = c_1\epsilon^{-jt} + c_2\epsilon^{jt}, \tag{4.2-16}$$

or

$$x^*(t) = c_3 \cos(t) + c_4 \sin(t), \tag{4.2-17}$$

where the c's are constants of integration.

To determine the constants that satisfy the boundary conditions

† $j \triangleq \sqrt{-1}$.

$x(0) = 0$, $x(\pi/2) = 1$, we use the form of the solution in (4.2-17) to obtain

$$0 = c_3 \cos(0) + c_4 \sin(0) \Longrightarrow c_3 = 0\dagger \tag{4.2-18}$$

and

$$1 = \overset{0}{c_3} \cos\left(\frac{\pi}{2}\right) + c_4 \sin\left(\frac{\pi}{2}\right) \Longrightarrow c_4 = 1. \tag{4.2-19}$$

Thus, the solution to the Euler equation is

$$x^*(t) = \sin(t). \tag{4.2-20}$$

The problem, as stated, has been solved, but let us investigate the increment for a neighboring curve to see if x^* is a minimum. As a comparison curve, consider the family

$$\begin{aligned} x(t) &= \sin(t) + \alpha \sin(2t) \\ &= x^*(t) + \delta x(t), \end{aligned} \tag{4.2-21}$$

with α as a real constant. Several curves for various values of α are shown in Fig. 4-8. Observe that each δx curve goes through zero at $t = 0$ and at $t = \pi/2$; thus $x^* + \delta x$ satisfies the required boundary conditions.

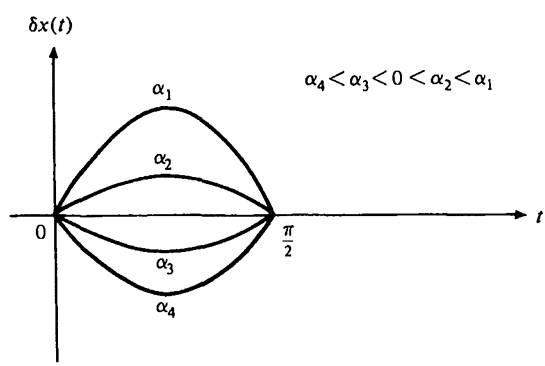


Figure 4-8 Several admissible δx curves

Substituting $x^*(t) = \sin(t)$ and $\dot{x}^*(t) = \cos(t)$ into the integrand of (4.2-11), we find that $J(x^*) = 0$. If $x(t) = \sin(t) + \alpha \sin(2t)$ and $\dot{x}(t) = \cos(t) + 2\alpha \cos(2t)$ are substituted into (4.2-11) and the integration performed, the result is

$$J(x^* + \delta x) = \left[\frac{3\pi}{4}\right]\alpha^2. \tag{4.2-22}$$

† $\Longrightarrow$ denotes "implies that."

Since $J(x^* + \delta x) > 0$ for all $\alpha \neq 0$, we conclude that

$$J(x^* + \delta x) > J(x^*) \quad \text{for } \alpha \neq 0. \tag{4.2-23}$$

What does this mean? It certainly indicates that x^* is not a maximizing curve, because we have just constructed a family of neighboring curves that gives larger values of J. Is x^* a minimizing curve? Our evidence is not conclusive, but it looks very much as if x^* does minimize J. We could try other neighboring curves to reinforce our suspicions, or else test x^* to see if it satisfies sufficient conditions for a minimum. Sufficient conditions for minima are beyond the scope of this book, so we shall content ourselves with investigating a few neighboring curves to ascertain whether a curve is maximal, minimal, or neither.

Now let us consider problems having end points that are not fixed. We shall consider only free end conditions at the final time; problems with unspecified boundary conditions at the initial time can be treated in a similar manner.

Final Time Specified, $x(t_f)$ Free

Problem 2: Find a necessary condition for a function to be an extremal for the functional

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\, dt; \tag{4.2-24}$$

t_0, $x(t_0)$, and t_f are specified, and $x(t_f)$ is free. The admissible curves all begin at the same point and terminate on a vertical line, as, for example, is the case in Fig. 4-9. To use the fundamental theorem, we first find the variation as in *Problem 1.* After integrating by parts, we have [see Eq. (4.2-7)]

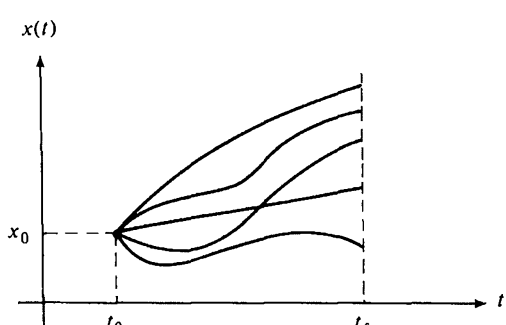


Figure 4-9 Several admissible curves for *Problem 2*

$$\delta J(x, \delta x) = \left[\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)\right] \delta x(t)\Big|_{t_0}^{t_f} + \int_{t_0}^{t_f} \left\{\left[\frac{\partial g}{\partial x}(x(t), \dot{x}(t), t)\right] - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)\right]\right\} \delta x(t)\, dt. \tag{4.2-25}$$

Now $\delta x(t_0) = 0$ for all admissible curves, but $\delta x(t_f)$ is arbitrary.

For an extremal x^*, we know that $\delta J(x^*, \delta x)$ must be zero. Let us next show that the integral in (4.2-25) must be zero on an extremal. *Suppose that* the curve x^* shown in Fig. 4-10 is an extremal for the free end point problem.

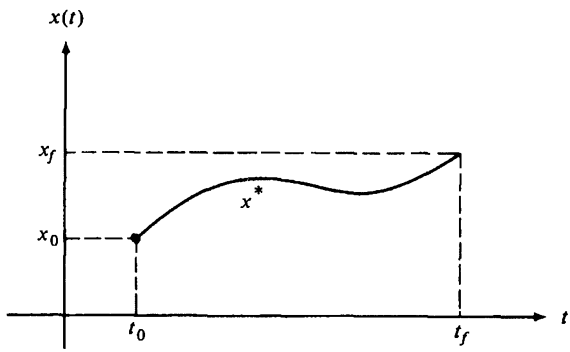


Figure 4-10 An extremal for a free end point problem

The value of $x^*(t_f)$ is x_f. Now consider a fixed end point problem with the same functional, the same initial and final times, and with *specified end points* $x(t_0) = x_0$ and $x(t_f) = x_f$ that are the same as for the extremal x^* in the free end point problem. The curve x^* in Fig. 4-10 must be an extremal for this fixed end point problem; therefore, x^* must be a solution of the Euler equation (4.2-10), and the integral term must be zero on an extremal. In other words, an extremal for a free end point problem is also an extremal for the fixed end point problem with the same end points, and the same functional; thus, *regardless of the boundary conditions, the Euler equation must be satisfied.*

Since

$$\delta J(x^*, \delta x) = 0, \quad \text{and} \quad \frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right] = 0$$

for all $t \in [t_0, t_f]$, from Eq. (4.2-25) we have

$$\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \delta x(t_f) = 0. \tag{4.2-26}$$

But since $x(t_f)$ is free, $\delta x(t_f)$ is arbitrary; therefore, it is necessary that

$$\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f) = 0. \tag{4.2-27}$$

The Euler equation is second order, and Eq. (4.2-27) provides the second required boundary condition [$x(t_0) = x_0$ is the other boundary condition]. We shall call Eq. (4.2-27) the *natural boundary condition;* notice that again we are confronted by a problem with split boundary values.

Example 4.2-2. Determine the smooth curve of smallest length connecting the point $x(0) = 1$ to the line $t = 5$.

It can be shown that the length of a curve lying in the $t - x(t)$ plane, with $t_0 = 0$ and $t_f = 5$, is

$$J(x) = \int_0^5 [1 + \dot{x}^2(t)]^{1/2}\, dt. \tag{4.2-28}$$

The Euler equation

$$-\frac{d}{dt}\left[\frac{\dot{x}^*(t)}{[1 + \dot{x}^{*2}(t)]^{1/2}}\right] = 0 \tag{4.2-29}$$

reduces to

$$\ddot{x}^*(t) = 0, \tag{4.2-30}$$

which has the solution

$$x^*(t) = c_1 t + c_2, \tag{4.2-31}$$

where c_1 and c_2 are constants of integration. $x^*(0) = 1$, so from (4.2-31) we have $c_2 = 1$. From Eq. (4.2-27),

$$\frac{\dot{x}^*(5)}{[1 + \dot{x}^{*2}(5)]^{1/2}} = 0, \tag{4.2-32}$$

which implies that $\dot{x}^*(5) = 0$. Substituting $\dot{x}^*(5) = 0$ into the equation

$$\dot{x}^*(t) = c_1, \tag{4.2-33}$$

obtained by differentiating (4.2-31), gives $c_1 = 0$. The solution then is

$$x^*(t) = 1, \tag{4.2-34}$$

a straight line parallel to the t axis.

Example 4.2-3. Determine an extremal for the functional

$$J(x) = \int_0^2 [\dot{x}^2(t) + 2x(t)\dot{x}(t) + 4x^2(t)]\,dt; \tag{4.2-35}$$

$x(0) = 1$, and $x(2)$ is free.

From (4.2-10) the Euler equation is

$$-\ddot{x}^*(t) + 4x^*(t) = 0. \tag{4.2-36}$$

The solution has the form

$$x^*(t) = c_1\epsilon^{-2t} + c_2\epsilon^{2t}. \tag{4.2-37}$$

To evaluate the constants of integration, use the boundary condition $x(0) = 1$, and the natural boundary condition

$$\frac{\partial g}{\partial \dot{x}}(x^*(2), \dot{x}^*(2)) = 0. \tag{4.2-38}$$

Equation (4.2-38) gives

$$\dot{x}^*(2) + x^*(2) = 0, \tag{4.2-39}$$

and from Eq. (4.2-37) we find that

$$\dot{x}^*(t) = -2c_1\epsilon^{-2t} + 2c_2\epsilon^{2t}. \tag{4.2-40}$$

Evaluating (4.2-37) and (4.2-40) with $t = 2$ and substituting in Eq. (4.2-39) we obtain

$$-c_1\epsilon^{-4} + 3c_2\epsilon^4 = 0. \tag{4.2-41}$$

The boundary value $x(0) = 1$ provides the equation

$$c_1 + c_2 = 1. \tag{4.2-42}$$

Solving these simultaneous algebraic equations for c_1 and c_2 yields

$$c_1 = \frac{3\epsilon^4}{\epsilon^{-4} + 3\epsilon^4}, \quad \text{and} \quad c_2 = \frac{\epsilon^{-4}}{\epsilon^{-4} + 3\epsilon^4}.$$

The final time was fixed in *Problems 1* and *2*; consequently, the variations of the functionals involved two integrals having the same limits of integration. If the final time is free, however, this is no longer the case; therefore, let us now generalize the results of our previous discussion. This is accomplished by separating the total variation of a functional into two

partial variations: the variation resulting from the difference $\delta x(t)$ in the interval $[t_0, t_f]$ and the variation resulting from the difference in end points of two curves. The sum of these two variations is called the *general variation* of a functional. First, let us consider the case where $x(t_f)$ is specified.

Final Time Free, $x(t_f)$ Specified

In *Problem 2* we considered the situation where $x(t_f)$ was free, but the final time t_f was specified. Let us now investigate problems in which $x(t_f)$ is specified, but t_f is free.

Problem 3: Find a necessary condition that must be satisfied by an extremal of the functional

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\, dt; \tag{4.2-43}$$

t_0, $x(t_0) = x_0$, and $x(t_f) = x_f$ are specified, and t_f is free.

The admissible curves, several of which are shown in Fig. 4-11, all begin at the point (x_0, t_0) and terminate on the horizontal line with ordinate x_f.

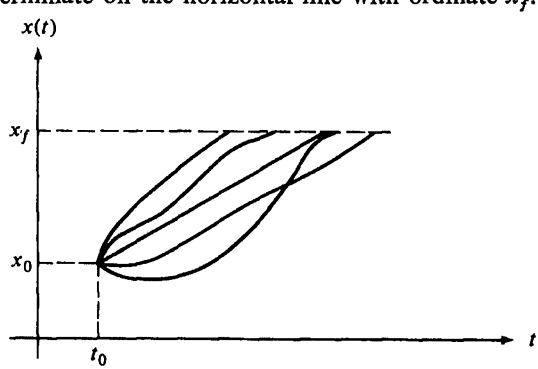


Figure 4-11 Several admissible curves for *Problem 3*

Because of the free final time, the development in *Problem 2* must be modified. In Fig. 4-12 an extremal curve x^*, terminating at the point (x_f, t_f), and a neighboring comparison curve x, terminating at the point $(x_f, t_f + \delta t_f)$, are shown.

From Fig. 4-12 it is apparent that $\delta x(t) = [x(t) - x^*(t)]$ has meaning only in the interval $[t_0, t_f]$, since x^* is not defined for $t \in (t_f, t_f + \delta t_f]$.†

† $t \in (t_f, t_f + \delta t_f]$ means $t_f < t \leq t_f + \delta t_f$.

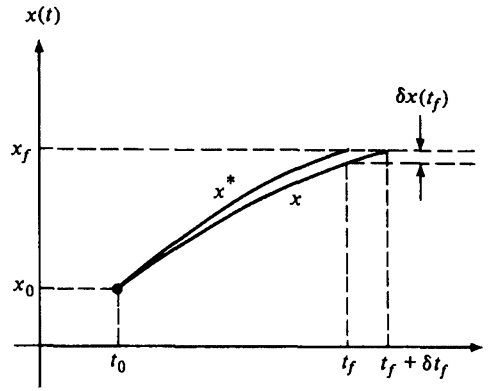


Figure 4-12 An extremal, x^*, and a neighboring comparison curve, x

First, we form the increment

$$\begin{aligned}\Delta J &= \int_{t_0}^{t_f+\delta t_f} g(x(t), \dot{x}(t), t)\, dt - \int_{t_0}^{t_f} g(x^*(t), \dot{x}^*(t), t)\, dt \\ &= \int_{t_0}^{t_f} \{g(x(t), \dot{x}(t), t) - g(x^*(t), \dot{x}^*(t), t)\}\, dt \\ &\quad + \int_{t_f}^{t_f+\delta t_f} g(x(t), \dot{x}(t), t)\, dt,\end{aligned} \tag{4.2-44}$$

or

$$\begin{aligned}\Delta J = \int_{t_0}^{t_f} &\{g(x^*(t) + \delta x(t), \dot{x}^*(t) + \delta \dot{x}(t), t) \\ &- g(x^*(t), \dot{x}^*(t), t)\}\, dt \\ &+ \int_{t_f}^{t_f+\delta t_f} g(x(t), \dot{x}(t), t)\, dt.\end{aligned} \tag{4.2-45}$$

The first integrand can be expanded about $x^*(t)$, $\dot{x}^*(t)$ in a Taylor series to give

$$\begin{aligned}\Delta J = \int_{t_0}^{t_f} &\left\{\left[\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t)\right] \delta x(t) \right. \\ &\left. + \left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right] \delta \dot{x}(t)\right\} dt \\ &+ o(\delta x(t), \delta \dot{x}(t)) + \int_{t_f}^{t_f+\delta t_f} g(x(t), \dot{x}(t), t)\, dt.\dagger\end{aligned} \tag{4.2-46}$$

† $o(\delta x(t), \delta \dot{x}(t))$ denotes terms of higher than first order in $\delta x(t)$ and $\delta \dot{x}(t)$; subsequently we will write simply $o(\cdot)$.

The second integral can be written

$$\int_{t_f}^{t_f+\delta t_f} g(x(t), \dot{x}(t), t)\, dt = [g(x(t_f), \dot{x}(t_f), t_f)]\, \delta t_f + o(\delta t_f). \quad (4.2\text{-}47)$$

Integrating by parts the term in Eq. (4.2-46) containing $\delta\dot{x}(t)$, and substituting (4.2-47), we obtain

$$\begin{aligned}\Delta J = &\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \delta x(t_f) + [g(x(t_f), \dot{x}(t_f), t_f)]\, \delta t_f \\ &+ \int_{t_0}^{t_f} \left\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t)\right. \\ &\left. - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]\right\} \delta x(t)\, dt + o(\cdot),\end{aligned} \quad (4.2\text{-}48)$$

where we have also used the fact that $\delta x(t_0) = 0$. Next, we shall express $g(x(t_f), \dot{x}(t_f), t_f)$ in terms of $g(x^*(t_f), \dot{x}^*(t_f), t_f)$ by the expansion

$$\begin{aligned}g(x(t_f), \dot{x}(t_f), t_f) = &\, g(x^*(t_f), \dot{x}^*(t_f), t_f) \\ &+ \left[\frac{\partial g}{\partial x}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \delta x(t_f) \\ &+ \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \delta\dot{x}(t_f) + o(\cdot).\end{aligned} \quad (4.2\text{-}49)$$

Substituting this expression in Eq. (4.2-48) yields

$$\begin{aligned}\Delta J = &\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \delta x(t_f) + [g(x^*(t_f), \dot{x}^*(t_f), t_f)]\, \delta t_f \\ &+ \int_{t_0}^{t_f} \left\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t)\right. \\ &\left. - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]\right\} \delta x(t)\, dt + o(\cdot).\end{aligned} \quad (4.2\text{-}50)$$

$\delta x(t_f)$, which is neither zero nor free, depends on δt_f. The variation of J, δJ, consists of the first-order terms in the increment ΔJ; therefore, the dependence of $\delta x(t_f)$ on δt_f must be linearly approximated. By inspection of Fig. 4-12 we have

$$\delta x(t_f) + \dot{x}^*(t_f)\, \delta t_f \doteq 0\dagger \quad (4.2\text{-}51)$$

or

$$\delta x(t_f) \doteq -\dot{x}^*(t_f)\, \delta t_f. \quad (4.2\text{-}52)$$

† $\doteq$ means "equal to first order."

Substituting (4.2-52) into Eq. (4.2-50), and retaining only first-order terms, we have the variation

$$\begin{aligned}\delta J(x^*, \delta x) = 0 = \Big\{\Big[-\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\Big]\dot{x}^*(t_f) \\ + g(x^*(t_f), \dot{x}^*(t_f), t_f)\Big\}\delta t_f \\ + \int_{t_0}^{t_f}\Big\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) \\ - \frac{d}{dt}\Big[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\Big]\Big\}\delta x(t)\, dt.\end{aligned} \tag{4.2-53}$$

Notice that the integral term represents the partial variation of J caused by $\delta x(t)$, $t \in [t_0, t_f]$, and the term involving δt_f is the partial variation of J caused by the difference in end points; together, these partial variations make up the general (or total) variation.

As in *Problem 2*, we argue that the extremal for this free end point problem is also an extremal for a particular fixed end point problem; therefore, the Euler equation

$$\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}\Big[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\Big] = 0 \tag{4.2-54}$$

must be satisfied, and the integral is zero. δt_f is arbitrary, so its coefficient must be zero, and the required boundary condition at t_f is

$$g(x^*(t_f), \dot{x}^*(t_f), t_f) - \Big[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\Big]\dot{x}^*(t_f) = 0. \tag{4.2-55}$$

The following example illustrates the procedure for solving a problem with $x(t_f)$ specified and t_f free.

Example 4.2-4. Find an extremal for the functional

$$J(x) = \int_1^{t_f}[2x(t) + \tfrac{1}{2}\dot{x}^2(t)]\, dt; \tag{4.2-56}$$

the boundary conditions are $x(1) = 4$, $x(t_f) = 4$, and $t_f > 1$ is free.

The Euler equation

$$\ddot{x}^*(t) = 2 \tag{4.2-57}$$

has the solution

$$x^*(t) = t^2 + c_1 t + c_2. \tag{4.2-58}$$

t_f is unspecified, so the relationship

$$0 = g(x^*(t_f), \dot{x}^*(t_f), t_f) - \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right]\dot{x}^*(t_f) \\ = 2x^*(t_f) - \tfrac{1}{2}\dot{x}^{*2}(t_f) \tag{4.2-59}$$

must be satisfied. From (4.2-59) and the specified values of $x(1)$ and $x(t_f)$ we obtain

$$x^*(1) = 4 = 1 + c_1 + c_2, \text{ or } c_1 + c_2 = 3 \tag{4.2-60a}$$

$$x^*(t_f) = 4 = t_f^2 + c_1 t_f + c_2 \tag{4.2-60b}$$

$$2x^*(t_f) - \frac{1}{2}\dot{x}^{*2}(t_f) = 0 = 2c_2 - \frac{c_1^2}{2}. \tag{4.2-60c}$$

Solving Eqs. (4.2-60) for c_1, c_2, and t_f gives the extremal

$$x^*(t) = t^2 - 6t + 9, \quad \text{and} \quad t_f = 5. \tag{4.2-61}$$

Problems with Both the Final Time t_f and $x(t_f)$ Free

We are now ready to consider problems having *both* t_f and $x(t_f)$ unspecified. Not surprisingly, we shall find that the necessary conditions of *Problems 2* and *3* are included as special cases.

Problem 4: Find a necessary condition that must be satisfied by an extremal for a functional of the form

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\, dt; \tag{4.2-62}$$

t_0 and $x(t_0) = x_0$ are specified, and t_f and $x(t_f)$ are free.

Figure 4-13 shows an extremal x^* and an admissible comparison curve x.

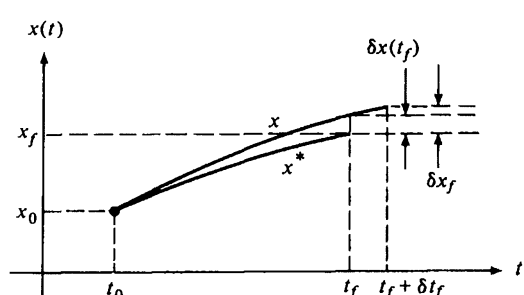


Figure 4-13 An extremal and a neighboring comparison curve for *Problem 4*

Notice that $\delta x(t_f)$ is the difference in ordinates at $t = t_f$ and δx_f is the difference in ordinates of the end points of the two curves. It is important to keep in mind that, in general, $\delta x(t_f) \neq \delta x_f$.

To use the fundamental theorem, we must first determine the variation by forming the increment. This is accomplished in exactly the same manner as in *Problem 3* as far as the equation

$$\begin{aligned}\Delta J = &\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \delta x(t_f) + [g(x^*(t_f), \dot{x}^*(t_f), t_f)]\, \delta t_f \\ &+ \int_{t_0}^{t_f} \left\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t)\right. \\ &\left. - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]\right\} \delta x(t)\, dt + o(\cdot).\end{aligned} \tag{4.2-50}$$

Next, we must relate $\delta x(t_f)$ to δt_f and δx_f. From Fig. 4-13 we have

$$\delta x_f \doteq \delta x(t_f) + \dot{x}^*(t_f)\, \delta t_f, \tag{4.2-63}$$

or

$$\delta x(t_f) = \delta x_f - \dot{x}^*(t_f)\, \delta t_f. \tag{4.2-64}$$

Substituting this in Eq. (4.2-50) and collecting terms, we obtain as the variation

$$\begin{aligned}\delta J(x^*, \delta x) = 0 = &\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \delta x_f \\ &+ \left[g(x^*(t_f), \dot{x}^*(t_f), t_f)\right. \\ &\left. - \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right] \dot{x}^*(t_f)\right] \delta t_f \\ &+ \int_{t_0}^{t_f} \left\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t)\right. \\ &\left. - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]\right\} \delta x(t)\, dt.\end{aligned} \tag{4.2-65}$$

As before, we argue that the Euler equation must be satisfied; therefore, the integral is zero. There may be a variety of end point conditions in practice; however, for the moment we shall consider only two possibilities:

1. t_f and $x(t_f)$ unrelated. In this case δx_f and δt_f are independent of one another and arbitrary, so their coefficients must each be zero. From Eq. (4.2-65), then,

$$\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f) = 0, \tag{4.2-66}$$

and

$$g(x^*(t_f), \dot{x}^*(t_f), t_f) - \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right]\dot{x}^*(t_f) = 0, \quad (4.2\text{-}67)$$

which together imply that

$$g(x^*(t_f), \dot{x}^*(t_f), t_f) = 0. \quad (4.2\text{-}68)$$

2. t_f and $x(t_f)$ related. For example, the final value of x may be constrained to lie on a specified moving point, $\theta(t)$; that is,

$$x(t_f) = \theta(t_f). \quad (4.2\text{-}69)$$

In this case the difference in end points δx_f is related to δt_f by

$$\delta x_f \doteq \frac{d\theta}{dt}(t_f)\,\delta t_f. \quad (4.2\text{-}70)$$

The geometric interpretation of this relationship is shown in Fig. 4-14. The distance a is a linear approximation to δx_f; that is,

$$a = \left[\frac{d\theta}{dt}(t_f)\right]\delta t_f. \quad (4.2\text{-}71)$$

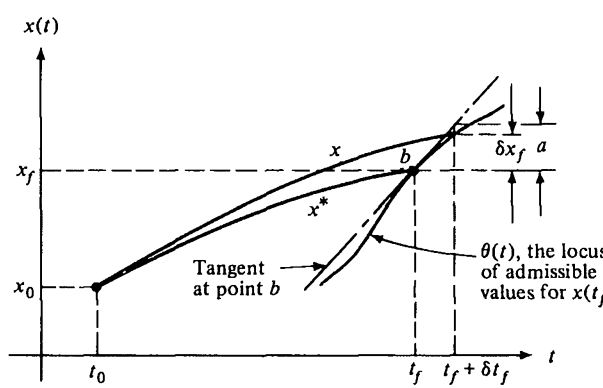


Figure 4-14 $x(t_f)$ and t_f free, but related

Substituting (4.2-70) into Eq. (4.2-65) and collecting terms gives

$$\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_f), \dot{x}^*(t_f), t_f)\right]\left[\frac{d\theta}{dt}(t_f) - \dot{x}^*(t_f)\right] + g(x^*(t_f), \dot{x}^*(t_f), t_f) = 0 \quad (4.2\text{-}72)$$

because δt_f is arbitrary. This equation is called the *transversality condition.*

In either of the two cases considered, integrating the Euler equation gives a solution $x^*(c_1, c_2, t)$, where c_1 and c_2 are constants of integration. c_1, c_2, and the unknown value of t_f can then be determined from $x^*(c_1, c_2, t_0) = x_0$ and Eqs. (4.2-66) and (4.2-68) if $x(t_f)$ and t_f are unrelated, or Eqs. (4.2-69) and (4.2-72) if $x(t_f)$ and t_f are related. Let us illustrate the use of these equations with the following examples.

Example 4.2-5. Find an extremal curve for the functional

$$J(x) = \int_{t_0}^{t_f} [1 + \dot{x}^2(t)]^{1/2}\, dt; \tag{4.2-73}$$

the boundary conditions $t_0 = 0$, $x(0) = 0$ are specified, t_f and $x(t_f)$ are free, but $x(t_f)$ is required to lie on the line

$$\theta(t) = -5t + 15. \tag{4.2-74}$$

The functional $J(x)$ is the length of the curve x; thus, the function that minimizes J is the shortest curve from the origin to the specified line. The Euler equation is

$$\frac{d}{dt}\left[\frac{\dot{x}^*(t)}{[1 + \dot{x}^{*2}(t)]^{1/2}}\right] = 0. \tag{4.2-75}$$

Performing the differentiation with respect to time and simplifying, we obtain

$$\ddot{x}^*(t) = 0, \tag{4.2-76}$$

which has the solution

$$x^*(t) = c_1 t + c_2. \tag{4.2-77}$$

We know that $x^*(0) = 0$, so $c_2 = 0$. To evaluate the other constant of integration, we use the transversality condition. From Eq. (4.2-72), since $x(t_f)$ and t_f are related,

$$\frac{\dot{x}^*(t_f)}{[1 + \dot{x}^{*2}(t_f)]^{1/2}} \cdot [-5 - \dot{x}^*(t_f)] + [1 + \dot{x}^{*2}(t_f)]^{1/2} = 0. \tag{4.2-78}$$

Simplifying, we have

$$-5\dot{x}^*(t_f) + 1 = 0, \tag{4.2-79}$$

from which, using Eq. (4.2-77), we obtain $c_1 = \frac{1}{5}$. The value of t_f,

found from

$$x^*(t_f) = \theta(t_f)$$
$$\tfrac{1}{5}t_f = -5t_f + 15, \tag{4.2-80}$$

is

$$t_f = \frac{75}{26} = 2.88. \tag{4.2-81}$$

Thus, the solution is

$$x^*(t) = \tfrac{1}{5}t. \tag{4.2-82}$$

Figure 4-15 shows what we knew all along: the shortest path is along the perpendicular to the line that passes through the origin.

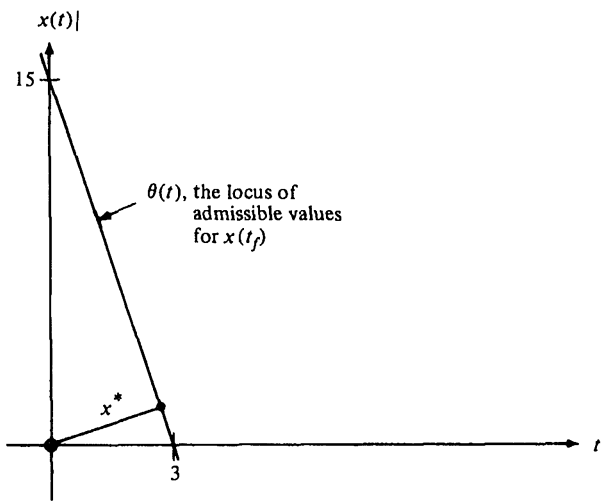


Figure 4-15 The extremal curve for Example 4.2-5

Example 4.2-6. Find an extremal for the functional in Eq. (4.2-73) which begins at the origin and terminates on the curve

$$\theta(t) = \tfrac{1}{2}[t - 5]^2 - \tfrac{1}{2}. \tag{4.2-83}$$

The Euler equation and its solution are the same as in the previous example, and since $x^*(0) = 0$ we again have $c_2 = 0$. From Eq. (4.2-72) the transversality condition is

$$\frac{\dot{x}^*(t_f)}{[1 + \dot{x}^{*2}(t_f)]^{1/2}} \cdot [t_f - 5 - \dot{x}^*(t_f)] + [1 + \dot{x}^{*2}(t_f)]^{1/2} = 0. \tag{4.2-84}$$

Simplifying, and substituting $\dot{x}^*(t_f) = c_1$, we obtain

$$c_1[t_f - 5] + 1 = 0. \tag{4.2-85}$$

Equating $x^*(t_f)$ and $\theta(t_f)$ yields

$$c_1 t_f = \tfrac{1}{2}[t_f - 5]^2 - \tfrac{1}{2}. \tag{4.2-86}$$

Solving the simultaneous equations (4.2-85) and (4.2-86), we find that $c_1 = \frac{1}{2}$ and $t_f = 3$, so the solution is

$$x^*(t) = \tfrac{1}{2}t. \tag{4.2-87}$$

Summary

We have now progressed from "the simplest variational problem" to problems having rather general boundary conditions. The key equation is (4.2-65), because from it we can deduce all of the results we have obtained so far. We have found that *regardless of the boundary conditions, the Euler equation must be satisfied;* thus, the integral term of (4.2-65) will be zero. If t_f and $x(t_f)$ are specified (*Problem 1*), then $\delta t_f = 0$ and $\delta x_f = \delta x(t_f) = 0$ in Eq. (4.2-65). To obtain the boundary condition equations for *Problem 2* [t_f specified, $x(t_f)$ free], simply let $\delta t_f = 0$ and $\delta x_f = \delta x(t_f)$ in (4.2-65). Similarly, to obtain the equations of *Problem 3*, substitute $\delta x_f = 0$ in Eq. (4.2-65).

Since the equations obtained for *Problems 1* through *3* can be obtained as special cases of Eq. (4.2-65), we suggest that the reader now consider the results of *Problem 4* as the starting point for solving problems of any of the foregoing types.

4.3 FUNCTIONALS INVOLVING SEVERAL INDEPENDENT FUNCTIONS

So far, the functionals considered have contained only a single function and its first derivative. We now wish to generalize our discussion to include functionals that may contain several *independent* functions and their first derivatives. We shall draw heavily on the results of Section 4.2—in fact, our terminal point will be the matrix version of Eq. (4.2-65).

Problems with Fixed End Points

Problem 1a: Consider the functional

$$J(x_1, x_2, \ldots, x_n) = \int_{t_0}^{t_f} g(x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t), t)\, dt, \tag{4.3-1}$$

where $x_1, x_2, \ldots, x_n$ are independent functions with continuous first derivatives, and g has continuous first and second partial derivatives with respect to all of its arguments. t_0 and t_f are specified, and the boundary conditions are

$$\begin{aligned} x_1(t_0) &= x_{1_0}; & x_1(t_f) &= x_{1_f}; \\ &\cdot & &\cdot \\ &\cdot & &\cdot \\ &\cdot & &\cdot \\ x_n(t_0) &= x_{n_0}; & x_n(t_f) &= x_{n_f}. \end{aligned}$$

We wish to use the fundamental theorem to determine a necessary condition for the functions $x_1^*, x_2^*, \ldots, x_n^*$ to be extremal.

To begin, we find the variation of J by introducing variations in $x_1, \ldots, x_n$, determining the increment, and retaining only the first-order terms:

$$\begin{aligned} \Delta J = \int_{t_0}^{t_f} \{ & g(x_1(t) + \delta x_1(t), \ldots, x_n(t) + \delta x_n(t), \\ & \dot{x}_1(t) + \delta \dot{x}_1(t), \ldots, \dot{x}_n(t) + \delta \dot{x}_n(t), t) \\ & - g(x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t), t)\} \, dt. \end{aligned} \tag{4.3-2}$$

Expanding in a Taylor series about $x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t)$ gives

$$\begin{aligned} \Delta J = \int_{t_0}^{t_f} \Bigg\{ & \sum_{i=1}^{n} \left[\left[\frac{\partial g}{\partial x_i}(x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t), t) \right] \delta x_i(t) \right] \\ & + \sum_{i=1}^{n} \left[\left[\frac{\partial g}{\partial \dot{x}_i}(x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t), t) \right] \delta \dot{x}_i(t) \right] \\ & + \sum_{i=1}^{n} [\text{terms of higher order in } \delta x_i(t), \delta \dot{x}_i(t)] \Bigg\} \, dt. \end{aligned} \tag{4.3-3}$$

The variation δJ is determined by retaining only the terms that are linear in δx_i and $\delta \dot{x}_i$. To eliminate the dependence of δJ on $\delta \dot{x}_i$, we integrate by parts the terms containing $\delta \dot{x}_i$ to obtain

$$\begin{aligned} \delta J = & \sum_{i=1}^{n} \left[\left[\frac{\partial g}{\partial \dot{x}_i}(x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t), t) \right] \delta x_i(t) \right]_{t_0}^{t_f} \\ & + \int_{t_0}^{t_f} \Bigg\{ \sum_{i=1}^{n} \left[\left[\frac{\partial g}{\partial x_i}(x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t), t) \right] \right. \\ & \left. - \frac{d}{dt} \left[\frac{\partial g}{\partial \dot{x}_i}(x_1(t), \ldots, x_n(t), \dot{x}_1(t), \ldots, \dot{x}_n(t), t) \right] \delta x_i(t) \right] \Bigg\} \, dt. \end{aligned} \tag{4.3-4}$$

Since the boundary conditions for all of the x_i's are fixed at t_0 and t_f, $\delta x_i(t_0) = 0$ and $\delta x_i(t_f) = 0$ $(i = 1, \ldots, n)$, and the terms outside the integral vanish. On an extremal [add*'s to the arguments in (4.3-4)], the variation

must be zero. The δx_i's are independent; let us select all of the δx_i's except δx_1 to be zero. Then

$$\delta J = 0 = \int_{t_0}^{t_f} \left\{ \frac{\partial g}{\partial x_1}(x_1^*(t), \ldots, x_n^*(t), \dot{x}_1^*(t), \ldots, \dot{x}_n^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}_1}(x_1^*(t), \ldots, x_n^*(t), \dot{x}_1^*(t), \ldots, \dot{x}_n^*(t), t)\right] \right\} \delta x_1(t)\, dt. \tag{4.3-5}$$

But δx_1 can assume arbitrary values as long as it is zero at the end points t_0 and t_f; therefore, the fundamental lemma applies, and the coefficient of $\delta x_1(t)$ must be zero everywhere in the interval $[t_0, t_f]$. Repeating this argument for each of the δx_i's in turn gives

$$\frac{\partial g}{\partial x_i}(x_1^*(t), \ldots, x_n^*(t), \dot{x}_1^*(t), \ldots, \dot{x}_n^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}_i}(x_1^*(t), \ldots, x_n^*(t), \dot{x}_1^*(t), \ldots, \dot{x}_n^*(t), t)\right] = 0 \quad \text{for all } t \in [t_0, t_f] \quad \text{and} \quad i = 1, \ldots, n. \tag{4.3-6}$$

We now have n Euler equations. Notice that the same adjectives apply to these equations as in *Problem 1*; that is, each equation is, in general, a nonlinear, ordinary, hard-to-solve, second-order differential equation with split boundary values. The situation is further complicated by the fact that these differential equations are simultaneous—each differential equation generally contains terms involving all of the functions and their first and second derivatives.

Throughout the preceding development we have painfully (very!) written out each of the arguments. It is much more convenient and compact to use matrix notation; in the future we shall do so. To gain familiarity with the notation, let us re-derive the preceding equations using vector-matrix notation. Starting with the problem statement, we have

$$J(\mathbf{x}) = \int_{t_0}^{t_f} g(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)\, dt \tag{4.3-1a}$$

and the boundary conditions $\mathbf{x}(t_0) = \mathbf{x}_0$, $\mathbf{x}(t_f) = \mathbf{x}_f$, where

$$\mathbf{x}(t) \triangleq \begin{bmatrix} x_1(t) \\ \cdot \\ \cdot \\ \cdot \\ x_n(t) \end{bmatrix} \quad \text{and} \quad \dot{\mathbf{x}}(t) \triangleq \begin{bmatrix} \frac{d}{dt}x_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \frac{d}{dt}x_n(t) \end{bmatrix}.$$

The expression for the increment becomes

$$\Delta J = \int_{t_0}^{t_f} \{g(\mathbf{x}(t) + \delta\mathbf{x}(t), \dot{\mathbf{x}}(t) + \delta\dot{\mathbf{x}}(t), t) - g(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)\}\, dt, \quad (4.3\text{-}2a)$$

which after expansion is

$$\Delta J = \int_{t_0}^{t_f} \left\{ \left[\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)\right]^T \delta\mathbf{x}(t) + \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)\right]^T \delta\dot{\mathbf{x}}(t) + [\text{terms of higher order in } \delta\mathbf{x}(t), \delta\dot{\mathbf{x}}(t)] \right\} dt, \quad (4.3\text{-}3a)$$

where

$$\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t) \triangleq \left[\frac{\partial g}{\partial x_1}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t), \ldots, \frac{\partial g}{\partial x_n}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)\right]^T$$

(an $n \times 1$ column matrix), and similarly for $\partial g/\partial \dot{\mathbf{x}}$. Discarding terms that are nonlinear in $\delta\mathbf{x}(t)$ and $\delta\dot{\mathbf{x}}(t)$ and integrating by parts, we have

$$\begin{aligned} \delta J(\mathbf{x}, \delta\mathbf{x}) = {} & \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}(t_f), \dot{\mathbf{x}}(t_f), t_f)\right]^T \cancelto{\mathbf{0}}{\delta\mathbf{x}(t_f)} \\ & - \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}(t_0), \dot{\mathbf{x}}(t_0), t_0)\right]^T \cancelto{\mathbf{0}}{\delta\mathbf{x}(t_0)} \\ & + \int_{t_0}^{t_f} \left\{ \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t) \right. \\ & \left. - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)\right] \right\}^T \delta\mathbf{x}(t)\, dt. \end{aligned} \quad (4.3\text{-}4a)$$

$\mathbf{0}$ is an $n \times 1$ matrix of zeros. Finally, the matrix representation of the Euler equations is

$$\boxed{\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)\right] = \mathbf{0};} \quad (4.3\text{-}6a)$$

Notice that Eq. (4.2-10), obtained previously, is the special case that results when $\mathbf{x}$ is a scalar.

Example 4.3-1. Find the Euler equations for the functional

$$J(\mathbf{x}) = \int_{t_0}^{t_f} [x_1^2(t)x_2(t) + t^2\dot{x}_1^2(t) - \dot{x}_2^2(t)\dot{x}_1(t)]\, dt; \quad (4.3\text{-}7)$$

the end points t_0, t_f, $x_1(t_0)$, $x_2(t_0)$, $x_1(t_f)$, and $x_2(t_f)$ are specified.

The Euler equations are given in Eq. (4.3-6a); writing out the indicated derivatives gives

$$\begin{aligned} 0 &= \frac{\partial g}{\partial x_1}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}_1}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)\right] \\ &= 2x_1^*(t)x_2^*(t) - \frac{d}{dt}[2t^2\dot{x}_1^*(t) - \dot{x}_2^{*2}(t)] \\ &= 2x_1^*(t)x_2^*(t) - 4t\dot{x}_1^*(t) - 2t^2\ddot{x}_1^*(t) + 2\dot{x}_2^*(t)\ddot{x}_2^*(t), \end{aligned} \tag{4.3-8}$$

and

$$\begin{aligned} 0 &= \frac{\partial g}{\partial x_2}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}_2}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)\right] \\ &= x_1^{*2}(t) - \frac{d}{dt}[-2\dot{x}_2^*(t)\dot{x}_1^*(t)] \\ &= x_1^{*2}(t) + 2\dot{x}_2^*(t)\ddot{x}_1^*(t) + 2\ddot{x}_2^*(t)\dot{x}_1^*(t). \end{aligned} \tag{4.3-9}$$

These differential equations are nonlinear and have time-varying coefficients.

Example 4.3-2. Find an extremal curve for the functional

$$J(\mathbf{x}) = \int_0^{\pi/4} [x_1^2(t) + 4x_2^2(t) + \dot{x}_1(t)\dot{x}_2(t)]\,dt \tag{4.3-10}$$

which satisfies the boundary conditions

$$\mathbf{x}(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}\left(\frac{\pi}{4}\right) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The Euler equations, found from (4.3-6a),

$$2x_1^*(t) - \ddot{x}_2^*(t) = 0 \tag{4.3-11a}$$

$$8x_2^*(t) - \ddot{x}_1^*(t) = 0, \tag{4.3-11b}$$

are linear, time-invariant, and homogeneous. Solving these equations by classical methods (or Laplace transforms) gives

$$x_1^*(t) = c_1\epsilon^{2t} + c_2\epsilon^{-2t} + c_3 \cos 2t + c_4 \sin 2t, \tag{4.3-12}$$

where c_1, c_2, c_3, and c_4 are constants of integration. Differentiating $x_1^*(t)$ twice and substituting into Eq. (4.3-11b) gives

$$x_2^*(t) = \tfrac{1}{2}c_1\epsilon^{2t} + \tfrac{1}{2}c_2\epsilon^{-2t} - \tfrac{1}{2}c_3 \cos 2t - \tfrac{1}{2}c_4 \sin 2t. \tag{4.3-13}$$

Putting $t = 0$ and $t = \pi/4$ in (4.3-12) and (4.3-13), we obtain four equations and four unknowns; that is,

$$x_1^*(0) = 0; \qquad x_2^*(0) = 1; \qquad x_1^*\left(\frac{\pi}{4}\right) = 1; \qquad x_2^*\left(\frac{\pi}{4}\right) = 0.$$

Solving these equations for the constants of integration yields

$$c_1 = \frac{-\frac{1}{2} + \epsilon^{-\pi/2}}{\epsilon^{-\pi/2} - \epsilon^{\pi/2}}; \qquad c_2 = \frac{\frac{1}{2} - \epsilon^{\pi/2}}{\epsilon^{-\pi/2} - \epsilon^{\pi/2}}; \qquad c_3 = -1; \qquad c_4 = \frac{1}{2}.$$

Problems with Free End Points

Problem 4a: Consider the functional

$$J(\mathbf{x}) = \int_{t_0}^{t_f} g(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)\, dt, \tag{4.3-14}$$

where $\mathbf{x}$ and g satisfy the continuity and differentiability requirements of *Problem 1a*. $\mathbf{x}(t_0)$ and t_0 are specified; $\mathbf{x}(t_f)$ and t_f are free. Find a necessary condition that must be satisfied by an extremal.

To obtain the generalized variation, we proceed in exactly the same manner as in *Problem 4* of Section 4.2. The only change is that now we are dealing with vector functions. Forming the increment, integrating by parts the term involving $\delta\dot{\mathbf{x}}(t)$, retaining terms of first order, and relating $\delta\mathbf{x}(t_f)$ to $\delta\mathbf{x}_f$ and δt_f [see Fig. 4-13 and Eq. (4.2-64)] by

$$\delta\mathbf{x}(t_f) = \delta\mathbf{x}_f - \dot{\mathbf{x}}^*(t_f)\, \delta t_f, \tag{4.3-15}$$

we obtain for the variation

$$\begin{aligned} \delta J(\mathbf{x}^*, \delta\mathbf{x}) = 0 = &\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \delta\mathbf{x}_f \\ &+ \Big[g(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f) \\ &- \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \dot{\mathbf{x}}^*(t_f)\Big] \delta t_f \\ &+ \int_{t_0}^{t_f} \Big\{\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) \\ &- \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)\right]\Big\}^T \delta\mathbf{x}(t)\, dt. \end{aligned} \tag{4.3-16}$$

As before, we argue that an extremal for this free end point problem must also be an extremal for a certain fixed end point problem; therefore, $\mathbf{x}^*$ must be a solution of the Euler equations

$$\boxed{\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)\right] = \mathbf{0}.} \tag{4.3-17}$$

The boundary conditions at the final time are then specified by the relationship

$$\begin{aligned}\delta J(\mathbf{x}^*, \delta\mathbf{x}) = 0 = &\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \delta\mathbf{x}_f \\ &+ \Big[g(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f) \\ &- \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \dot{\mathbf{x}}^*(t_f)\Big] \delta t_f.\end{aligned} \tag{4.3-18}$$

Equations (4.3-17) and (4.3-18) are the key equations, because they summarize necessary conditions that must be satisfied by an extremal curve. The boundary condition equations are obtained by making the appropriate substitutions in Eq. (4.3-18). The equations obtained by making these substitutions, which are contained in Table 4-1, are simply the vector analogs of the equations derived in Section 4.2. Notice that regardless of the problem specifications the boundary conditions are always split; thus, to find an optimal trajectory, in general, a nonlinear, two-point boundary-value problem must be solved.

Situations not included in Table 4-1 may arise; however, these can be handled by returning to Eq. (4.3-18). For example, suppose that t_f is fixed, $x_i(t_f)$, $i = 1, 2, \ldots, r$ are specified, and $x_j(t_f), j = r + 1, \ldots, n$ are free. In this case, the appropriate substitutions are

$$\begin{aligned}\delta t_f &= 0; \\ \delta x_i(t_f) &= 0, \qquad i = 1, 2, \ldots, r; \\ \delta x_j(t_f) &\text{ arbitrary}, \qquad j = r + 1, \ldots, n.\end{aligned}$$

Let us now consider several examples that illustrate the use of Table 4-1 and the key equations (4.3-17) and (4.3-18).

Example 4.3-3. Find an extremal for the functional

$$J(\mathbf{x}) = \int_0^{\pi/4} [x_1^2(t) + \dot{x}_1(t)\dot{x}_2(t) + \dot{x}_2^2(t)]\, dt. \tag{4.3-19}$$

The functions x_1 and x_2 are independent, and the boundary conditions are

$$\begin{aligned}x_1(0) &= 1; \qquad x_1\left(\frac{\pi}{4}\right) = 2; \\ x_2(0) &= \frac{3}{2}; \qquad x_2\left(\frac{\pi}{4}\right) \text{ free.}\end{aligned}$$

The Euler equations are, from Eq. (4.3-17),

$$2x_1^*(t) - \dot{x}_2^*(t) = 0; \tag{4.3-20a}$$

$$-\dot{x}_1^*(t) - 2\dot{x}_2^*(t) = 0. \tag{4.3-20b}$$

Multiplying Eq. (4.3-20a) by 2 and subtracting (4.3-20b), we obtain

$$\ddot{x}_1^*(t) + 4x_1^*(t) = 0, \tag{4.3-21}$$

which has the solution

$$x_1^*(t) = c_1 \cos 2t + c_2 \sin 2t; \tag{4.3-22}$$

therefore,

$$\dot{x}_2^*(t) = 2c_1 \cos 2t + 2c_2 \sin 2t. \tag{4.3-23}$$

Integrating twice yields

$$x_2^*(t) = -\frac{c_1}{2} \cos 2t - \frac{c_2}{2} \sin 2t + c_3 t + c_4. \tag{4.3-24}$$

Notice that the boundary conditions are such that this problem does not fit into any of the categories of Table 4-1, so we return to Eq. (4.3-18). $x_1(t_f)$ is specified, which means that $\delta x_{1_f} = \delta x_1(t_f) = 0$. $x_2(t_f)$, however, is free, so $\delta x_2(t_f)$ is arbitrary. We also have that $\delta t_f = 0$ because t_f is specified. Making these substitutions in Eq. (4.3-18) gives

$$\left[\frac{\partial g}{\partial \dot{x}_2}\left(\mathbf{x}^*\left(\frac{\pi}{4}\right), \dot{\mathbf{x}}^*\left(\frac{\pi}{4}\right)\right)\right] \delta x_2(t_f) = 0, \tag{4.3-25}$$

which implies [since $\delta x_2(t_f)$ is arbitrary] that

$$\frac{\partial g}{\partial \dot{x}_2}\left(\mathbf{x}^*\left(\frac{\pi}{4}\right), \dot{\mathbf{x}}^*\left(\frac{\pi}{4}\right)\right) = 0. \tag{4.3-26}$$

But

$$\frac{\partial g}{\partial \dot{x}_2}\left(\mathbf{x}^*\left(\frac{\pi}{4}\right), \dot{\mathbf{x}}^*\left(\frac{\pi}{4}\right)\right) = \dot{x}_1^*\left(\frac{\pi}{4}\right) + 2\dot{x}_2^*\left(\frac{\pi}{4}\right) = 2 \cdot c_3,$$

so $c_3 = 0$. From the specified boundary conditions we have

$$x_1(0) = 1 = c_1 \cdot 1 + c_2 \cdot 0: \qquad c_1 = 1;$$

$$x_2(0) = \frac{3}{2} = -\frac{c_1}{2} \cdot 1 - \frac{c_2}{2} \cdot 0 + c_3 \cdot 0 + c_4: \quad c_4 = 1.5 + \frac{c_1}{2} = 2;$$

$$x_1\left(\frac{\pi}{4}\right) = c_1 \cdot 0 + c_2 \cdot 1 = 2: \qquad c_2 = 2.$$

Table 4-1 DETERMINATION OF BOUNDARY-VALUE RELATIONSHIPS

Problem description	*Substitution*	*Boundary conditions*	*Remarks*
1. $\mathbf{x}(t_f)$, t_f both specified (*Problem 1*)	$\delta\mathbf{x}_f = \delta\mathbf{x}(t_f) = \mathbf{0}$ $\delta t_f = 0$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\mathbf{x}^*(t_f) = \mathbf{x}_f$	$2n$ equations to determine $2n$ constants of integration
2. $\mathbf{x}(t_f)$ free; t_f specified (*Problem 2*)	$\delta\mathbf{x}_f = \delta\mathbf{x}(t_f)$ $\delta t_f = 0$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f) = \mathbf{0}$	$2n$ equations to determine $2n$ constants of integration
3. t_f free; $\mathbf{x}(t_f)$ specified (*Problem 3*)	$\delta\mathbf{x}_f = \mathbf{0}$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\mathbf{x}^*(t_f) = \mathbf{x}_f$ $g(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)$ $- \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \dot{\mathbf{x}}^*(t_f) = 0$	$(2n + 1)$ equations to determine $2n$ constants of integration and t_f
4. t_f, $\mathbf{x}(t_f)$ free and independent (*Problem 4*)	—	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f) = \mathbf{0}$ $g(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f) = 0$	$(2n + 1)$ equations to determine $2n$ constants of integration and t_f
5. t_f, $\mathbf{x}(t_f)$ free but related by $\mathbf{x}(t_f) = \boldsymbol{\theta}(t_f)$ (*Problem 4*)	$\delta\mathbf{x}_f = \frac{d\boldsymbol{\theta}}{dt}(t_f)\,\delta t_f$†	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\mathbf{x}^*(t_f) = \boldsymbol{\theta}(t_f)$ $g(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)$ $+ \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \left[\frac{d\boldsymbol{\theta}}{dt}(t_f) - \dot{\mathbf{x}}^*(t_f)\right] = 0$†	$(2n + 1)$ equations to determine $2n$ constants of integration and t_f

† $\frac{d\boldsymbol{\theta}}{dt}$ denotes the $n \times 1$ column vector $\left[\frac{d\theta_1}{dt} \quad \frac{d\theta_2}{dt} \quad \cdots \quad \frac{d\theta_n}{dt}\right]^T$.

The extremal curve is, then,

$$\begin{aligned} x_1^*(t) &= \cos 2t + 2 \sin 2t \\ x_2^*(t) &= -\tfrac{1}{2} \cos 2t - \sin 2t + 2. \end{aligned} \tag{4.3-27}$$

Example 4.3-4. Find the Euler equations for the functional

$$J(\mathbf{x}) = \int_0^{t_f} [x_1^3(t) + x_1^2(t) + 2\dot{x}_1(t)\,\dot{x}_2(t) + x_1(t)\,x_2^2(t)]\,dt, \tag{4.3-28}$$

and determine the relationships required to evaluate the constants of integration. The specified boundary conditions are

$$\mathbf{x}(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad \mathbf{x}(t_f) = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

and t_f is free. The functions x_1 and x_2 are independent.

From Eq. (4.3-17) the Euler equations are

$$\begin{aligned} 3x_1^{*2}(t) + 2x_1^*(t) + x_2^{*2}(t) - 2\ddot{x}_2^*(t) &= 0 \\ 2x_1^*(t)x_2^*(t) - 2\ddot{x}_1^*(t) &= 0. \end{aligned} \tag{4.3-29}$$

The solution of these two nonlinear second-order differential equations, $\mathbf{x}^*(c_1, c_2, c_3, c_4, t)$, will contain the four constants of integration, c_1, c_2, c_3, c_4. From the specified boundary conditions we have

$$\begin{aligned} x_1^*(c_1, c_2, c_3, c_4, 0) &= 2 \\ x_2^*(c_1, c_2, c_3, c_4, 0) &= 1 \\ x_1^*(c_1, c_2, c_3, c_4, t_f) &= -1 \\ x_2^*(c_1, c_2, c_3, c_4, t_f) &= 4, \end{aligned} \tag{4.3-30}$$

but since t_f is unspecified, there are five unknowns. The other relationship that must be satisfied is obtained from Eq. (4.3-18) with $\delta\mathbf{x}_f = \mathbf{0}$:

$$x_1^{*3}(t_f) + x_1^{*2}(t_f) - 2\dot{x}_1^*(t_f)\dot{x}_2^*(t_f) + x_1^*(t_f)x_2^{*2}(t_f) = 0. \tag{4.3-31}$$

Thus, to determine c_1, c_2, c_3, c_4, and t_f the (nonlinear) algebraic equations (4.3-30) and (4.3-31) would have to be solved.

Example 4.3-5. Find the equation of the curve that is an extremal for the functional

$$J(x) = \int_0^{t_f} [t\dot{x}(t) + \dot{x}^2(t)]\,dt \qquad (t_f > 0) \tag{4.3-32}$$

for the boundary conditions specified below.

From (4.3-17) the Euler equation is

$$-\frac{d}{dt}[t + 2\dot{x}^*(t)] = 0, \tag{4.3-33}$$

or

$$1 + 2\dot{x}^*(t) = 0. \tag{4.3-34}$$

The solution of this equation is

$$x^*(t) = -\tfrac{1}{4}t^2 + c_1 t + c_2. \tag{4.3-35}$$

(a) What is the extremal if the boundary conditions are $t_f = 1$, $x(0) = 1$, $x(1) = 2.75$?

$$\begin{aligned} x^*(0) &= 1 = c_2 \\ x^*(1) &= 2.75 = -0.25 + c_1 + c_2, \quad \text{and} \quad c_1 = 2, \end{aligned} \tag{4.3-36}$$

so

$$x^*(t) = -\tfrac{1}{4}t^2 + 2t + 1. \tag{4.3-37}$$

(b) Find the extremal curve if $x(0) = 1$, $t_f = 2$, and $x(2)$ is free.

Again we have

$$x^*(0) = 1, \quad \text{so} \quad c_2 = 1.$$

From entry 2 of Table 4-1,

$$\begin{aligned} t_f + 2\dot{x}^*(t_f) &= 0 \\ 2 + 2[-\tfrac{1}{2}(2) + c_1] &= 0; \end{aligned} \tag{4.3-38}$$

therefore, $c_1 = 0$, so

$$x^*(t) = -\tfrac{1}{4}t^2 + 1. \tag{4.3-39}$$

(c) Find the extremal curve if $x(0) = 1$, $x(t_f) = 5$, and t_f is free.

As before, $x^*(0) = 1$ implies that $c_2 = 1$. From entry 3 of Table 4-1

$$t_f[\dot{x}^*(t_f)] + \dot{x}^{*2}(t_f) - [t_f + 2\dot{x}^*(t_f)]\dot{x}^*(t_f) = 0 \tag{4.3-40}$$

or

$$[t_f + \dot{x}^*(t_f) - t_f - 2\dot{x}^*(t_f)]\dot{x}^*(t_f) = 0, \tag{4.3-41}$$

which implies that $\dot{x}^*(t_f) = 0$, so

$$-\tfrac{1}{2}t_f + c_1 = 0, \tag{4.3-42}$$

and

$$5 = -\tfrac{1}{4}t_f^2 + c_1 t_f + 1, \tag{4.3-43}$$

since $c_2 = 1$. Solving these equations simultaneously gives $t_f = 4$ and $c_1 = 2$; therefore,

$$x^*(t) = -\tfrac{1}{4}t^2 + 2t + 1. \tag{4.3-44}$$

Summary

In Sections 4.2 and 4.3 we have progressed from the very restricted problem of a functional of one function with fixed end points to a rather general problem in which there can be several (independent) functions and free end points. Equations (4.3-17) and (4.3-18) are the important equations, because from them we can obtain the necessary conditions derived for more restricted problems.

To recapitulate, we have found that:

1. *Regardless of the boundary conditions*, the Euler equations

$$\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)\right] = \mathbf{0} \tag{4.3-17}$$

must be satisfied.

2. The required boundary condition equations are found from the equation

$$\begin{aligned}\left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \delta\mathbf{x}_f + \Big[g(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f) \\ - \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), t_f)\right]^T \dot{\mathbf{x}}^*(t_f)\Big]\delta t_f = 0\end{aligned} \tag{4.3-18}$$

by making the appropriate substitutions for $\delta\mathbf{x}_f$ and δt_f.

4.4 PIECEWISE-SMOOTH EXTREMALS

In the preceding sections we have derived necessary conditions that must be satisfied by extremal curves. The admissible curves were assumed to be continuous and to have continuous first derivatives; that is, the admissible curves were *smooth*. This is a very restrictive requirement for many practical problems. For example, if a control signal is the output of a relay, we know that this signal will contain discontinuities and that when such a control discontinuity occurs, one or more of the components of $\dot{\mathbf{x}}(t)$ will be discontinuous. Thus, we wish to enlarge the class of admissible curves to include

functions that have only *piecewise-continuous* first derivatives; that is, $\dot{\mathbf{x}}$ will be continuous except at a finite number of times in the interval (t_0, t_f).† At a time when $\dot{\mathbf{x}}$ is discontinuous, $\mathbf{x}$ is said to have a *corner*. Let us begin by considering functionals involving only a single function.

The problem is to find a necessary condition that must be satisfied by extrema of the functional

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\, dt. \tag{4.4-1}$$

It is assumed that g has continuous first and second partial derivatives with respect to all of its arguments, and that t_0, t_f, $x(t_0)$, and $x(t_f)$ are specified. $\dot{x}$ is a piecewise-continuous function (or we say that x is a *piecewise-smooth* curve). Assume that $\dot{x}$ has a discontinuity at some point $t_1 \in (t_0, t_f)$; t_1 is not fixed, nor is it usually known in advance.

Let us first express the functional J as

$$\begin{aligned} J(x) &= \int_{t_0}^{t_1} g(x(t), \dot{x}(t), t)\, dt + \int_{t_1}^{t_f} g(x(t), \dot{x}(t), t)\, dt \\ &\triangleq J_1(x) + J_2(x). \end{aligned} \tag{4.4-2}$$

We assert that if x^* is a minimizing extremal for J, then $x^*(t)$, $t \in [t_0, t_1]$, is an extremal for J_1 and $x^*(t)$, $t \in [t_1, t_f]$, is an extremal for J_2. To show this, assume that the final segment of an extremal for J is known; that is, we know $x^*(t)$, $t \in [t_1, t_f]$. Then to minimize J, we seek a curve defined in the interval $[t_0, t_1]$ which minimizes J_1; this curve is, by definition, an extremal of J_1. Similarly, if $x^*(t)$, $t \in [t_0, t_1]$, is known, to minimize J we seek a curve that minimizes J_2—an extremal for J_2.

Figure 4-16 shows an extremal curve x^* and a neighboring comparison

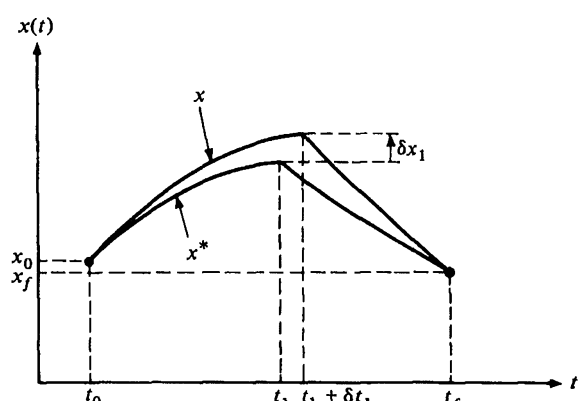


Figure 4-16 A piecewise-smooth extremal and a neighboring comparison curve

† The notation $t \in (t_0, t_f)$ means $t_0 < t < t_f$.

curve x. δt_1 and δx_1 are free, and from the fundamental theorem we know it is necessary that $\delta J(x^*, \delta x) = 0$. Since the coordinates of the corner point are free, we can use the results of *Problem 4* in Section 4.2 [see Eq. (4.2-65)] to obtain

$$\begin{aligned}
\delta J(x^*, \delta x) = 0 = &\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1^-), \dot{x}^*(t_1^-), t_1^-)\right] \delta x_1 + \Big\{g(x^*(t_1^-), \dot{x}^*(t_1^-), t_1^-) \\
&- \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1^-), \dot{x}^*(t_1^-), t_1^-)\right] \dot{x}^*(t_1^-)\Big\} \delta t_1 \\
&+ \int_{t_0}^{t_1} \left\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t)\right. \\
&\left. - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]\right\} \delta x(t)\, dt \\
&- \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1^+), \dot{x}^*(t_1^+), t_1^+)\right] \delta x_1 \qquad (4.4\text{-}3) \\
&- \Big\{g(x^*(t_1^+), \dot{x}^*(t_1^+), t_1^+) \\
&- \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1^+), \dot{x}^*(t_1^+), t_1^+)\right] \dot{x}^*(t_1^+)\Big\} \delta t_1 \\
&+ \int_{t_1}^{t_f} \left\{\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t)\right. \\
&\left. - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right]\right\} \delta x(t)\, dt.
\end{aligned}$$

δx_1 is the difference $x(t_1 + \delta t_1) - x^*(t_1)$, and t_1^- and t_1^+ denote the times just before and just after the discontinuity of $\dot{x}^*$. The terms that multiply δt_1 and δx_1 are due to the presence of t_1 as the upper limit of the first integral and as the lower limit of the second integral. We have shown that x^* is an extremal in both of the intervals $[t_0, t_1]$, and $[t_1, t_f]$; thus the Euler equation must be satisfied, and the integral terms are zero. In order that $\delta J(x^*, \delta x)$ be zero, it is then necessary that

$$\begin{aligned}
&\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^-), t_1) - \frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^+), t_1)\right] \delta x_1 \\
&+ \Big\{g(x^*(t_1), \dot{x}^*(t_1^-), t_1) - \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^-), t_1)\right] \dot{x}^*(t_1^-) \qquad (4.4\text{-}4) \\
&- g(x^*(t_1), \dot{x}^*(t_1^+), t_1) + \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^+), t_1)\right] \dot{x}^*(t_1^+)\Big\} \delta t_1 = 0.\dagger
\end{aligned}$$

If t_1 and $x(t_1)$ are unrelated, δx_1 and δt_1 are independently arbitrary, so their coefficients must each be zero and we have

† Notice that we have retained the t_1^+, t_1^- notation only where the distinction needs to be made.

$$\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^-), t_1) = \frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^+), t_1), \tag{4.4-5a}$$

and

$$\begin{aligned} &g(x^*(t_1), \dot{x}^*(t_1^-), t_1) - \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^-), t_1)\right]\dot{x}^*(t_1^-) \\ &= g(x^*(t_1), \dot{x}^*(t_1^+), t_1) - \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^+), t_1)\right]\dot{x}^*(t_1^+). \end{aligned} \tag{4.4-5b}$$

These two equations, called the *Weierstrass-Erdmann corner conditions*, are necessary conditions for an extremal. If there are several times $t_1, t_2, \ldots, t_r$ when corners exist, then at each such time these corner conditions must be satisfied.

It may be that $x(t_1)$ and t_1 are related by $x(t_1) = \theta(t_1)$. If so, δx_1 and δt_1 in Eq. (4.4-4) are not independently arbitrary; they are related by

$$\delta x_1 \doteq \frac{d\theta}{dt}(t_1)\,\delta t_1.\dagger \tag{4.4-6}$$

Substituting (4.4-6) into (4.4-4) and equating the coefficient of δt_1 equal to zero (since δt_1 is arbitrary), we obtain

$$\begin{aligned} &\left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^-), t_1)\right]\left[\frac{d\theta}{dt}(t_1) - \dot{x}^*(t_1^-)\right] + g(x^*(t_1), \dot{x}^*(t_1^-), t_1) \\ &= \left[\frac{\partial g}{\partial \dot{x}}(x^*(t_1), \dot{x}^*(t_1^+), t_1)\right]\left[\frac{d\theta}{dt}(t_1) - \dot{x}^*(t_1^+)\right] + g(x^*(t_1), \dot{x}^*(t_1^+), t_1). \end{aligned} \tag{4.4-7}$$

The extension of the Weierstrass-Erdmann corner conditions to the case where J involves several functions is straightforward. The reader can show that

$$\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_1), \dot{\mathbf{x}}^*(t_1^-), t_1) = \frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_1), \dot{\mathbf{x}}^*(t_1^+), t_1), \tag{4.4-8a}$$

and

$$\begin{aligned} &g(\mathbf{x}^*(t_1), \dot{\mathbf{x}}^*(t_1^-), t_1) - \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_1), \dot{\mathbf{x}}^*(t_1^-), t_1)\right]^T \dot{\mathbf{x}}^*(t_1^-) \\ &= g(\mathbf{x}^*(t_1), \dot{\mathbf{x}}^*(t_1^+), t_1) - \left[\frac{\partial g}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_1), \dot{\mathbf{x}}^*(t_1^+), t_1)\right]^T \dot{\mathbf{x}}^*(t_1^+) \end{aligned} \tag{4.4-8b}$$

are the appropriate equations when $\mathbf{x}$ represents n independent functions and $\mathbf{x}(t_1)$ and t_1 are not constrained by any relationship.

† For a geometric interpretation of this relationship, refer to *Problem 4*, Fig. 4-14.

To illustrate the role of the corner conditions, let us consider the following examples.

Example 4.4-1. Find a piecewise-smooth curve that begins at the point $x(0) = 0$, ends at the point $x(2) = 1$, and minimizes the functional

$$J(x) = \int_0^2 \dot{x}^2(t)[1 - \dot{x}(t)]^2 \, dt. \tag{4.4-9}$$

The integrand g depends only on $\dot{x}(t)$; therefore, the solution of the Euler equation is (see Appendix 3, Case 1)

$$\dot{x}^*(t) = c_1 t + c_2. \tag{4.4-10}$$

The Weierstrass-Erdmann corner conditions are

$$\begin{aligned} &2\dot{x}^*(t_1^-)[1 - 2\dot{x}^*(t_1^-)][1 - \dot{x}^*(t_1^-)] \\ &= 2\dot{x}^*(t_1^+)[1 - 2\dot{x}^*(t_1^+)][1 - \dot{x}^*(t_1^+)] \end{aligned} \tag{4.4-11a}$$

and

$$\begin{aligned} &\dot{x}^{*2}(t_1^-)[1 - \dot{x}^*(t_1^-)][3\dot{x}^*(t_1^-) - 1] \\ &= \dot{x}^{*2}(t_1^+)[1 - \dot{x}^*(t_1^+)][3\dot{x}^*(t_1^+) - 1]. \end{aligned} \tag{4.4-11b}$$

Equation (4.4-11a) is satisfied by $\dot{x}^*(t_1^-) = 0, \frac{1}{2}, 1$ and $\dot{x}^*(t_1^+) = 0, \frac{1}{2}, 1$ in any combinations. Equation (4.4-11b) is satisfied by $\dot{x}^*(t_1^-) = 0, 1, \frac{1}{3}$ and $\dot{x}^*(t_1^+) = 0, 1, \frac{1}{3}$ in any combinations. Together these requirements give

$$\dot{x}^*(t_1^-) = 0 \quad \text{and} \quad \dot{x}^*(t_1^+) = 1,$$

or

$$\dot{x}^*(t_1^-) = 1 \quad \text{and} \quad \dot{x}^*(t_1^+) = 0$$

as the only nontrivial possibilities.

The curves labeled *a*, *b*, *c* in Fig. 4-17 are all extremals for this example. By inspection of the functional we see that each of these curves makes $J = 0$. Notice that if the admissible curves had been required to have continuous derivatives, the extremal would have been the straight line joining the points $x(0) = 0$ and $x(2) = 1$ (curve *d* in Fig. 4-17). The reader can verify that this curve makes $J = 0.125$.

Example 4.4-2. Find an extremal for the functional

$$J(x) = \int_0^{\pi/2} [\dot{x}^2(t) - x^2(t)] \, dt \tag{4.4-12}$$

with $x(0) = 0$ and $x(\pi/2) = 1$. Assume that $\dot{x}$ may have corners.

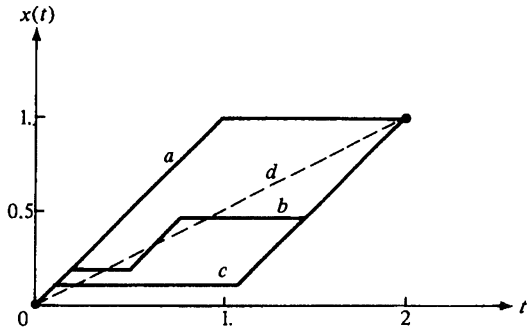


Figure 4-17 Extremal curves for Example 4.4-1

This problem was previously solved (see Example 4.2-1) under the assumption that x was required to be a smooth curve. The Euler equation

$$\ddot{x}^*(t) + x^*(t) = 0 \tag{4.4-13}$$

has a solution of the form

$$x^*(t) = c_3 \cos t + c_4 \sin t. \tag{4.4-14}$$

The Weierstrass-Erdmann corner conditions are

$$\not{2}\dot{x}^*(t_1^-) = \not{2}\dot{x}^*(t_1^+), \tag{4.4-15a}$$

and

$$\begin{aligned} &\dot{x}^{*2}(t_1^-) - x^{*2}(t_1) - [2\dot{x}^*(t_1^-)]\dot{x}^*(t_1^-) \\ &= \dot{x}^{*2}(t_1^+) - x^{*2}(t_1) - [2\dot{x}^*(t_1^+)]\dot{x}^*(t_1^+). \end{aligned} \tag{4.4-15b}$$

From Eq. (4.4-15) we see that there can be no corners, because $\dot{x}^*(t_1^-)$ must equal $\dot{x}^*(t_1^+)$. So the extremal is, as in Example 4.2-1,

$$x^*(t) = \sin t. \tag{4.4-16}$$

Let us now consider an example in which the coordinates of the corner are constrained.

Example 4.4-3. Find the shortest piecewise-smooth curve joining the points $x(0) = 1.5$ and $x(1.5) = 0$ which intersects the line $x(t) = -t + 2$ at one point.

The functional to be minimized is (see Example 4.2-2)

$$J(x) = \int_0^{1.5} [1 + \dot{x}^2(t)]^{1/2}\, dt. \tag{4.4-17}$$

The solutions of the Euler equation are of the form

$$x^*(t) = c_1 t + c_2. \tag{4.4-18}$$

In this case the corner condition of Eq. (4.4-7) becomes

$$\begin{aligned} &\frac{\dot{x}^*(t_1^-)}{[1 + \dot{x}^{*2}(t_1^-)]^{1/2}}[-1 - \dot{x}^*(t_1^-)] + [1 + \dot{x}^{*2}(t_1^-)]^{1/2} \\ &= \frac{\dot{x}^*(t_1^+)}{[1 + \dot{x}^{*2}(t_1^+)]^{1/2}}[-1 - \dot{x}^*(t_1^+)] + [1 + \dot{x}^{*2}(t_1^+)]^{1/2}. \end{aligned} \tag{4.4-19}$$

Putting both sides over common denominators and reducing, we obtain

$$\frac{1 - \dot{x}^*(t_1^-)}{[1 + \dot{x}^{*2}(t_1^-)]^{1/2}} = \frac{1 - \dot{x}^*(t_1^+)}{[1 + \dot{x}^{*2}(t_1^+)]^{1/2}}. \tag{4.4-20}$$

The extremal subarcs have the form given by Eq. (4.4-18), but the constants of integration will generally be different on the two sides of the corner, so let

$$x^*(t) = c_1 t + c_2 \qquad \text{for } t \in [0, t_1] \tag{4.4-21a}$$

$$x^*(t) = c_3 t + c_4 \qquad \text{for } t \in [t_1, 1.5]. \tag{4.4-21b}$$

Substituting the derivatives of Eqs. (4.4-21) into (4.4-20) yields

$$\frac{1 - c_1}{[1 + c_1^2]^{1/2}} = \frac{1 - c_3}{[1 + c_3^2]^{1/2}} \tag{4.4-22}$$

The extremals must also satisfy the boundary conditions $x(0) = 1.5$ and $x(1.5) = 0$, so

$$c_1 \cdot 0 + c_2 = 1.5 \Longrightarrow c_2 = 1.5 \tag{4.4-23}$$

$$1.5c_3 + c_4 = 0. \tag{4.4-24}$$

At a corner, it must also be true that $x(t_1) = -t_1 + 2$; therefore, we have the additional equations

$$c_1 t_1 + c_2 = -t_1 + 2 \tag{4.4-25}$$

$$c_3 t_1 + c_4 = -t_1 + 2. \tag{4.4-26}$$

Equations (4.4-22) through (4.4-26) are a set of five nonlinear algebraic equations in the five unknowns c_1, c_2, c_3, c_4, and t_1. These equations can be solved by using (4.4-23) through (4.4-26) to express c_1 and c_3 solely in terms of t_1, substituting these expressions in Eq. (4.4-22), and solving for t_1. Doing this gives

$$\begin{aligned} x^*(t) &= -0.5t + 1.5, \qquad t \in [0, 1.0] \\ x^*(t) &= -2t + 3, \qquad\quad t \in [1.0, 1.5] \end{aligned} \tag{4.4-27}$$

and $t_1 = 1.0$. This solution is shown in Fig. 4-18. The reader can show that we have found the shortest path to be the one whose angle of incidence θ_1 equals its angle of reflection θ_2. For further generalizations see reference [E-1], Chapter 2.

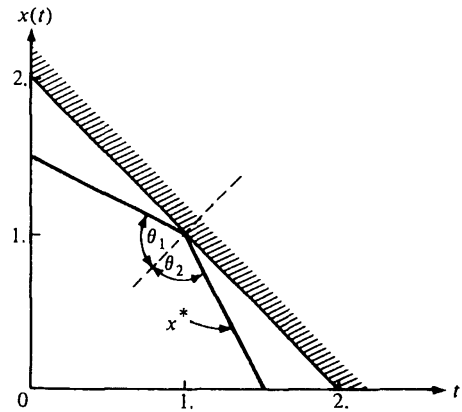


Figure 4-18 An extremal with a reflection

4.5 CONSTRAINED EXTREMA

So far, we have discussed functionals involving $\mathbf{x}$ and $\dot{\mathbf{x}}$, and we have derived necessary conditions for extremals *assuming that the components of* $\mathbf{x}$ *are independent*. In control problems the situation is more complicated, because the state trajectory is determined by the control $\mathbf{u}$; thus, we wish to consider functionals of $n + m$ functions, $\mathbf{x}$ and $\mathbf{u}$, but only m of the functions are independent—the controls. Let us now extend the necessary conditions we have derived to include problems with constraints.

To begin, we shall review the analogous problem from the calculus, and introduce some new variables—the Lagrange multipliers—that will be required for our subsequent discussion.

Constrained Minimization of Functions

Example 4.5-1. Find the point on the line $y_1 + y_2 = 5$ that is nearest the origin.

To solve this problem we need only apply elementary plane geometry to Fig. 4-19 to obtain the result that the minimum distance is $5/\sqrt{2}$, and the extreme point is $y_1^* = 2.5$, $y_2^* = 2.5$.

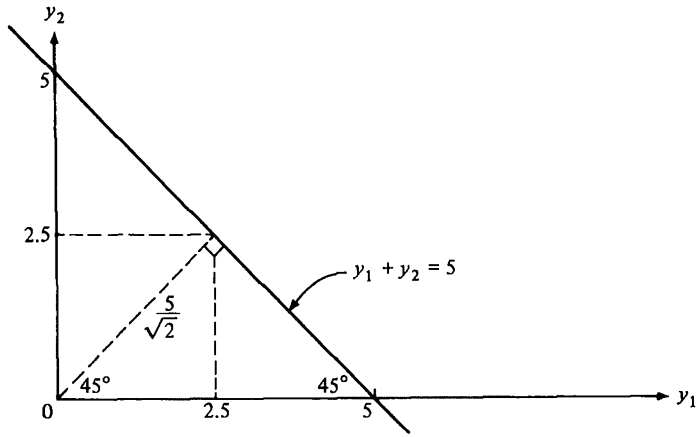


Figure 4-19 Geometrical interpretation of Example 4.5-1

Most problems cannot be solved by inspection, so let us consider alternative methods of solving this simple example.

The Elimination Method. If $\mathbf{y}^*$ is an extreme point of a function, it is necessary that the differential of the function, evaluated at $\mathbf{y}^*$, be zero.† In our example, the function

$$f(y_1, y_2) = y_1^2 + y_2^2 \quad \text{(the square of the distance)} \tag{4.5-1}$$

is to be minimized subject to the constraint

$$y_1 + y_2 = 5. \tag{4.5-2}$$

The differential is

$$df(y_1, y_2) = \left[\frac{\partial f}{\partial y_1}(y_1, y_2)\right] \Delta y_1 + \left[\frac{\partial f}{\partial y_2}(y_1, y_2)\right] \Delta y_2, \tag{4.5-3}$$

and if (y_1^*, y_2^*) is an extreme point,

$$df(y_1^*, y_2^*) = \left[\frac{\partial f}{\partial y_1}(y_1^*, y_2^*)\right] \Delta y_1 + \left[\frac{\partial f}{\partial y_2}(y_1^*, y_2^*)\right] \Delta y_2 = 0. \tag{4.5-4}$$

If y_1 and y_2 were independent, then Δy_1 and Δy_2 could be selected arbitrarily and Eq. (4.5-4) would imply that the partial derivatives must both be zero. In this example, however, y_1 and y_2 are constrained to lie on the specified line, so Δy_1 and Δy_2 are not independent. Solving Eq. (4.5-2) for y_1 and substituting into (4.5-1), we obtain

† Only interior points of bounded regions are considered.

$$\begin{aligned} f(y_2) &= [5 - y_2]^2 + y_2^2 \\ &= 25 - 10y_2 + 2y_2^2 \end{aligned} \tag{4.5-5}$$

The differential of f at the point y_2^* is then

$$df(y_2^*) = [-10 + 4y_2^*]\,\Delta y_2 = 0, \tag{4.5-6}$$

so $y_2^* = 2.5$. From (4.5-2) we then find that $y_1^* = 2.5$. The minimum value of the function is $\frac{25}{2}$, and the minimum distance is $5/\sqrt{2}$.†

The Lagrange Multiplier Method. Consider the *augmented* function

$$f_a(y_1, y_2, p) \triangleq y_1^2 + y_2^2 + p[y_1 + y_2 - 5], \tag{4.5-7}$$

with p a variable (the Lagrange multiplier) whose value is yet to be determined. For values of y_1 and y_2 that satisfy the constraining relation (4.5-2) (these are the only values of interest), the augmented function f_a equals f regardless of the value of p—we have simply added zero to f to obtain f_a. By satisfying the constraint and minimizing f_a, the constrained extreme point of f can be found. To find an extreme point of f_a, we use the necessary condition

$$\begin{aligned} df_a(y_1^*, y_2^*, p) = 0 = [2y_1^* + p]\,\Delta y_1 + [2y_2^* + p]\,\Delta y_2 \\ + [y_1^* + y_2^* - 5]\,\Delta p. \end{aligned} \tag{4.5-8}$$

Since only points that satisfy the constraining relation are acceptable,

$$y_1^* + y_2^* - 5 = 0, \tag{4.5-9}$$

but this is the coefficient of Δp. The remaining two terms must add to zero, but Δy_1 and Δy_2 are not independent—if Δy_1 is selected Δy_2 is determined, and vice versa; however, p comes to the rescue. Since the constraint must be satisfied, p can be any value, so we make a convenient choice—we select p so that the coefficient of Δy_2 (or Δy_1) is zero, and we denote this value of p by p^*. Then we have

$$2y_2^* + p^* = 0. \tag{4.5-10}$$

Δy_1 can assume arbitrary values; for each value of Δy_1 there is an associated dependent value of Δy_2, but this does not matter, because p was selected to make the coefficient of Δy_2 equal to zero. Since df_a must be zero and Δy_1 is arbitrary, the coefficient of Δy_1 must be zero; therefore,

$$2y_1^* + p^* = 0. \tag{4.5-11}$$

† Alternatively, we could reach the same final result by substituting $y_1 = 5 - y_2$ and $\Delta y_1 = \Delta y_2$ into Eq. (4.5-4), setting the coefficient of Δy_2 to zero, and solving for y_2^*.

Solving (4.5-9), (4.5-10), and (4.5-11) simultaneously gives

$$y_1^* = 2.5, \qquad y_2^* = 2.5, \qquad p^* = -5. \tag{4.5-12}$$

The reasoning that led to Eqs. (4.5-9), (4.5-10), and (4.5-11) is very important; we shall use it again shortly. Notice, however, that the same equations are obtained by forming $f_a(y_1, y_2, p)$ and then treating the three variables *as if* they were independent.

Let us now consider the "elimination method" and the method of Lagrange multipliers as they are applied in a general problem.

The problem is to find the extreme values for a function of $(n + m)$ variables, $y_1, \ldots, y_{n+m}$. The function that is to be extremized is given by $f(y_1, y_2, \ldots, y_{n+m})$. There are n constraints among the variables of the form

$$\begin{aligned} &a_1(y_1, \ldots, y_{n+m}) = 0 \\ &\cdot \\ &\cdot \\ &\cdot \\ &a_n(y_1, \ldots, y_{n+m}) = 0; \end{aligned} \tag{4.5-13}$$

thus, only $(n + m) - n = m$ of the variables are independent. Using the elimination method, we solve Eq. (4.5-13) for n of the variables in terms of the remaining m variables. For example, solving for the first n variables gives

$$\begin{aligned} &y_1 = e_1(y_{n+1}, \ldots, y_{n+m}) \\ &\cdot \\ &\cdot \\ &\cdot \\ &y_n = e_n(y_{n+1}, \ldots, y_{n+m}). \end{aligned} \tag{4.5-14}$$

Substituting these relations into f, we obtain a function of m *independent* variables, $f(y_{n+1}, \ldots, y_{n+m})$. To find the minimum value of this function, we solve the equations

$$\begin{aligned} &\frac{\partial f}{\partial y_{n+1}}(y_{n+1}^*, \ldots, y_{n+m}^*) = 0 \\ &\cdot \\ &\cdot \\ &\cdot \\ &\frac{\partial f}{\partial y_{n+m}}(y_{n+1}^*, \ldots, y_{n+m}^*) = 0 \end{aligned} \tag{4.5-15}$$

for $y_{n+1}^*, \ldots, y_{n+m}^*$, and substitute these values in (4.5-14) to obtain y_1^*, $\ldots, y_n^*$. The extreme value of f can then also be obtained. This procedure is conceptually straightforward; the principal difficulty is in obtaining the

relations (4.5-14). The solution of (4.5-15) may also be difficult, but this problem is also present in the method of Lagrange multipliers.

Now let us consider the method of Lagrange multipliers. First, we form the augmented function

$$\begin{aligned} f_a(y_1, \ldots, y_{n+m}, p_1, \ldots, p_n) &\triangleq f(y_1, \ldots, y_{n+m}) \\ &+ p_1[a_1(y_1, \ldots, y_{n+m})] + \cdots + p_n[a_n(y_1, \ldots, y_{n+m})]. \end{aligned} \tag{4.5-16}$$

Then

$$\begin{aligned} df_a &= \frac{\partial f_a}{\partial y_1}\Delta y_1 + \cdots + \frac{\partial f_a}{\partial y_{n+m}}\Delta y_{n+m} + \frac{\partial f_a}{\partial p_1}\Delta p_1 + \cdots + \frac{\partial f_a}{\partial p_n}\Delta p_n \\ &= \frac{\partial f_a}{\partial y_1}\Delta y_1 + \cdots + \frac{\partial f_a}{\partial y_{n+m}}\Delta y_{n+m} + a_1\,\Delta p_1 + \cdots + a_n\,\Delta p_n. \end{aligned} \tag{4.5-17}$$

If the constraints are satisfied, the coefficients of $\Delta p_1, \ldots, \Delta p_n$ are zero. We then select the n p_i's so that the coefficients of Δy_i $(i = 1, \ldots, n)$ are zero. The remaining m Δy_i's are independent, and for df_a to equal zero their coefficients must vanish. The result is that the extreme point $y_1^*, \ldots, y_{n+m}^*$ is found by solving the equations

$$\left.\begin{aligned} &a_i(y_1^*, \ldots, y_{n+m}^*) = 0, && i = 1, 2, \ldots, n \\ &\frac{\partial f_a}{\partial y_j}(y_1^*, \ldots, y_{n+m}^*, p_1^*, \ldots, p_n^*) = 0, && j = 1, 2, \ldots, n+m \end{aligned}\right\} \begin{array}{l} 2n+m \\ \text{equations} \end{array} \tag{4.5-18}$$

We shall now conclude our consideration of the calculus problem with another illustrative example.

Example 4.5-2 [H-1]. Find the point in three-dimensional Euclidean space that is nearest the origin and lies on the intersection of the surfaces

$$\begin{aligned} y_3 &= y_1 y_2 + 5 \\ y_1 + y_2 + y_3 &= 1. \end{aligned} \tag{4.5-19}$$

The function to be minimized is

$$f(y_1, y_2, y_3) = y_1^2 + y_2^2 + y_3^2. \tag{4.5-20}$$

The elimination method is left as an exercise for the reader. To use the method involving Lagrange multipliers, first form the augmented function

$$\begin{aligned} f_a(y_1, y_2, y_3, p_1, p_2) &= y_1^2 + y_2^2 + y_3^2 + p_1[y_1 y_2 + 5 - y_3] \\ &+ p_2[y_1 + y_2 + y_3 - 1]. \end{aligned} \tag{4.5-21}$$

Using the same reasoning as before, we find that the equations corresponding to (4.5-18) are

$$\begin{aligned} y_1^* + y_2^* + y_3^* - 1 &= 0 \\ y_1^* y_2^* + 5 - y_3^* &= 0 \\ 2y_1^* + p_1^* y_2^* + p_2^* &= 0 \\ 2y_2^* + p_1^* y_1^* + p_2^* &= 0 \\ 2y_3^* - p_1^* + p_2^* &= 0. \end{aligned} \tag{4.5-22}$$

Solving these five equations gives

$$(y_1^*, y_2^*, y_3^*) = \begin{cases} (2, -2, 1) \\ \text{or} \\ (-2, 2, 1) \end{cases} \tag{4.5-23}$$

and $f_{a_{\min}} = 9$, so the distance is 3.

Constrained Minimization of Functionals

We are now ready to consider the presence of constraints in variational problems. To simplify the variational equations, it will be assumed that the admissible curves are smooth.

Point Constraints. Let us determine a set of necessary conditions for a function $\mathbf{w}^*$ to be an extremal for a functional of the form

$$J(\mathbf{w}) = \int_{t_0}^{t'} g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\, dt; \tag{4.5-24}$$

$\mathbf{w}$ is an $(n + m) \times 1$ vector of functions $(n, m \geq 1)$ that is required to satisfy n relationships of the form

$$f_i(\mathbf{w}(t), t) = 0, \qquad i = 1, 2, \ldots, n, \tag{4.5-25}$$

which are called *point constraints*. Constraints of this type would be present if, for example, the admissible trajectories were required to lie on a specified surface in the $n + m + 1$-dimensional $\mathbf{w}(t) - t$ space. The presence of these n constraining relations means that only m of the $n + m$ components of $\mathbf{w}$ are independent.

We have previously found that the Euler equations must be satisfied regardless of the boundary conditions, so we will ignore, temporarily, terms that enter only into the determination of boundary conditions.

One way to attack this problem might be to solve Eqs. (4.5-25) for n

of the components of $\mathbf{w}(t)$ in terms of the remaining m components—which can then be regarded as m independent functions—and use these equations to eliminate the n dependent components of $\mathbf{w}(t)$ and $\dot{\mathbf{w}}(t)$ from J. If this can be done, then the equations of Sections 4.2 and 4.3 apply. Unfortunately, the constraining equations (4.5-25) are generally nonlinear algebraic equations, which may be quite difficult to solve.

As an alternative approach we can use Lagrange multipliers. The first step is to form the *augmented functional* by adjoining the constraining relations to J, which yields

$$\begin{aligned} J_a(\mathbf{w}, \mathbf{p}) &= \int_{t_0}^{t_f} \{g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + p_1(t)[f_1(\mathbf{w}(t), t)] \\ &\quad + p_2(t)[f_2(\mathbf{w}(t), t)] + \cdots + p_n(t)[f_n(\mathbf{w}(t), t)]\}\, dt \qquad (4.5\text{-}26) \\ &= \int_{t_0}^{t_f} \{g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + \mathbf{p}^T(t)[\mathbf{f}(\mathbf{w}(t), t)]\}\, dt. \end{aligned}$$

Since the constraints must be satisfied for all $t \in [t_0, t_f]$, the Lagrange multipliers $p_1, \ldots, p_n$ are assumed to be functions of time. This allows us the flexibility of multiplying the constraining relations by a *different* real number for each value of t; the reason for desiring this flexibility will become clear as we proceed.

Notice that if the constraints are satisfied, $J_a = J$ for any function $\mathbf{p}$. The variation of the functional J_a,

$$\begin{aligned} \delta J_a(\mathbf{w}, \delta\mathbf{w}, \mathbf{p}, \delta\mathbf{p}) = \int_{t_0}^{t_f} \Bigg\{ &\left[\frac{\partial g^T}{\partial \mathbf{w}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + \mathbf{p}^T(t)\left[\frac{\partial \mathbf{f}}{\partial \mathbf{w}}(\mathbf{w}(t), t)\right]\right] \delta\mathbf{w}(t) \\ &+ \left[\frac{\partial g^T}{\partial \dot{\mathbf{w}}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\right] \delta\dot{\mathbf{w}}(t) + [\mathbf{f}^T(\mathbf{w}(t), t)]\, \delta\mathbf{p}(t) \Bigg\}\, dt, \end{aligned} \qquad (4.5\text{-}27)$$

is found in the usual manner by introducing variations in the functions $\mathbf{w}$, $\dot{\mathbf{w}}$, and $\mathbf{p}$. $\partial\mathbf{f}/\partial\mathbf{w}$ denotes the $n \times (n + m)$ matrix

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial w_1} & \cdots & \dfrac{\partial f_1}{\partial w_{n+m}} \\ \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial w_1} & \cdots & \dfrac{\partial f_n}{\partial w_{n+m}} \end{bmatrix}.$$

Integrating by parts the term containing $\delta\dot{\mathbf{w}}$ and retaining only the terms inside the integral, we obtain

$$\delta J_a(\mathbf{w}, \delta\mathbf{w}, \mathbf{p}, \delta\mathbf{p}) = \int_{t_0}^{t_f} \left\{ \left[\frac{\partial g^T}{\partial \mathbf{w}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + \mathbf{p}^T(t) \left[\frac{\partial \mathbf{f}}{\partial \mathbf{w}}(\mathbf{w}(t), t) \right] \right. \right.$$
$$\left. - \frac{d}{dt} \left[\frac{\partial g^T}{\partial \dot{\mathbf{w}}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) \right] \right] \delta\mathbf{w}(t) \qquad (4.5\text{-}28)$$
$$\left. + [\mathbf{f}^T(\mathbf{w}(t), t)]\, \delta\mathbf{p}(t) \right\} dt.$$

On an extremal, the variation must be zero; that is, $\delta J_a(\mathbf{w}^*, \mathbf{p}) = 0$. In addition, the point constraints must also be satisfied by an extremal; therefore,

$$\mathbf{f}(\mathbf{w}^*(t), t) = \mathbf{0}, \qquad t \in [t_0, t_f], \qquad (4.5\text{-}29)$$

and the coefficient of $\delta\mathbf{p}(t)$ in Eq. (4.5-28) is zero. Since the constraints are satisfied, we can select the n Lagrange multipliers arbitrarily—let us choose the p's so that the coefficients of n of the components of $\delta\mathbf{w}(t)$ are zero throughout the interval $[t_0, t_f]$. The remaining $(n + m) - n = m$ components of $\delta\mathbf{w}$ are then independent; hence, the coefficients of these components of $\delta\mathbf{w}(t)$ must be zero. The final result is that, in addition to Eq. (4.5-29), the equations

$$\frac{\partial g}{\partial \mathbf{w}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t) + \left[\frac{\partial \mathbf{f}}{\partial \mathbf{w}}(\mathbf{w}^*(t), t) \right]^T \mathbf{p}^*(t)$$
$$- \frac{d}{dt} \left[\frac{\partial g}{\partial \dot{\mathbf{w}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t) \right] = \mathbf{0} \qquad (4.5\text{-}30)$$

must be satisfied.

If we define the *augmented integrand function* as

$$g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t), t) \triangleq g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + \mathbf{p}^T(t)[\mathbf{f}(\mathbf{w}(t), t)], \qquad (4.5\text{-}31)$$

then Eq. (4.5-30) can be written

$$\boxed{\begin{aligned} &\frac{\partial g_a}{\partial \mathbf{w}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), t) \\ &- \frac{d}{dt} \left[\frac{\partial g_a}{\partial \dot{\mathbf{w}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), t) \right] = \mathbf{0} \end{aligned}} \qquad (4.5\text{-}30a)$$

Equations (4.5-30a) are a set of $n + m$ second-order differential equations, and the constraining relations (4.5-29) are a set of n algebraic equations. Together, these $2n + m$ equations constitute a set of necessary conditions for $\mathbf{w}^*$ to be an extremal.

The reader may have already noticed that Eqs. (4.5-29) and (4.5-30a) are

the same as if the results from *Problem 1a* had been applied to the functional

$$J_a(\mathbf{w}, \mathbf{p}) = \int_{t_0}^{t_f} g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t), t)\, dt \tag{4.5-32}$$

with the assumption that the functions $\mathbf{w}$ and $\mathbf{p}$ are independent. It should be emphasized that, although the results are the same, the reasoning used is quite different.

Example 4.5-3. Find necessary conditions that must be satisfied by the curve of smallest length which lies on the sphere $w_1^2(t) + w_2^2(t) + t^2 = R^2$, for $t \in [t_0, t_f]$, and joins the specified points $\mathbf{w}_0, t_0$, and $\mathbf{w}_f, t_f$.

The functional to be minimized is

$$J(\mathbf{w}) = \int_{t_0}^{t_f} [1 + \dot{w}_1^2(t) + \dot{w}_2^2(t)]^{1/2}\, dt, \tag{4.5-33}$$

so the augmented integrand function is

$$\begin{aligned} &g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), p(t), t) \\ &= [1 + \dot{w}_1^2(t) + \dot{w}_2^2(t)]^{1/2} + p(t)[w_1^2(t) + w_2^2(t) + t^2 - R^2]. \end{aligned} \tag{4.5-34}$$

Performing the operations indicated by Eq. (4.5-30a) gives

$$2w_1^*(t)p^*(t) - \frac{d}{dt}\left\{\frac{\dot{w}_1^*(t)}{[1 + \dot{w}_1^{*2}(t) + \dot{w}_2^{*2}(t)]^{1/2}}\right\} = 0 \tag{4.5-35a}$$

$$2w_2^*(t)p^*(t) - \frac{d}{dt}\left\{\frac{\dot{w}_2^*(t)}{[1 + \dot{w}_1^{*2}(t) + \dot{w}_2^{*2}(t)]^{1/2}}\right\} = 0. \tag{4.5-35b}$$

In addition, of course, it is necessary that the constraining relation

$$w_1^{*2}(t) + w_2^{*2}(t) + t^2 = R^2 \tag{4.5-35c}$$

be satisfied.

Differential Equation Constraints. Let us now find necessary conditions for a function $\mathbf{w}^*$ to be an extremal for a functional

$$J(\mathbf{w}) = \int_{t_0}^{t_f} g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\, dt. \tag{4.5-36}$$

$\mathbf{w}$ is an $(n + m) \times 1$ vector of functions ($n, m \geq 1$) which must satisfy the n differential equations

$$f_i(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) = 0, \qquad i = 1, 2, \ldots, n. \tag{4.5-37}$$

Because of the n differential equation constraints, only m of the $n + m$

components of $\mathbf{w}$ are independent. Constraints of this type may represent the state equation constraints in optimal control problems where $\mathbf{w}$ corresponds to the $n + m$ vector $[\mathbf{x} \vdots \mathbf{u}]^T$.

As with point constraints, it is generally not feasible to eliminate n dependent functions and their derivatives from the functional J, so we shall again use the method of Lagrange multipliers. The derivation proceeds along the same lines as for problems with point constraints; that is, we first form the augmented functional

$$\begin{aligned} J_a(\mathbf{w}, \mathbf{p}) &= \int_{t_0}^{t_f} \{g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + p_1(t)[f_1(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)] \\ &\quad + p_2(t)[f_2(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)] + \cdots \\ &\quad + p_n(t)[f_n(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)]\}\, dt \\ &= \int_{t_0}^{t_f} \{g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + \mathbf{p}^T(t)[\mathbf{f}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)]\}\, dt. \end{aligned} \tag{4.5-38}$$

Again notice that if the constraints are satisfied, $J_a = J$ for any $\mathbf{p}(t)$. The variation of the functional J_a,

$$\begin{aligned} \delta J_a(\mathbf{w}, \delta\mathbf{w}, \mathbf{p}, \delta\mathbf{p}) = \int_{t_0}^{t_f} \Bigg\{ &\left[\frac{\partial g^T}{\partial \mathbf{w}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) \right. \\ &\left. + \mathbf{p}^T(t)\left[\frac{\partial \mathbf{f}}{\partial \mathbf{w}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\right]\right] \delta\mathbf{w}(t) \\ &+ \left[\frac{\partial g^T}{\partial \dot{\mathbf{w}}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) \right. \\ &\left. + \mathbf{p}^T(t)\left[\frac{\partial \mathbf{f}}{\partial \dot{\mathbf{w}}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\right]\right] \delta\dot{\mathbf{w}}(t) \\ &+ [\mathbf{f}^T(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)]\, \delta\mathbf{p}(t) \Bigg\}\, dt, \end{aligned} \tag{4.5-39}$$

is found in the usual manner by introducing variations in the functions $\mathbf{w}$, $\dot{\mathbf{w}}$, and $\mathbf{p}$. The notation $\partial\mathbf{f}/\partial\dot{\mathbf{w}}$ means

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial \dot{w}_1} & \cdots & \dfrac{\partial f_1}{\partial \dot{w}_{n+m}} \\ \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial \dot{w}_1} & \cdots & \dfrac{\partial f_n}{\partial \dot{w}_{n+m}} \end{bmatrix}.$$

Integrating by parts the terms containing $\delta\dot{\mathbf{w}}$ and retaining only the terms inside the integral, we obtain

$$\begin{aligned}\delta J_a(\mathbf{w}, \delta\mathbf{w}, \mathbf{p}, \delta\mathbf{p}) = \int_{t_0}^{t_f} \Bigg\{ &\left[\frac{\partial g^T}{\partial \mathbf{w}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\right. \\ &\left. + \mathbf{p}^T(t)\left[\frac{\partial \mathbf{f}}{\partial \mathbf{w}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\right]\right. \\ &- \frac{d}{dt}\left[\frac{\partial g^T}{\partial \dot{\mathbf{w}}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\right. \\ &\left.\left. + \mathbf{p}^T(t)\left[\frac{\partial \mathbf{f}}{\partial \dot{\mathbf{w}}}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\right]\right]\right] \delta\mathbf{w}(t) \\ &+ [\mathbf{f}^T(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)]\,\delta\mathbf{p}(t)\Bigg\}\, dt.\end{aligned} \tag{4.5-40}$$

On an extremal, the variation must be zero, that is, $\delta J_a(\mathbf{w}^*, \mathbf{p}) = 0$, and the differential equation constraints must also be satisfied; therefore,

$$\mathbf{f}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t) = \mathbf{0}, \tag{4.5-41}$$

and the coefficient of $\delta\mathbf{p}(t)$ in Eq. (4.5-40) is zero. Since the constraints are satisfied, we can choose the n Lagrange multipliers arbitrarily—let us select the p's so that the coefficients of n of the components of $\delta\mathbf{w}(t)$ are zero throughout the interval $[t_0, t_f]$. The remaining $(n + m) - n = m$ components of $\delta\mathbf{w}$ are then independent; hence, the coefficients of these components of $\delta\mathbf{w}(t)$ must be zero. The final result is that, in addition to Eq. (4.5-41), the equations

$$\begin{aligned}&\frac{\partial g}{\partial \mathbf{w}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t) + \left[\frac{\partial \mathbf{f}}{\partial \mathbf{w}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t)\right]^T \mathbf{p}^*(t) \\ &- \frac{d}{dt}\left\{\frac{\partial g}{\partial \dot{\mathbf{w}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t) + \left[\frac{\partial \mathbf{f}}{\partial \dot{\mathbf{w}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t)\right]^T \mathbf{p}^*(t)\right\} = \mathbf{0}\end{aligned} \tag{4.5-42}$$

must be satisfied.

If we define the augmented integrand function as

$$\begin{aligned}&g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t), t) \\ &\quad = g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) + \mathbf{p}^T(t)[\mathbf{f}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)]\end{aligned} \tag{4.5-43}$$

then Eq. (4.5-42) can be written

$$\boxed{\frac{\partial g_a}{\partial \mathbf{w}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g_a}{\partial \dot{\mathbf{w}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), t)\right] = \mathbf{0}.} \tag{4.5-42a}$$

Equations (4.5-41) and (4.5-42a) comprise a set of $(2n + m)$ second-order

differential equations. We shall see in Chapter 5 that in optimal control problems m of these equations are algebraic, and the remaining $2n$ differential equations are first order.

Equations (4.5-41) and (4.5-42a) are the same as if the results of *Problem 1a* had been applied to the functional

$$J_a(\mathbf{w}, \mathbf{p}) = \int_{t_0}^{t_f} g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t), t)\, dt \tag{4.5-44}$$

with the assumption that the functions $\mathbf{w}$ and $\mathbf{p}$ are independent. Again we emphasize that although the results are the same, the reasoning is quite different!

Example 4.5-4. Find the equations that must be satisfied by an extremal for the functional

$$J(\mathbf{w}) = \int_{t_0}^{t_f} \tfrac{1}{2}[w_1^2(t) + w_2^2(t)]\, dt, \tag{4.5-45}$$

where the functions w_1 and w_2 are related by

$$\dot{w}_1(t) = w_2(t). \tag{4.5-46}$$

There is one constraint, so the function $\mathbf{f}$ in Eq. (4.5-41) is

$$f(\mathbf{w}(t), \dot{\mathbf{w}}(t)) = w_2(t) - \dot{w}_1(t), \tag{4.5-47}$$

and one Lagrange multiplier $p(t)$ is required. The function g_a in Eq. (4.5-43) is

$$g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), p(t)) = \tfrac{1}{2}w_1^2(t) + \tfrac{1}{2}w_2^2(t) + p(t)w_2(t) - p(t)\dot{w}_1(t). \tag{4.5-48}$$

From Eq. (4.5-42a) we have

$$\begin{aligned} w_1^*(t) + \dot{p}^*(t) &= 0 \\ w_2^*(t) + p^*(t) &= 0, \end{aligned} \tag{4.5-49}$$

and satisfaction of (4.5-46) requires that

$$\dot{w}_1^*(t) = w_2^*(t). \tag{4.5-46a}$$

Equations (4.5-49) and (4.5-46a) are necessary conditions for $\mathbf{w}^*$ to be an extremal.

Example 4.5-5. Suppose that the system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) - x_1(t) \\ \dot{x}_2(t) &= -2x_1(t) - 3x_2(t) + u(t) \end{aligned} \tag{4.5-50}$$

is to be controlled to minimize the performance measure

$$J(\mathbf{x}, u) = \int_{t_0}^{t_f} \tfrac{1}{2}[x_1^2(t) + x_2^2(t) + u^2(t)]\, dt. \tag{4.5-51}$$

Find a set of necessary conditions for optimal control.

If we define $x_1 \triangleq w_1$, $x_2 \triangleq w_2$, and $u \triangleq w_3$, the problem statement and solution, using the notation of this section, are the following.

Find the equations that must be satisfied for a function $\mathbf{w}^*$ to be an extremal for the functional

$$J(\mathbf{w}) = \int_{t_0}^{t_f} \tfrac{1}{2}[w_1^2(t) + w_2^2(t) + w_3^2(t)]\, dt, \tag{4.5-52}$$

where the function $\mathbf{w}$ must satisfy the differential equation constraints

$$\begin{aligned} \dot{w}_1(t) &= w_2(t) - w_1(t) \\ \dot{w}_2(t) &= -2w_1(t) - 3w_2(t) + w_3(t). \end{aligned} \tag{4.5-53}$$

The function $\mathbf{f}$ is

$$\begin{aligned} f_1(\mathbf{w}(t), \dot{\mathbf{w}}(t)) &= w_2(t) - w_1(t) - \dot{w}_1(t) = 0 \\ f_2(\mathbf{w}(t), \dot{\mathbf{w}}(t)) &= -2w_1(t) - 3w_2(t) + w_3(t) - \dot{w}_2(t) = 0, \end{aligned} \tag{4.5-54}$$

and g_a is given by

$$\begin{aligned} g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t)) = {} & \tfrac{1}{2}w_1^2(t) + \tfrac{1}{2}w_2^2(t) + \tfrac{1}{2}w_3^2(t) \\ & + p_1(t)[w_2(t) - w_1(t) - \dot{w}_1(t)] \\ & + p_2(t)[-2w_1(t) - 3w_2(t) + w_3(t) - \dot{w}_2(t)]. \end{aligned} \tag{4.5-55}$$

From Eq. (4.5-42a), we obtain the differential equations

$$\begin{aligned} \dot{p}_1^*(t) &= -w_1^*(t) + p_1^*(t) + 2p_2^*(t) \\ \dot{p}_2^*(t) &= -w_2^*(t) - p_1^*(t) + 3p_2^*(t), \end{aligned} \tag{4.5-56}$$

and the algebraic equation (since $\dot{w}_3$ does not appear in g_a),

$$w_3^*(t) + p_2^*(t) = 0. \tag{4.5-57}$$

The two additional equations that must be satisfied by an extremal are the constraints

$$\begin{aligned} \dot{w}_1^*(t) &= w_2^*(t) - w_1^*(t) \\ \dot{w}_2^*(t) &= -2w_1^*(t) - 3w_2^*(t) + w_3^*(t). \end{aligned} \tag{4.5-58}$$

Isoperimetric Constraints. Queen Dido's land transaction was perhaps the original problem with an *isoperimetric constraint*—she attempted to find the

curve having a fixed length which enclosed the maximum area. Today, we say that any constraints of the form

$$\int_{t_0}^{t_f} e_i(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\, dt = c_i \qquad (i = 1, 2, \ldots, r) \tag{4.5-59}$$

are isoperimetric constraints. The c_i's are specified constants. In control problems such constraints often enter in the form of total fuel or energy available to perform a required task.

Suppose that it is desired to find necessary conditions for $\mathbf{w}^*$ to be an extremal for

$$J(\mathbf{w}) = \int_{t_0}^{t_f} g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\, dt \tag{4.5-60}$$

subject to the isoperimetric constraints given in Eq. (4.5-59).

These constraints can be put into the form of differential equation constraints by defining the new variables

$$z_i(t) \triangleq \int_{t_0}^{t} e_i(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)\, dt, \qquad i = 1, 2, \ldots, r.\dagger \tag{4.5-61}$$

The required boundary conditions for these additional variables are $z_i(t_0) = 0$ and $z_i(t_f) = c_i$. Differentiating Eq. (4.5-61) with respect to time gives

$$\dot{z}_i(t) = e_i(\mathbf{w}(t), \dot{\mathbf{w}}(t), t), \qquad i = 1, 2, \ldots, r, \tag{4.5-62}$$

or, in vector notation,

$$\dot{\mathbf{z}}(t) = \mathbf{e}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t). \tag{4.5-62a}$$

Equation (4.5-62a) is a set of r differential equation constraints which we treat, as before, by forming the augmented function

$$\begin{aligned} g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t), \dot{\mathbf{z}}(t), t) &\triangleq g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) \\ &+ \mathbf{p}^T(t)[\mathbf{e}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) - \dot{\mathbf{z}}(t)]. \end{aligned} \tag{4.5-63}$$

Corresponding to Eq. (4.5-42a), we now have the set of $n + m$ equations

$$\frac{\partial g_a}{\partial \mathbf{w}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), \dot{\mathbf{z}}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g_a}{\partial \dot{\mathbf{w}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), \dot{\mathbf{z}}^*(t), t)\right] = \mathbf{0}, \tag{4.5-64}$$

and the set of r equations

$$\frac{\partial g_a}{\partial \mathbf{z}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), \dot{\mathbf{z}}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g_a}{\partial \dot{\mathbf{z}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), \dot{\mathbf{z}}^*(t), t)\right] = \mathbf{0}, \tag{4.5-65}$$

† Notice that the upper limit on the integral is t, *not* t_f.

a total of $(n + m + r)$ equations involving $(n + m + r + r)$ functions $(\mathbf{w}^*, \mathbf{p}^*, \mathbf{z}^*)$. The additional r equations required are

$$\dot{\mathbf{z}}^*(t) = \mathbf{e}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t) \tag{4.5-66}$$

whose solution must satisfy the boundary conditions $z_i^*(t_f) = c_i$, $i = 1, \ldots, r$.

Notice that g_a does not contain $\mathbf{z}(t)$, so $\partial g_a/\partial \mathbf{z} \equiv \mathbf{0}$. In addition, $\partial g_a/\partial \dot{\mathbf{z}} = -\mathbf{p}^*(t)$; therefore, Eq. (4.5-65) always gives

$$\dot{\mathbf{p}}^*(t) = \mathbf{0}, \tag{4.5-67}$$

which implies that the Lagrange multipliers are constants.

To summarize, for problems with isoperimetric constraints, the necessary conditions for an extremal are given by Eqs. (4.5-64), (4.5-66), and (4.5-67). The following examples illustrate the use of these equations.

Example 4.5-6. Find necessary conditions for $\mathbf{w}^*$ to be an extremal of the functional

$$J(\mathbf{w}) = \int_{t_0}^{t_f} \tfrac{1}{2}[w_1^2(t) + w_2^2(t) + 2\dot{w}_1(t)\dot{w}_2(t)]\, dt \tag{4.5-68}$$

subject to the constraint

$$\int_{t_0}^{t_f} w_2^2(t)\, dt = c; \tag{4.5-69}$$

c is a specified constant.

Let $\dot{z}(t) \triangleq w_2^2(t)$; then

$$\begin{aligned} g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), p(t), \dot{z}(t)) = {} & \tfrac{1}{2}w_1^2(t) + \tfrac{1}{2}w_2^2(t) + \dot{w}_1(t)\dot{w}_2(t) \\ & + p(t)[w_2^2(t) - \dot{z}(t)]. \end{aligned} \tag{4.5-70}$$

From (4.5-64),

$$\begin{aligned} & w_1^*(t) - \ddot{w}_2^*(t) = 0 \\ & w_2^*(t) + 2w_2^*(t)p^*(t) - \dot{w}_1^*(t) = 0, \end{aligned} \tag{4.5-71}$$

and Eq. (4.5-65) gives

$$\dot{p}^*(t) = 0. \tag{4.5-72}$$

In addition, the solution of the differential equation

$$\dot{z}^*(t) = w_2^{*2}(t), \qquad z^*(t_0) = 0 \tag{4.5-73}$$

must satisfy the boundary condition

$$z^*(t_f) = c. \tag{4.5-74}$$

In control problems, there are always state differential equation constraints, in addition to any isoperimetric constraints. Let us now consider an example having both types of constraints.

Example 4.5-7. The system with state equations

$$\begin{aligned} \dot{x}_1(t) &= -x_1(t) + x_2(t) + u(t) \\ \dot{x}_2(t) &= -2x_1(t) - 3x_2(t) + u(t) \end{aligned} \tag{4.5-75}$$

is to be controlled to minimize the performance measure

$$J(\mathbf{x}, u) = \int_{t_0}^{t_f} \tfrac{1}{2}[x_1^2(t) + x_2^2(t)]\, dt. \tag{4.5-76}$$

The total control energy to be expended is

$$\int_{t_0}^{t_f} u^2(t)\, dt = c, \tag{4.5-77}$$

where c is a specified constant. Find a set of necessary conditions for optimal control.

If we define $x_1 \triangleq w_1$, $x_2 \triangleq w_2$, and $u \triangleq w_3$, then the problem stated in the notation of this section is as follows.

Find necessary conditions that must be satisfied by an extremal for the functional

$$J(\mathbf{w}) = \int_{t_0}^{t_f} \tfrac{1}{2}[w_1^2(t) + w_2^2(t)]\, dt. \tag{4.5-78}$$

The constraining relations are

$$\begin{aligned} \dot{w}_1(t) &= -w_1(t) + w_2(t) + w_3(t) \\ \dot{w}_2(t) &= -2w_1(t) - 3w_2(t) + w_3(t) \end{aligned} \tag{4.5-79}$$

and

$$\int_{t_0}^{t_f} w_3^2(t)\, dt = c. \tag{4.5-80}$$

First, we form the function

$$\begin{aligned} g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t), \dot{z}(t)) = {} & \tfrac{1}{2}w_1^2(t) + \tfrac{1}{2}w_2^2(t) \\ & + p_1(t)[-w_1(t) + w_2(t) + w_3(t) - \dot{w}_1(t)] \\ & + p_2(t)[-2w_1(t) - 3w_2(t) + w_3(t) - \dot{w}_2(t)] \\ & + p_3(t)[w_3^2(t) - \dot{z}(t)]. \end{aligned}$$

The required equations are

$$
\begin{aligned}
\dot{p}_1^*(t) &= p_1^*(t) + 2p_2^*(t) - w_1^*(t) \\
\dot{p}_2^*(t) &= -p_1^*(t) + 3p_2^*(t) - w_2^*(t) \\
p_1^*(t) &+ p_2^*(t) + 2w_3^*(t)p_3^*(t) = 0 \\
\dot{p}_3^*(t) &= 0 \\
\dot{w}_1^*(t) &= -w_1^*(t) + w_2^*(t) + w_3^*(t) \\
\dot{w}_2^*(t) &= -2w_1^*(t) - 3w_2^*(t) + w_3^*(t) \\
\dot{z}^*(t) &= w_3^{*2}(t), \qquad z^*(t_0) = 0.
\end{aligned}
\tag{4.5-81}
$$

The boundary condition $z^*(t_f) = c$ must also be satisfied.

To recapitulate, the important result of this section is that a necessary condition for problems with differential equation constraints, or point constraints, is

$$
\boxed{\frac{\partial g_a}{\partial \mathbf{w}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g_a}{\partial \dot{\mathbf{w}}}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), \mathbf{p}^*(t), t)\right] = \mathbf{0},} \tag{4.5-42a}
$$

where

$$
\begin{aligned}
g_a(\mathbf{w}(t), \dot{\mathbf{w}}(t), \mathbf{p}(t), t) &\triangleq g(\mathbf{w}(t), \dot{\mathbf{w}}(t), t) \\
&+ \mathbf{p}^T(t)[\mathbf{f}(\mathbf{w}(t), \dot{\mathbf{w}}(t), t)].
\end{aligned}
\tag{4.5-43}
$$

This means that to determine the necessary conditions for an extremal we simply form the function g_a and write the Euler equations *as if* there were no constraints among the functions $\mathbf{w}$. Naturally, the constraining relations

$$
\mathbf{f}(\mathbf{w}^*(t), \dot{\mathbf{w}}^*(t), t) = \mathbf{0} \qquad \dagger(4.5\text{-}41)
$$

must also be satisfied.

4.6 SUMMARY

In this chapter, some basic ideas of the calculus of variations have been introduced. The analogy between familiar results of the calculus and corresponding results in the calculus of variations has been established and

† If $\dot{\mathbf{w}}(t)$ does not appear explicitly in $\mathbf{f}$, then we have point constraints.

exploited. First, some basic definitions were stated, and used to prove the fundamental theorem of the calculus of variations. The fundamental theorem was then applied to determine necessary conditions to be satisfied by an extremal. Initially, the problems considered were assumed to have trajectories with fixed end points; subsequently, problems with free end points were considered. We found that regardless of the boundary conditions, the fundamental theorem yields a set of differential equations (the Euler equations) that are the same for a specified functional. Furthermore, we observed that the Euler equations are generally *nonlinear* differential equations with *split boundary values*; these two characteristics combine to make the solution of optimal control problems a challenging task.

In control problems the system trajectory is determined by the applied control—we say that the optimization problem is *constrained* by the dynamics of the process. In the concluding section of this chapter we considered constrained problems and introduced the method of Lagrange multipliers.

With this background material, we are at last ready to tackle "the optimal control problem."

REFERENCES

B-4 Bartle, R. G., *The Elements of Real Analysis*. New York: John Wiley & Sons, Inc., 1964.

E-1 Elsgolc, L. E., *Calculus of Variations*. Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1962.

G-1 Gelfand, I. M., and S. V. Fomin, *Calculus of Variations*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963.

H-1 Hildebrand, F. B., *Methods of Applied Mathematics*, 1st ed. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1952.

M-2 Menger, K., "What Is Calculus of Variations and What Are Its Applications?", *The World of Mathematics*, Vol. 2. New York: Simon and Schuster, Inc., 1956.

O-2 Olmstead, J. M. H., *Real Variables*. New York: Appleton-Century-Crofts, Inc., 1959.

PROBLEMS

4-1. f is a differentiable function of n variables defined on the domain $\mathcal{D}$. If $\mathbf{q}^*$ is an interior point of $\mathcal{D}$ and $f(\mathbf{q}^*)$ is a relative extremum, prove that the differential of f must be zero at the point $\mathbf{q}^*$.

4-2. Prove the fundamental lemma; that is, show that if $h(t)$ is continuous for $t \in [t_0, t_f]$, and if

$$\int_{t_0}^{t_f} h(t)\,\delta x(t)\,dt = 0$$

for every function $\delta x(t)$ that is continuous in the interval $[t_0, t_f]$ with $\delta x(t_0) = \delta x(t_f) = 0$, then $h(t)$ must be identically zero in the interval $[t_0, t_f]$.

4-3. Using the definition, find the differentials of the following functions:
(a) $f(t) = 4t^3 + 5/t,\ t > 0$.
(b) $f(q_1, q_2) = 5q_1^2, + 6q_1q_2 + 2q_2^2$.
(c) $f(\mathbf{q}) = q_1^2 + q_2^2 + 5q_1q_2q_3 + 2q_1q_2 + 3q_3$.
Compare your answers with the results obtained by using formal procedures for determining the differential.

4-4. Determine the variations of the functionals:
(a) $J(x) = \int_{t_0}^{t_f} [x^3(t) - x^2(t)\dot{x}(t)]\,dt$.
(b) $J(\mathbf{x}) = \int_{t_0}^{t_f} [x_1^2(t) + x_1(t)x_2(t) + x_2^2(t) + 2\dot{x}_1(t)\dot{x}_2(t)]\,dt$.
(c) $J(x) = \int_{t_0}^{t_f} \epsilon^{x(t)}\,dt$.
Assume that the end points are specified.

4-5. Consider *Problem 1* of Section 4.2 and let η be a specified continuously differentiable function that is arbitrary in the interval $[t_0, t_f]$ except at the end points, where $\eta(t_0) = \eta(t_f) = 0$. If ϵ is an arbitrary real parameter, then $x^* + \epsilon\eta$ represents a family of curves. Evaluating the functional

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\,dt$$

on the family $x^* + \epsilon\eta$ makes J a *function* of ϵ, and if x^* is an extremal this function must have a relative extremum at the point $\epsilon = 0$.

Show that the Euler equation (4.2-10) is obtained from the necessary condition

$$\left.\frac{dJ(x^* + \epsilon\eta)}{d\epsilon}\right|_{\epsilon=0} = 0.$$

4-6. Euler derived necessary conditions to be satisfied by an extremal using finite differences. The first step in the finite-difference approach to *Problem 1* of Section 4.2 is to approximate the functional

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)\,dt$$

by the summation

$$J_d \approx \sum_{k=0}^{N-1} g(x(k), \dot{x}(k), k)\,\Delta t,$$

where $x(k) \triangleq x(t_0 + k\,\Delta t)$. The derivative is approximated by

$$\dot{x}(k) \approx \frac{x(k+1) - x(k)}{\Delta t}.$$

By making these approximations, the problem becomes one of determining the $N-1$ independent parameters $x(1), \ldots, x(N-1)$ [$x(0)$ and $x(N)$ are fixed] that minimize (or maximize) J_d. Show that by using the necessary conditions

$$\frac{\partial J_d}{\partial x(k)} = 0, \qquad k = 1, \ldots, N-1,$$

and by letting $N \to \infty$, the Euler equation (4.2-10) is obtained.

4-7. Find the extrema for the functions:
(a) $f(t) = 0.333t^3 + 1.5t^2 + 2.0t + 5$.
(b) $f(t) = te^{-2t}$, $t \geq 0$.
(c) $f(\mathbf{q}) = q_1^2 + 2q_2^2 + 9q_1 - q_2 + q_1q_2 + 22$.

4-8. Find the extremals for the following functionals:
(a) $J(x) = \int_0^1 [x^2(t) + \dot{x}^2(t)]\,dt$; $x(0) = 0$, $x(1) = 1$.
(b) $J(x) = \int_0^2 [x^2(t) + 2\dot{x}(t)x(t) + \dot{x}^2(t)]\,dt$; $x(0) = 1$, $x(2) = -3$.
(c) $J(\mathbf{x}) = \int_0^{\pi/2} [\dot{x}_1^2(t) + \dot{x}_2^2(t) + 2x_1(t)x_2(t)]\,dt$; $x_1(0) = 0$, $x_1(\pi/2) = 1$, $x_2(0) = 0$, $x_2(\pi/2) = 1$.

4-9. Find the curve x^* that minimizes the functional

$$J(x) = \int_0^1 [\tfrac{1}{2}\dot{x}^2(t) + 3x(t)\dot{x}(t) + 2x^2(t) + 4x(t)]\,dt$$

and passes through the points $x(0) = 1$, $x(1) = 4$.

4-10. Find extremals for the functionals:
(a) $J(x) = \int_0^1 [x^2(t) + \dot{x}^2(t)]\,dt$; $x(0) = 1$, $x(1)$ free.
(b) $J(x) = \int_0^1 [\tfrac{1}{2}\dot{x}^2(t) + x(t)\dot{x}(t) + \dot{x}(t) + x(t)]\,dt$; $x(0) = \tfrac{1}{2}$, $x(1)$ free.
(c) $J(\mathbf{x}) = \int_0^{\pi/2} [\dot{x}_1^2(t) + \dot{x}_2^2(t) + 2x_1(t)x_2(t)]\,dt$; $x_1(0) = 0$, $x_1(\pi/2)$ free, $x_2(0) = 0$, $x_2(\pi/2) = 1$.

4-11. Consider the functional

$$J(x) = \int_{t_0}^{t_f} g\left(x(t), \dot{x}(t), \ldots, \frac{d^r x(t)}{dt^r}, t\right) dt.$$

t_0 and t_f are specified, and $2r$ boundary conditions ($x(t_0)$, $x(t_f)$ and the first $(r-1)$ derivatives of x at t_0 and t_f) are given.

Show that the Euler equation is

$$\sum_{k=0}^{r} (-1)^k \frac{d^k}{dt^k}\left[\frac{\partial g}{\partial x^{(k)}}\left(x^*(t), \ldots, \frac{d^r x^*(t)}{dt^r}, t\right)\right] = 0,$$

where $x^{(k)}$ has been used to denote $d^k x(t)/dt^k$.

4-12. Use the Euler equation from Problem 4-11 to find extremals for the functionals:

(a) $J(x) = \int_0^1 [x(t)\dot{x}(t) + \ddot{x}^2(t)]\, dt$; $x(0) = 0,\ \dot{x}(0) = 1,$ $x(1) = 2,\ \dot{x}(1) = 4.$

(b) $J(x) = \int_0^\infty \{\ddot{x}^2(t) + x^2(t) + [\dot{x}(t) + \dot{x}(t)]^2\}\, dt$; $x(0) = 1,\quad \dot{x}(0) = 2,$ $x(\infty) = 0,\ \dot{x}(\infty) = 0.$

4-13. Determine an extremal for the functional

$$J(x) = \int_0^{t_f} \sqrt{1 + \dot{x}^2(t)}\, dt,$$

which has $x(0) = 2$ and terminates on the curve $\theta(t) = -4t + 5$.

4-14. Find the extremal curves for the functional

$$J(x) = \int_0^{t_f} \left[\frac{\sqrt{1 + \dot{x}^2(t)}}{x(t)}\right] dt.$$

$x(0) = 0$, and $x(t_f)$ must lie on the line $\theta(t) = t - 5$.

4-15. Find a curve that is an extremal for the functional

$$J(x) = \int_0^{t_f} \sqrt{1 + \dot{x}^2(t)}\, dt.$$

$x(0) = 5$, and the end points must lie on the circle $x^2(t) + (t - 5)^2 - 4 = 0$. Verify your solution geometrically.

Hint: This is an end point constraint of the form $m(x(t_f), t_f) = 0$. Draw a picture to determine the relationship between δx_f and δt_f.

4-16. Repeat Problem 4-14 with $x(0) = 0$ and $x(t_f)$ lying on the circle $(t - 9)^2 + x^2(t) = 9$.

4-17. Determine the equations that would have to be solved to find the constants of integration for the functional in Problem 4-8(c) if the boundary conditions are

(a) $x_1(0) = 0$, $x_2(0) = 0$, t_f is free, and $\mathbf{x}(t_f)$ must lie on the curve

$$\boldsymbol{\theta}(t) = \begin{bmatrix} 5t + 3 \\ \frac{1}{2}t^2 \end{bmatrix}.$$

(b) $x_1(0) = 0$, $x_2(0) = 0$, t_f is free, and $\mathbf{x}(t_f)$ must lie on the surface $x_1(t) + 3x_2(t) + 5t = 15$.

4-18. Find the shortest piecewise-smooth curve joining the points $x(-2) = 0$ and $x(1) = 0$ that intersects the curve $x(t) = t^2 + 2$ at one point.

4-19. Show that extremals for the functional

$$J(x) = \int_{t_0}^{t_f} [a\dot{x}^2(t) + bx(t)\dot{x}(t) + cx^2(t)]\, dt$$

can have no corners. a, b, and c are constants, and it is given that $a \neq 0$, $x(t_0) = x_0$, and $x(t_f) = x_f$.

4-20. Determine the extremals for the functional

$$J(x) = \int_0^4 [\dot{x}(t) - 1]^2[\dot{x}(t) + 1]^2\, dt$$

which have only one corner. The boundary conditions are $x(0) = 0$, $x(4) = 2$.

4-21. Find a point on the curve

$$y_2 = y_1^2 - 4.5$$

that minimizes the function $f(y_1, y_2) = y_1^2 + y_2^2$.

4-22. Using calculus, find the point in three-dimensional Euclidean space that satisfies the constraints

$$y_1 + y_2 + y_3 = 5$$
$$y_1^2 + y_2^2 + y_3 = 9$$

and is nearest the origin.

4-23. In Section 4.5, functions constrained by differential equations were considered. In deriving necessary conditions for an extremal, the terms that determine boundary values at $t = t_f$ were ignored. Suppose that t_f and $\mathbf{w}(t_f)$ are free; what are the terms that would appear in addition to the integral in Eq. (4.5-40)?

4-24. Determine necessary conditions (excluding boundary conditions) that must be satisfied by extremals for the functionals:

(a) $J(\mathbf{w}) = \int_{t_0}^{t_f} [w_1^2(t) + w_1(t)w_2(t) + w_2^2(t) + w_3^2(t)]\, dt$,
where the constraining equations
$\dot{w}_1(t) = w_2(t)$
$\dot{w}_2(t) = -w_1(t) + [1 - w_1^2(t)]w_2(t) + w_3(t)$
must be satisfied.

(b) $J(\mathbf{w}) = \int_{t_0}^{t_f} [\lambda + w_3^2(t)]\, dt, \qquad \lambda > 0,$
and the differential equation constraints
$\dot{w}_1(t) = w_2(t)$
$\dot{w}_2(t) = w_3(t)$
must be satisfied.

(c) $J(\mathbf{w}) = \int_{t_0}^{t_f} [\lambda + w_3^2(t)]\, dt, \qquad \lambda > 0,$
and the differential equations
$\dot{w}_1(t) = w_2(t)$
$\dot{w}_2(t) = -w_2(t)|w_2(t)| + w_3(t)$
must be satisfied.

4-25. Find the extremals for the functional

$$J(x) = \int_0^1 [\dot{x}^2(t) + t^2]\, dt$$

which satisfy the boundary conditions $x(0) = 0$, $x(1) = 0$, and the constraint

$$\int_0^1 x^2(t)\, dt = 2.$$

4-26. A particle of unit mass moves on the surface $f(w_1(t), w_2(t), w_3(t)) = 0$ from the point $(w_{1_0}, w_{2_0}, w_{3_0})$ to the point $(w_{1_f}, w_{2_f}, w_{3_f})$ in fixed time T. Show that if the particle moves so that the integral of the kinetic energy is minimized, then the motion satisfies the equations

$$\frac{\ddot{w}_1}{\dfrac{\partial f}{\partial w_1}} = \frac{\ddot{w}_2}{\dfrac{\partial f}{\partial w_2}} = \frac{\ddot{w}_3}{\dfrac{\partial f}{\partial w_3}}.$$

5

The Variational Approach to Optimal Control Problems

In this chapter we shall apply variational methods to optimal control problems. We shall first derive necessary conditions for optimal control assuming that the admissible controls are not bounded. These necessary conditions are then employed to find the optimal control law for the important linear regulator problem. Next, Pontryagin's minimum principle is introduced heuristically as a generalization of the fundamental theorem of the calculus of variations, and problems with bounded control and state variables are discussed. The three concluding sections of the chapter are devoted to time-optimal problems, minimum control-effort systems, and problems involving singular intervals.

5.1 NECESSARY CONDITIONS FOR OPTIMAL CONTROL

Let us now employ the techniques introduced in Chapter 4 to determine necessary conditions for optimal control. As stated in Chapter 1, the problem is to find an admissible control $\mathbf{u}^*$ that causes the system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{5.1-1}$$

to follow an admissible trajectory $\mathbf{x}^*$ that minimizes the performance measure

$$J(\mathbf{u}) = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt.\dagger \tag{5.1-2}$$

We shall initially assume that the admissible state and control regions are not bounded, and that the initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$ and the initial time t_0 are specified. As usual, $\mathbf{x}$ is the $n \times 1$ state vector and $\mathbf{u}$ is the $m \times 1$ vector of control inputs.

In the terminology of Chapter 4, we have a problem involving $n + m$ functions which must satisfy the n differential equation constraints (5.1-1). The m control inputs are the independent functions.

The only difference between Eq. (5.1-2) and the functionals considered in Chapter 4 is the term involving the final states and final time. However, assuming that h is a differentiable function, we can write

$$h(\mathbf{x}(t_f), t_f) = \int_{t_0}^{t_f} \frac{d}{dt}\left[h(\mathbf{x}(t), t)\right] dt + h(\mathbf{x}(t_0), t_0), \tag{5.1-3}$$

so that the performance measure can be expressed as

$$J(\mathbf{u}) = \int_{t_0}^{t_f} \left\{ g(\mathbf{x}(t), \mathbf{u}(t), t) + \frac{d}{dt}\left[h(\mathbf{x}(t), t)\right] \right\} dt + h(\mathbf{x}(t_0), t_0). \tag{5.1-4}$$

Since $\mathbf{x}(t_0)$ and t_0 are fixed, the minimization does not affect the $h(\mathbf{x}(t_0), t_0)$ term, so we need consider only the functional

$$J(\mathbf{u}) = \int_{t_0}^{t_f} \left\{ g(\mathbf{x}(t), \mathbf{u}(t), t) + \frac{d}{dt}\left[h(\mathbf{x}(t), t)\right] \right\} dt. \tag{5.1-5}$$

Using the chain rule of differentiation, we find that this becomes

$$J(\mathbf{u}) = \int_{t_0}^{t_f} \left\{ g(\mathbf{x}(t), \mathbf{u}(t), t) + \left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}(t), t)\right]^T \dot{\mathbf{x}}(t) + \frac{\partial h}{\partial t}(\mathbf{x}(t), t) \right\} dt. \tag{5.1-6}$$

To include the differential equation constraints, we form the augmented functional

$$J_a(\mathbf{u}) = \int_{t_0}^{t_f} \left\{ g(\mathbf{x}(t), \mathbf{u}(t), t) + \left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}(t), t)\right]^T \dot{\mathbf{x}}(t) + \frac{\partial h}{\partial t}(\mathbf{x}(t), t) + \mathbf{p}^T(t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) - \dot{\mathbf{x}}(t)] \right\} dt \tag{5.1-7}$$

by introducing the Lagrange multipliers $p_1(t), \ldots, p_n(t)$. Let us define

† In general, the functional J depends on $\mathbf{x}(t_0)$, t_0, $\mathbf{x}$, $\mathbf{u}$, the target set S, and t_f. However, here it is assumed that $\mathbf{x}(t_0)$ and t_0 are specified; hence, $\mathbf{x}$ is determined by $\mathbf{u}$ and we write $J(\mathbf{u})$—the dependence of J on S and t_f will not be explicitly indicated.

$$g_a(\mathbf{x}(t), \dot{\mathbf{x}}(t), \mathbf{u}(t), \mathbf{p}(t), t) \triangleq g(\mathbf{x}(t), \mathbf{u}(t), t) + \mathbf{p}^T(t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) - \dot{\mathbf{x}}(t)] + \left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}(t), t)\right]^T \dot{\mathbf{x}}(t) + \frac{\partial h}{\partial t}(\mathbf{x}(t), t)$$

so that

$$J_a(\mathbf{u}) = \int_{t_0}^{t_f} \{g_a(\mathbf{x}(t), \dot{\mathbf{x}}(t), \mathbf{u}(t), \mathbf{p}(t), t)\}\, dt. \tag{5.1-8}$$

We shall assume that the end points at $t = t_f$ can be specified or free. To determine the variation of J_a, we introduce the variations $\delta\mathbf{x}$, $\delta\dot{\mathbf{x}}$, $\delta\mathbf{u}$, $\delta\mathbf{p}$, and δt_f. From *Problem 4a* in the preceding chapter this gives [see Eq. (4.3-16)], on an extremal,

$$\begin{aligned}
\delta J_a(\mathbf{u}^*) = 0 = &\left[\frac{\partial g_a}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f)\right]^T \delta\mathbf{x}_f \\
&+ \Big[g_a(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) \\
&- \left[\frac{\partial g_a}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t_f), \dot{\mathbf{x}}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f)\right]^T \dot{\mathbf{x}}^*(t_f)\Big] \delta t_f \\
&+ \int_{t_0}^{t_f} \Big\{\Big[\left[\frac{\partial g_a}{\partial \mathbf{x}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T \\
&- \frac{d}{dt}\left[\frac{\partial g_a}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T\Big] \delta\mathbf{x}(t) \\
&+ \left[\frac{\partial g_a}{\partial \mathbf{u}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T \delta\mathbf{u}(t) \\
&+ \left[\frac{\partial g_a}{\partial \mathbf{p}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T \delta\mathbf{p}(t)\Big\}\, dt.
\end{aligned} \tag{5.1-9}$$

Notice that the above result is obtained because $\dot{\mathbf{u}}(t)$ and $\dot{\mathbf{p}}(t)$ do not appear in g_a.

Next, let us consider only those terms inside the integral which involve the function h; these terms contain

$$\frac{\partial}{\partial \mathbf{x}}\left[\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t), t)\right]^T \dot{\mathbf{x}}^*(t) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t), t)\right] - \frac{d}{dt}\left\{\frac{\partial}{\partial \dot{\mathbf{x}}}\left[\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t), t)\right]^T \dot{\mathbf{x}}^*(t)\right]\right\}. \tag{5.1-10}$$

Writing out the indicated partial derivatives gives

$$\left[\frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}^*(t), t)\right]\dot{\mathbf{x}}^*(t) + \left[\frac{\partial^2 h}{\partial t \partial \mathbf{x}}(\mathbf{x}^*(t), t)\right] - \frac{d}{dt}\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t), t)\right], \tag{5.1-11}$$

or, if we apply the chain rule to the last term,

$$\left[\frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}^*(t), t)\right]\dot{\mathbf{x}}^*(t) + \left[\frac{\partial^2 h}{\partial t \partial \mathbf{x}}(\mathbf{x}^*(t), t)\right] - \left[\frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}^*(t), t)\right]\dot{\mathbf{x}}^*(t) - \left[\frac{\partial^2 h}{\partial \mathbf{x} \partial t}(\mathbf{x}^*(t), t)\right]. \tag{5.1-12}$$

If it is assumed that the second partial derivatives are continuous, the order of differentiation can be interchanged, and these terms add to zero. In the integral term we have, then,

$$\int_{t_0}^{t_f} \left\{\left[\left[\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T + \mathbf{p}^{*T}(t)\left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right] - \frac{d}{dt}[-\mathbf{p}^{*T}(t)]\right]\delta\mathbf{x}(t) + \left[\left[\frac{\partial g}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T + \mathbf{p}^{*T}(t)\left[\frac{\partial \mathbf{a}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]\right]\delta\mathbf{u}(t) + \left[[\mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) - \dot{\mathbf{x}}^*(t)]^T\right]\delta\mathbf{p}(t)\right\} dt. \tag{5.1-13}$$

This integral must vanish on an extremal regardless of the boundary conditions. We first observe that the constraints

$$\dot{\mathbf{x}}^*(t) = \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \tag{5.1-14a}$$

must be satisfied by an extremal so that the coefficient of $\delta\mathbf{p}(t)$ is zero. The Lagrange multipliers are arbitrary, so let us select them to make the coefficient of $\delta\mathbf{x}(t)$ equal to zero, that is,

$$\dot{\mathbf{p}}^*(t) = -\left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T \mathbf{p}^*(t) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t). \tag{5.1-14b}$$

We shall henceforth call (5.1-14b) the *costate equations* and $\mathbf{p}(t)$ the *costate*.

The remaining variation $\delta\mathbf{u}(t)$ is independent, so its coefficient must be zero; thus,

$$\mathbf{0} = \frac{\partial g}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) + \left[\frac{\partial \mathbf{a}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T \mathbf{p}^*(t). \tag{5.1-14c}$$

Equations (5.1-14) are important equations; we shall be using them throughout the remainder of this chapter. We shall find that even when the admissible controls are bounded, only Eq. (5.1-14c) is modified.

There are still the terms outside the integral to deal with; since the variation must be zero, we have

$$\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f)\right]^T \delta\mathbf{x}_f + \left[g(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) + \mathbf{p}^{*T}(t_f)[\mathbf{a}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), t_f)]\right]\delta t_f = 0. \tag{5.1-15}$$

In writing (5.1-15), we have used the fact that $\dot{\mathbf{x}}^*(t_f) = \mathbf{a}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), t_f)$. Equation (5.1-15) admits a variety of situations, which we shall discuss shortly.

Equations (5.1-14) are the necessary conditions we set out to determine. Notice that these necessary conditions consist of a set of $2n$, *first-order* differential equations—the state and costate equations (5.1-14a) and (5.1-14b)—and a set of m algebraic relations—(5.1-14c)—which must be satisfied throughout the interval $[t_0, t_f]$. The solution of the state and costate equations will contain $2n$ constants of integration. To evaluate these constants we use the n equations $\mathbf{x}^*(t_0) = \mathbf{x}_0$ and an additional set of n or $(n + 1)$ relationships —depending on whether or not t_f is specified—from Eq. (5.1-15). Notice that, as expected, we are again confronted by a two-point boundary-value problem.

In the following we shall find it convenient to use the function $\mathscr{H}$, called the *Hamiltonian*, defined as

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) \triangleq g(\mathbf{x}(t), \mathbf{u}(t), t) + \mathbf{p}^T(t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)]. \quad (5.1\text{-}16)$$

Using this notation, we can write the necessary conditions (5.1-14) through (5.1-15) as follows:

$$\left.\begin{aligned} \dot{\mathbf{x}}^*(t) &= \frac{\partial \mathscr{H}}{\partial \mathbf{p}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) && \text{(5.1-17a)} \\ \dot{\mathbf{p}}^*(t) &= -\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) && \text{(5.1-17b)} \\ \mathbf{0} &= \frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) && \text{(5.1-17c)} \end{aligned}\right\} \text{for all } t \in [t_0, t_f]$$

and

$$\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f)\right]^T \delta \mathbf{x}_f + \left[\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f)\right] \delta t_f = 0. \quad (5.1\text{-}18)$$

Let us now consider the boundary conditions that may occur.

Boundary Conditions

In a particular problem either g or h may be missing; in this case, we simply strike out the terms involving the missing function. To determine the boundary conditions is a matter of making the appropriate substitutions in Eq. (5.1-18). In all cases it will be assumed that we have the n equations $\mathbf{x}^*(t_0) = \mathbf{x}_0$.

Problems with Fixed Final Time. If the final time t_f is specified, $\mathbf{x}(t_f)$ may be specified, free, or required to lie on some surface in the state space.

CASE I. *Final state specified.* Since $\mathbf{x}(t_f)$ and t_f are specified, we substitute $\delta\mathbf{x}_f = \mathbf{0}$ and $\delta t_f = 0$ in (5.1-18). The required n equations are

$$\mathbf{x}^*(t_f) = \mathbf{x}_f. \tag{5.1-19}$$

CASE II. *Final state free.* We substitute $\delta t_f = 0$ in Eq. (5.1-18); since $\delta\mathbf{x}_f$ is arbitrary, the n equations

$$\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) - \mathbf{p}^*(t_f) = \mathbf{0}\dagger \tag{5.1-20}$$

must be satisfied.

CASE III. *Final state lying on the surface defined by* $\mathbf{m}(\mathbf{x}(t)) = \mathbf{0}$. Since this is a new situation, let us consider an introductory example. Suppose that the final state of a second-order system is required to lie on the circle

$$m(\mathbf{x}(t)) = [x_1(t) - 3]^2 + [x_2(t) - 4]^2 - 4 = 0 \tag{5.1-21}$$

shown in Fig. 5-1. Notice that admissible changes in $\mathbf{x}(t_f)$ are (to first-order) tangent to the circle at the point $(\mathbf{x}^*(t_f), t_f)$. The tangent line is normal to the gradient vector

$$\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) = \begin{bmatrix} 2[x_1^*(t_f) - 3] \\ 2[x_2^*(t_f) - 4] \end{bmatrix} \tag{5.1-22}$$

at the point $(\mathbf{x}^*(t_f), t_f)$. Thus, $\delta\mathbf{x}(t_f)$ must be normal to the gradient (5.1-22), so that

$$\left[\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right]^T \delta\mathbf{x}(t_f) = 2[x_1^*(t_f) - 3]\,\delta x_1(t_f) + 2[x_2^*(t_f) - 4]\,\delta x_2(t_f) = 0. \tag{5.1-23}$$

† Since the final time is fixed, h will not depend on t_f.

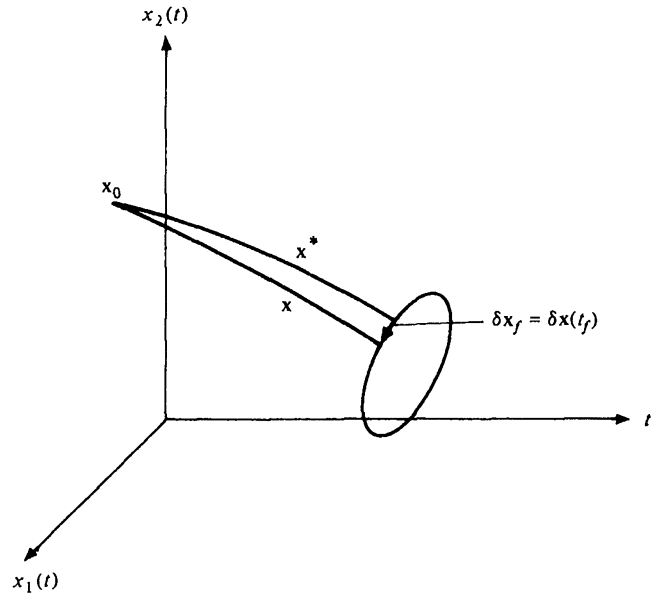


Figure 5-1 An extremal and a comparison curve that terminate on the curve $[x_1(t) - 3]^2 + [x_2(t) - 4]^2 - 4 = 0$ at the specified final time, t_f

Solving for $\delta x_2(t_f)$ gives

$$\delta x_2(t_f) = \frac{-[x_1^*(t_f) - 3]}{[x_2^*(t_f) - 4]} \delta x_1(t_f), \tag{5.1-24}$$

which, when substituted in Eq. (5.1-18), gives

$$\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) - \mathbf{p}^*(t_f)\right]^T \begin{bmatrix} 1 \\ \dfrac{-[x_1^*(t_f) - 3]}{[x_2^*(t_f) - 4]} \end{bmatrix} = 0 \tag{5.1-25}$$

since $\delta t_f = 0$ and $\delta x_1(t_f)$ is arbitrary. The second required equation at the final time is

$$m(\mathbf{x}^*(t_f)) = [x_1^*(t_f) - 3]^2 + [x_2^*(t_f) - 4]^2 - 4 = 0. \tag{5.1-26}$$

In the general situation there are n state variables and $1 \leq k \leq n - 1$ relationships that the states must satisfy at $t = t_f$. In this case we write

$$\mathbf{m}(\mathbf{x}(t)) = \begin{bmatrix} m_1(\mathbf{x}(t)) \\ \cdot \\ \cdot \\ \cdot \\ m_k(\mathbf{x}(t)) \end{bmatrix} = \mathbf{0}, \tag{5.1-27}$$

and each component of $\mathbf{m}$ represents a hypersurface in the n-dimensional state space. Thus, the final state lies on the intersection of these k hypersurfaces, and $\delta\mathbf{x}(t_f)$ is tangent to each of the hypersurfaces at the point $(\mathbf{x}^*(t_f), t_f)$. This means that $\delta\mathbf{x}(t_f)$ is normal to each of the gradient vectors

$$\frac{\partial m_1}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)), \ldots, \frac{\partial m_k}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)), \tag{5.1-28}$$

which are assumed to be linearly independent. From Eq. (5.1-18) we have, since $\delta t_f = 0$,

$$\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) - \mathbf{p}^*(t_f)\right]^T \delta\mathbf{x}(t_f) \triangleq \mathbf{v}^T \delta\mathbf{x}(t_f) = 0. \tag{5.1-29}$$

It can be shown that this equation is satisfied if and only if the vector $\mathbf{v}$ is a linear combination of the gradient vectors in Eq. (5.1-28), that is,

$$\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) - \mathbf{p}^*(t_f) = d_1\left[\frac{\partial m_1}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right] + \cdots + d_k\left[\frac{\partial m_k}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right]. \tag{5.1-30}$$

To determine the $2n$ constants of integration in the solution of the state-costate equations, and $d_1, \ldots, d_k$, we have the n equations $\mathbf{x}^*(t_0) = \mathbf{x}_0$, the n equations (5.1-30), and the k equations

$$\mathbf{m}(\mathbf{x}^*(t_f)) = \mathbf{0}. \tag{5.1-31}$$

Let us show that Eqs. (5.1-30) and (5.1-31) lead to the results obtained in our introductory example. The constraining relation is

$$m(\mathbf{x}(t)) = [x_1(t) - 3]^2 + [x_2(t) - 4]^2 - 4 = 0. \tag{5.1-21}$$

From Eq. (5.1-30) we obtain the two equations

$$\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) - \mathbf{p}^*(t_f) = d\begin{bmatrix} 2[x_1^*(t_f) - 3] \\ 2[x_2^*(t_f) - 4] \end{bmatrix}, \tag{5.1-32}$$

and (5.1-31) gives

$$m(\mathbf{x}^*(t_f)) = [x_1^*(t_f) - 3]^2 + [x_2^*(t_f) - 4]^2 - 4 = 0. \quad (5.1\text{-}33)$$

By solving the second of Eqs. (5.1-32) for d and substituting this into the first equation of (5.1-32), Eq. (5.1-25) is obtained.

Problems with Free Final Time. If the final time is free, there are several situations that may occur.

CASE I. *Final state fixed.* The appropriate substitution in Eq. (5.1-18) is $\delta\mathbf{x}_f = \mathbf{0}$. δt_f is arbitrary, so the $(2n + 1)$st relationship is

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = 0. \quad (5.1\text{-}34)$$

CASE II. *Final state free.* $\delta\mathbf{x}_f$ and δt_f are arbitrary and independent; therefore, their coefficients must be zero; that is,

$$\mathbf{p}^*(t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) \qquad (n \text{ equations}) \quad (5.1\text{-}35)$$

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = 0 \qquad (1 \text{ equation}). \quad (5.1\text{-}36)$$

Notice that if $h = 0$

$$\mathbf{p}^*(t_f) = \mathbf{0} \quad (5.1\text{-}37)$$

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) = 0. \quad (5.1\text{-}38)$$

CASE III. $\mathbf{x}(t_f)$ *lies on the moving point* $\boldsymbol{\theta}(t)$. Here $\delta\mathbf{x}_f$ and δt_f are related by

$$\delta\mathbf{x}_f = \left[\frac{d\boldsymbol{\theta}}{dt}(t_f)\right] \delta t_f;$$

making this substitution in Eq. (5.1-18) yields the equation

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) + \left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f)\right]^T \times \left[\frac{d\boldsymbol{\theta}}{dt}(t_f)\right] = 0. \quad (5.1\text{-}39)$$

This gives one equation; the remaining n required relationships are

$$\mathbf{x}^*(t_f) = \boldsymbol{\theta}(t_f).$$

CASE IV. *Final state lying on the surface defined by* $\mathbf{m}(\mathbf{x}(t)) = \mathbf{0}$. As an example of this type of end point constraint, suppose that the final state is required to lie on the curve

$$m(\mathbf{x}(t)) = [x_1(t) - 3]^2 + [x_2(t) - 4]^2 - 4 = 0. \tag{5.1-40}$$

Since the final time is free, the admissible end points lie on the cylindrical surface shown in Fig. 5-2. Notice that

1. To first-order, the change in $\mathbf{x}(t_f)$ must be in the plane tangent to the cylindrical surface at the point $(\mathbf{x}^*(t_f), t_f)$.
2. The change in $\mathbf{x}(t_f)$ is independent of δt_f.

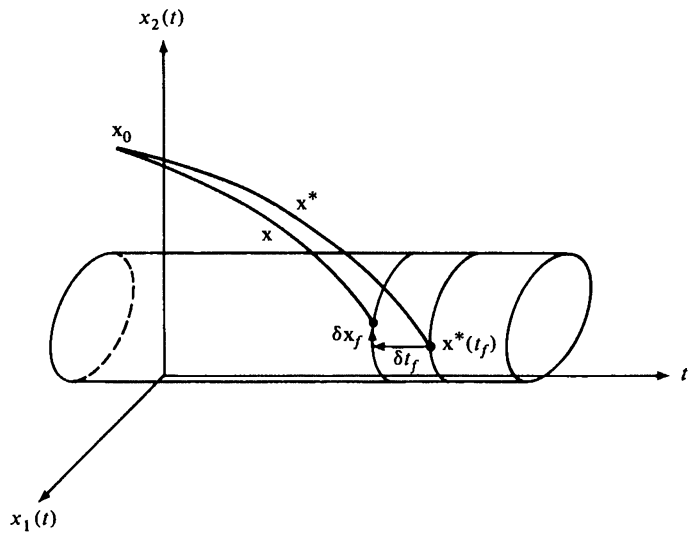


Figure 5-2 An extremal and a comparison curve that terminate on the surface $[x_1(t) - 3]^2 + [x_2(t) - 4]^2 - 4 = 0$

Since $\delta \mathbf{x}_f$ is independent of δt_f, the coefficient of δt_f must be zero, and

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = 0. \tag{5.1-41}$$

The plane that is tangent to the cylinder at the point $(\mathbf{x}^*(t_f), t_f)$ is described by its normal vector or gradient; that is, every vector in the plane is normal to the vector

$$\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) = \begin{bmatrix} 2[x_1^*(t_f) - 3] \\ 2[x_2^*(t_f) - 4] \end{bmatrix}. \tag{5.1-42}$$

This means that

$$\left[\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right]^T \delta\mathbf{x}_f = 2[x_1^*(t_f) - 3]\,\delta x_{1_f} + 2[x_2^*(t_f) - 4]\,\delta x_{2_f} = 0. \quad (5.1\text{-}43)$$

Solving for δx_{2_f} gives

$$\delta x_{2_f} = \frac{-[x_1^*(t_f) - 3]}{[x_2^*(t_f) - 4]}\,\delta x_{1_f}. \quad (5.1\text{-}44)$$

Substituting this for δx_{2_f} in Eq. (5.1-18) gives

$$\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f)\right]^T \begin{bmatrix} 1 \\ \dfrac{-[x_1^*(t_f) - 3]}{[x_2^*(t_f) - 4]} \end{bmatrix} \delta x_{1_f} = 0. \quad (5.1\text{-}45)$$

Since δx_{1_f} is arbitrary, its coefficient must be zero. Equations (5.1-41) and (5.1-45) give two relationships; the third is the constraint

$$m(\mathbf{x}^*(t_f)) = [x_1^*(t_f) - 3]^2 + [x_2^*(t_f) - 4]^2 - 4 = 0. \quad (5.1\text{-}46)$$

In the general situation we have n state variables, and there may be $1 \leq k \leq n - 1$ relationships that the states are required to satisfy at the terminal time. In this case we write

$$\mathbf{m}(\mathbf{x}(t)) = \begin{bmatrix} m_1(\mathbf{x}(t)) \\ \cdot \\ \cdot \\ \cdot \\ m_k(\mathbf{x}(t)) \end{bmatrix} = \mathbf{0}, \quad (5.1\text{-}47)$$

and each component of $\mathbf{m}$ describes a hypersurface in the n-dimensional state space. This means that the final state lies on the intersection of the hypersurfaces defined by $\mathbf{m}$, and that $\delta\mathbf{x}_f$ is (to first order) tangent to each of the hypersurfaces at the point $(\mathbf{x}^*(t_f), t_f)$. Thus, $\delta\mathbf{x}_f$ is normal to each of the gradient vectors

$$\frac{\partial m_1}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)), \ldots, \frac{\partial m_k}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)), \quad (5.1\text{-}48)$$

which we assume to be linearly independent. It is left as an exercise for the reader to show that the reasoning used in Case III with *fixed* final time also applies in the present situation and leads to the $(2n + k + 1)$ equations

$$\mathbf{x}^*(t_0) = \mathbf{x}_0$$

$$\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f) = d_1\left[\frac{\partial m_1}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right] + \cdots + d_k\left[\frac{\partial m_k}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right]$$

$$\mathbf{m}(\mathbf{x}^*(t_f)) = \mathbf{0}$$

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = 0 \qquad (5.1\text{-}49)$$

involving the $2n$ constants of integration, the variables $d_1, \ldots, d_k$, and t_f. It is also easily shown that Eqs. (5.1-49) give Eqs. (5.1-41), (5.1-45), and (5.1-46) in the preceding example.

CASE V. *Final state lying on the moving surface defined by* $\mathbf{m}(\mathbf{x}(t), t) = \mathbf{0}$. Suppose that the final state must lie on the surface

$$m(\mathbf{x}(t), t) = [x_1(t) - 3]^2 + [x_2(t) - 4 - t]^2 - 4 = 0 \qquad (5.1\text{-}50)$$

shown in Fig. 5-3. Notice that δt_f does influence the admissible values of $\delta \mathbf{x}_f$; that is, to remain on the surface $m(\mathbf{x}(t), t) = 0$ the value of $\delta \mathbf{x}_f$ depends on δt_f. The vector with components $\delta x_{1_f}, \delta x_{2_f}, \delta t_f$ must be contained in a plane tangent to the surface at the point $(\mathbf{x}^*(t_f), t_f)$. This means that the normal to this tangent plane is the vector

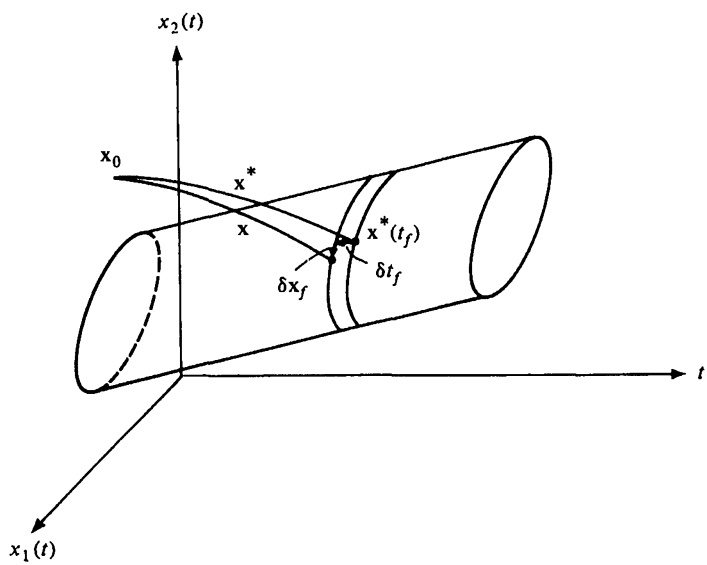


Figure 5-3 An extremal and a comparison curve that terminate on the surface $[x_1(t) - 3]^2 + [x_2(t) - 4 - t]^2 - 4 = 0$

$$\begin{bmatrix} \dfrac{\partial m}{\partial x_1}(\mathbf{x}^*(t_f), t_f) \\ \dfrac{\partial m}{\partial x_2}(\mathbf{x}^*(t_f), t_f) \\ \dfrac{\partial m}{\partial t}(\mathbf{x}^*(t_f), t_f) \end{bmatrix} \triangleq \begin{bmatrix} \dfrac{\partial m}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) \\ \hline \dfrac{\partial m}{\partial t}(\mathbf{x}^*(t_f), t_f) \end{bmatrix} \tag{5.1-51}$$

in the three-dimensional space. Thus, admissible variations must be normal to the vector (5.1-51), so

$$\left[\frac{\partial m}{\partial x_1}(\mathbf{x}^*(t_f), t_f)\right] \delta x_{1_f} + \left[\frac{\partial m}{\partial x_2}(\mathbf{x}^*(t_f), t_f)\right] \delta x_{2_f} + \left[\frac{\partial m}{\partial t}(\mathbf{x}^*(t_f), t_f)\right] \delta t_f = 0. \tag{5.1-52}$$

For the surface specified we have

$$2[x_1^*(t_f) - 3]\, \delta x_{1_f} + 2[x_2^*(t_f) - 4 - t_f]\, \delta x_{2_f} - 2[x_2^*(t_f) - 4 - t_f]\, \delta t_f = 0. \tag{5.1-53}$$

Solving for δt_f gives

$$\delta t_f = \frac{[x_1^*(t_f) - 3]}{[x_2^*(t_f) - 4 - t_f]} \delta x_{1_f} + \delta x_{2_f}. \tag{5.1-54}$$

Substituting in Eq. (5.1-18) and collecting terms, we obtain

$$\begin{aligned} &\left[\frac{\partial h}{\partial x_1}(\mathbf{x}^*(t_f), t_f) - p_1^*(t_f) + \left[\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f)\right.\right. \\ &\quad \left.\left. + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f)\right]\left[\frac{x_1^*(t_f) - 3}{x_2^*(t_f) - 4 - t_f}\right]\right] \delta x_{1_f} \\ &\quad + \left[\frac{\partial h}{\partial x_2}(\mathbf{x}^*(t_f), t_f) - p_2^*(t_f) + \mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f)\right. \\ &\quad \left. + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f)\right] \delta x_{2_f} = 0. \end{aligned} \tag{5.1-55}$$

Since there is one constraint involving the three variables $(\delta x_{1_f}, \delta x_{2_f}, \delta t_f)$, δx_{1_f} and δx_{2_f} can be varied independently; therefore, the coefficients of δx_{1_f} and δx_{2_f} must be zero. This gives two equations; the third equation is

$$m(\mathbf{x}^*(t_f), t_f) = 0. \tag{5.1-56}$$

In general, we may have $1 \leq k \leq n$ relationships

$$\mathbf{m}(\mathbf{x}(t), t) = \begin{bmatrix} m_1(\mathbf{x}(t), t) \\ \cdot \\ \cdot \\ \cdot \\ m_k(\mathbf{x}(t), t) \end{bmatrix} = \mathbf{0}, \tag{5.1-57}$$

which must be satisfied by the $(n + 1)$ variables $\mathbf{x}(t_f)$ and t_f. Reasoning as in the situation where $\mathbf{m}$ is not dependent on time, we deduce that the admissible values of the $(n + 1)$ vector

$$\begin{bmatrix} \delta\mathbf{x}_f \\ \hline \delta t_f \end{bmatrix}$$

are normal to each of the gradient vectors

$$\begin{bmatrix} \dfrac{\partial m_1}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) \\ \hline \dfrac{\partial m_1}{\partial t}(\mathbf{x}^*(t_f), t_f) \end{bmatrix}, \ldots, \begin{bmatrix} \dfrac{\partial m_k}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) \\ \hline \dfrac{\partial m_k}{\partial t}(\mathbf{x}^*(t_f), t_f) \end{bmatrix}, \tag{5.1-58}$$

which are assumed to be linearly independent. Writing Eq. (5.1-18) as

$$\begin{bmatrix} \dfrac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f) \\ \hline \mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \dfrac{\partial h}{\partial t}(\mathbf{x}^*(t_f)\, t_f) \end{bmatrix}^T \begin{bmatrix} \delta\mathbf{x}_f \\ \hline \delta t_f \end{bmatrix} = 0 \triangleq \mathbf{v}^T \begin{bmatrix} \delta\mathbf{x}_f \\ \hline \delta t_f \end{bmatrix} \tag{5.1-59}$$

and again using the result that $\mathbf{v}$ must be a linear combination of the gradient vectors in (5.1-58), we obtain

$$\mathbf{v} = d_1 \begin{bmatrix} \dfrac{\partial m_1}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) \\ \hline \dfrac{\partial m_1}{\partial t}(\mathbf{x}^*(t_f), t_f) \end{bmatrix} + \cdots + d_k \begin{bmatrix} \dfrac{\partial m_k}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) \\ \hline \dfrac{\partial m_k}{\partial t}(\mathbf{x}^*(t_f), t_f) \end{bmatrix}, \tag{5.1-60}$$

or

$$\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f) = d_1 \left[\frac{\partial m_1}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f)\right] + \cdots + d_k \left[\frac{\partial m_k}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f)\right]$$

and

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = d_1\left[\frac{\partial m_1}{\partial t}(\mathbf{x}^*(t_f), t_f)\right] + \cdots + d_k\left[\frac{\partial m_k}{\partial t}(\mathbf{x}^*(t_f), t_f)\right]. \tag{5.1-61}$$

Equations (5.1-61), the k equations

$$\mathbf{m}(\mathbf{x}^*(t_f), t_f) = \mathbf{0}, \tag{5.1-62}$$

and the n equations $\mathbf{x}^*(t_0) = \mathbf{x}_0$ comprise a set of $(2n + k + 1)$ equations in the $2n$ constants of integration, the variables $d_1, d_2, \ldots, d_k$, and t_f. It is left as an exercise for the reader to verify that (5.1-62) and (5.1-61) yield Eqs. (5.1-55) and (5.1-56).

The boundary conditions which we have discussed are summarized in Table 5-1. Of course, mixed situations can arise, but these can be handled by returning to Eq. (5.1-18) and applying the ideas introduced in the preceding discussion.

Although the boundary condition relationships may look foreboding, setting up the equations is not difficult; obtaining solutions is another matter. This should not surprise us, however, for we already suspect that numerical techniques are required to solve most problems of practical interest. Let us now illustrate the determination of the boundary-condition equations by considering several examples.

Example 5.1-1. The system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -x_2(t) + u(t) \end{aligned} \tag{5.1-63}$$

is to be controlled so that its control effort is conserved; that is, the performance measure

$$J(u) = \int_{t_0}^{t_f} \tfrac{1}{2}u^2(t)\, dt \tag{5.1-64}$$

is to be minimized. The admissible states and controls are not bounded. Find necessary conditions that must be satisfied for optimal control.

The first step is to form the Hamiltonian

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = \tfrac{1}{2}u^2(t) + p_1(t)x_2(t) - p_2(t)x_2(t) + p_2(t)u(t). \tag{5.1-65}$$

From Eqs. (5.1-17b) and (5.1-17c) necessary conditions for optimality are

$$\begin{aligned} \dot{p}_1^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_1} = 0 \\ \dot{p}_2^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_2} = -p_1^*(t) + p_2^*(t), \end{aligned} \tag{5.1-66}$$

and

$$0 = \frac{\partial \mathscr{H}}{\partial u} = u^*(t) + p_2^*(t). \tag{5.1-67}$$

If Eq. (5.1-67) is solved for $u^*(t)$ and substituted into the state equations (5.1-63), we have

$$\begin{aligned} \dot{x}_1^*(t) &= x_2^*(t) \\ \dot{x}_2^*(t) &= -x_2^*(t) - p_2^*(t). \end{aligned} \tag{5.1-68}$$

Equations (5.1-68) and (5.1-66)—the state and costate equations—are a set of $2n$ linear first-order, homogeneous, constant-coefficient differential equations. Solving these equations gives

$$\begin{aligned} x_1^*(t) &= c_1 + c_2[1 - \epsilon^{-t}] + c_3[-t - \tfrac{1}{2}\epsilon^{-t} + \tfrac{1}{2}\epsilon^{t}] \\ &\quad + c_4[1 - \tfrac{1}{2}\epsilon^{-t} - \tfrac{1}{2}\epsilon^{t}] \\ x_2^*(t) &= c_2\epsilon^{-t} + c_3[-1 + \tfrac{1}{2}\epsilon^{-t} + \tfrac{1}{2}\epsilon^{t}] + c_4[\tfrac{1}{2}\epsilon^{-t} - \tfrac{1}{2}\epsilon^{t}] \\ p_1^*(t) &= c_3 \\ p_2^*(t) &= c_3[1 - \epsilon^{t}] + c_4\epsilon^{t}. \end{aligned} \tag{5.1-69}$$

Now let us consider several possible sets of boundary conditions.

a. Suppose $\mathbf{x}(0) = \mathbf{0}$ and $\mathbf{x}(2) = [5 \quad 2]^T$. From $\mathbf{x}(0) = \mathbf{0}$ we obtain $c_1 = c_2 = 0$; the remaining two equations to be solved are

$$\begin{aligned} 5 &= c_3[-2 - \tfrac{1}{2}\epsilon^{-2} + \tfrac{1}{2}\epsilon^{2}] + c_4[1 - \tfrac{1}{2}\epsilon^{-2} - \tfrac{1}{2}\epsilon^{2}] \\ 2 &= c_3[-1 + \tfrac{1}{2}\epsilon^{-2} + \tfrac{1}{2}\epsilon^{2}] + c_4[\tfrac{1}{2}\epsilon^{-2} - \tfrac{1}{2}\epsilon^{2}]. \end{aligned} \tag{5.1-70}$$

Solving these linear algebraic equations gives $c_3 = -7.289$ and $c_4 = -6.103$, so the optimal trajectory is

$$\begin{aligned} x_1^*(t) &= 7.289t - 6.103 + 6.696\epsilon^{-t} - 0.593\epsilon^{t} \\ x_2^*(t) &= 7.289 - 6.696\epsilon^{-t} - 0.593\epsilon^{t}. \end{aligned} \tag{5.1-71}$$

b. Let $\mathbf{x}(0) = \mathbf{0}$ and $\mathbf{x}(2)$ be unspecified; consider the performance measure

$$J(u) = \tfrac{1}{2}[x_1(2) - 5]^2 + \tfrac{1}{2}[x_2(2) - 2]^2 + \tfrac{1}{2}\int_0^2 u^2(t)\,dt. \tag{5.1-72}$$

Table 5-1 SUMMARY OF BOUNDARY CONDITIONS IN OPTIMAL CONTROL PROBLEMS

Problem	*Description*	*Substitution in Eq. (5.1-18)*	*Boundary-condition equations*	*Remarks*
t_f fixed	1. $\mathbf{x}(t_f) = \mathbf{x}_f$ specified final state	$\delta\mathbf{x}_f = \delta\mathbf{x}(t_f) = \mathbf{0}$ $\delta t_f = 0$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\mathbf{x}^*(t_f) = \mathbf{x}_f$	$2n$ equations to determine $2n$ constants of integration
	2. $\mathbf{x}(t_f)$ free	$\delta\mathbf{x}_f = \delta\mathbf{x}(t_f)$ $\delta t_f = 0$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) - \mathbf{p}^*(t_f) = \mathbf{0}$	$2n$ equations to determine $2n$ constants of integration
	3. $\mathbf{x}(t_f)$ on the surface $\mathbf{m}(\mathbf{x}(t)) = \mathbf{0}$	$\delta\mathbf{x}_f = \delta\mathbf{x}(t_f)$ $\delta t_f = 0$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)) - \mathbf{p}^*(t_f) = \sum_{i=1}^{k} d_i\left[\frac{\partial m_i}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right]$ $\mathbf{m}(\mathbf{x}^*(t_f)) = \mathbf{0}$	$(2n + k)$ equations to determine the $2n$ constants of integration and the variables $d_1, \ldots, d_k$
t_f free	4. $\mathbf{x}(t_f) = \mathbf{x}_f$ specified final state	$\delta\mathbf{x}_f = \mathbf{0}$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\mathbf{x}^*(t_f) = \mathbf{x}_f$ $\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = 0$	$(2n + 1)$ equations to determine the $2n$ constants of integration and t_f
	5. $\mathbf{x}(t_f)$ free		$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f) = \mathbf{0}$ $\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = 0$	$(2n + 1)$ equations to determine the $2n$ constants of integration and t_f
	6. $\mathbf{x}(t_f)$ on the moving point $\boldsymbol{\theta}(t)$	$\delta\mathbf{x}_f = \left[\frac{d\boldsymbol{\theta}}{dt}(t_f)\right]\delta t_f$	$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\mathbf{x}^*(t_f) = \boldsymbol{\theta}(t_f)$ $\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f)$ $+ \left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f)\right]^T \left[\frac{d\boldsymbol{\theta}}{dt}(t_f)\right] = 0$	$(2n + 1)$ equations to determine the $2n$ constants of integration and t_f

7. $\mathbf{x}(t_f)$ on the surface $\mathbf{m}(\mathbf{x}(t)) = \mathbf{0}$		$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f) = \sum_{i=1}^{k} d_i \left[\frac{\partial m_i}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right]$ $\mathbf{m}(\mathbf{x}^*(t_f)) = \mathbf{0}$ $\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f) = 0$	$(2n + k + 1)$ equations to determine the $2n$ constants of integration, the variables $d_1, \ldots, d_k$, and t_f
8. $\mathbf{x}(t_f)$ on the moving surface $\mathbf{m}(\mathbf{x}(t), t) = \mathbf{0}$		$\mathbf{x}^*(t_0) = \mathbf{x}_0$ $\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f) = \sum_{i=1}^{k} d_i \left[\frac{\partial m_i}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f)\right]$ $\mathbf{m}(\mathbf{x}^*(t_f), t_f) = \mathbf{0}$ $\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f)$ $= \sum_{i=1}^{k} d_i \left[\frac{\partial m_i}{\partial t}(\mathbf{x}^*(t_f), t_f)\right]$	$(2n + k + 1)$ equations to determine the $2n$ constants of integration, the variables $d_1, \ldots, d_k$, and t_f.

The modified performance measure affects only the boundary conditions at $t = 2$. From entry 2 of Table 5-1 we have

$$\begin{aligned} p_1^*(2) &= x_1^*(2) - 5 \\ p_2^*(2) &= x_2^*(2) - 2. \end{aligned} \tag{5.1-73}$$

c_1 and c_2 are again zero because $\mathbf{x}^*(0) = \mathbf{0}$. Putting $t = 2$ in Eq. (5.1-69) and substituting in (5.1-73), we obtain the linear algebraic equations

$$\begin{bmatrix} 0.627 & -2.762 \\ 9.151 & -11.016 \end{bmatrix} \begin{bmatrix} c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}. \tag{5.1-74}$$

Solving these equations, we find that $c_3 = -2.697$, $c_4 = -2.422$; hence,

$$\begin{aligned} x_1^*(t) &= 2.697t - 2.422 + 2.560\epsilon^{-t} - 0.137\epsilon^{t} \\ x_2^*(t) &= 2.697 - 2.560\epsilon^{-t} - 0.137\epsilon^{t}. \end{aligned} \tag{5.1-75}$$

c. Next, suppose that the system is to be transferred from $\mathbf{x}(0) = \mathbf{0}$ to the line

$$x_1(t) + 5x_2(t) = 15 \tag{5.1-76}$$

while the original performance measure (5.1-64) is minimized. As before, the solution of the state and costate equations is given by Eq. (5.1-69), and $c_1 = c_2 = 0$. The boundary conditions at $t = 2$ are, from entry 3 of Table 5-1,

$$\begin{aligned} x_1^*(2) + 5x_2^*(2) &= 15 \\ -p_1^*(2) &= d \\ -p_2^*(2) &= 5d. \end{aligned} \tag{5.1-77}$$

Eliminating d and substituting $t = 2$ in (5.1-69), we obtain the equations

$$\begin{bmatrix} 15.437 & -20.897 \\ 11.389 & -7.389 \end{bmatrix} \begin{bmatrix} c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 15 \\ 0 \end{bmatrix}, \tag{5.1-78}$$

which have the solution $c_3 = -0.894$, $c_4 = -1.379$. The optimal trajectory is then

$$\begin{aligned} x_1^*(t) &= 0.894t - 1.379 + 1.136\epsilon^{-t} + 0.242\epsilon^{t} \\ x_2^*(t) &= 0.894 - 1.136\epsilon^{-t} + 0.242\epsilon^{t}. \end{aligned} \tag{5.1-79}$$

Example 5.1-2. The space vehicle shown in Fig. 5-4 is in the gravity field of the moon. Assume that the motion is planar, that aerodynamic forces are negligible, and that the thrust magnitude T is constant. The control

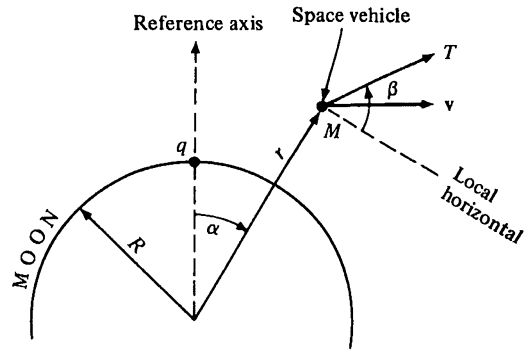


Figure 5-4 A space vehicle in the gravity field of the moon

variable is the thrust direction $\beta(t)$, which is measured from the local horizontal. To simplify the state equations, we shall approximate the vehicle as a particle of mass M. The gravitational force exerted on the vehicle is $F_g(t) = Mg_0R^2/r^2(t)$; g_0 is the gravitational constant at the surface of the moon, R is the radius of the moon, and r is the distance of the spacecraft from the center of the moon. The instantaneous velocity of the vehicle is the vector v, and α is the angular displacement from the reference axis. Selecting $x_1 \triangleq r$, $x_2 \triangleq \alpha$, $x_3 \triangleq \dot{r}$, and $x_4 \triangleq r\dot{\alpha}$ as the states of the system, letting $u \triangleq \beta$, and neglecting the change in mass resulting from fuel consumption, we find that the state equations are

$$\begin{aligned}
\dot{x}_1(t) &= x_3(t) \\
\dot{x}_2(t) &= \frac{x_4(t)}{x_1(t)} \\
\dot{x}_3(t) &= \frac{x_4^2(t)}{x_1(t)} - \frac{g_0R^2}{x_1^2(t)} + \left[\frac{T}{M}\right] \sin u(t) \\
\dot{x}_4(t) &= -\frac{x_3(t)x_4(t)}{x_1(t)} + \left[\frac{T}{M}\right] \cos u(t).
\end{aligned} \tag{5.1-80}$$

Notice that these differential equations are nonlinear in both the states and the control variable. Let us consider several possible missions for the space vehicle.

Mission a. Suppose that the spacecraft is to be launched from the point q on the reference axis at $t = 0$ into a circular orbit of altitude D, as shown in Fig. 5-5(a), in minimum time. $\alpha(t_f)$ is unspecified, and the vehicle starts from rest; thus, the initial conditions are $\mathbf{x}(0) = [R \quad 0 \quad 0 \quad 0]^T$.

From the performance measure

$$J(u) = \int_0^{t_f} dt \tag{5.1-81}$$

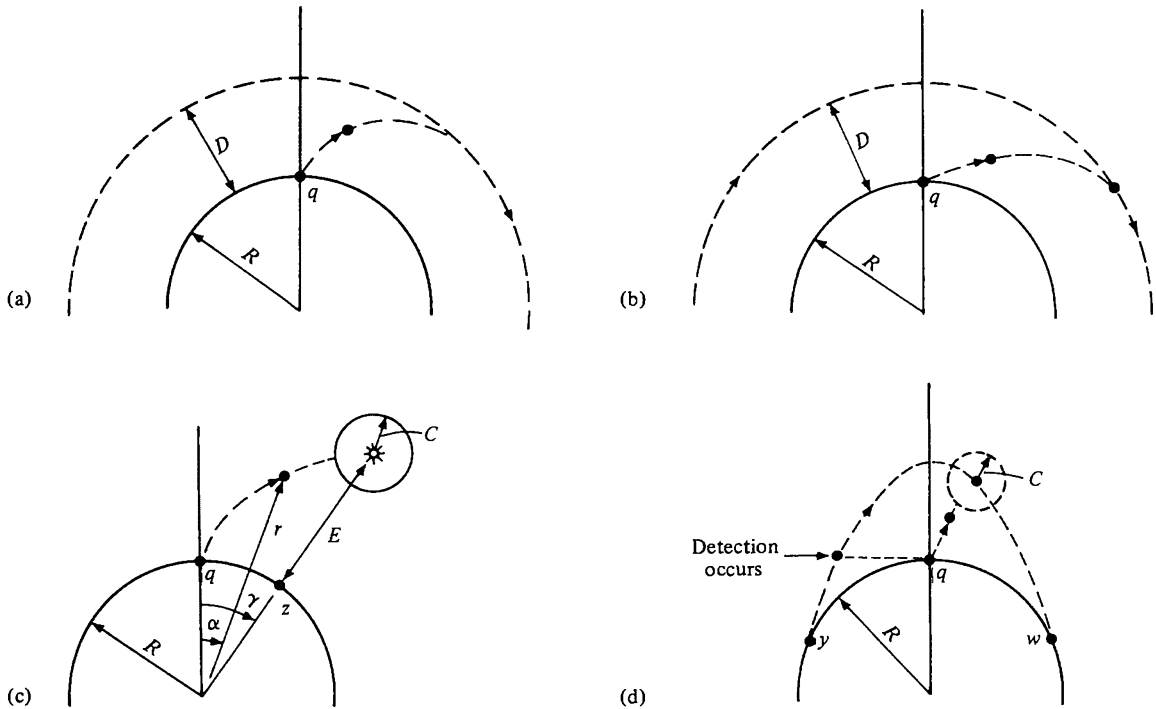


Figure 5-5 (a) Orbit injection. (b) Rendezvous. (c) Reconnaissance of synchronous satellite. (d) Reconnaissance of approaching spacecraft

and the state equations, the Hamiltonian is

$$\begin{aligned}\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) &= 1 + p_1(t)x_3(t) + \frac{p_2(t)x_4(t)}{x_1(t)} \\ &+ p_3(t)\left[\frac{x_4^2(t)}{x_1(t)} - \frac{g_0R^2}{x_1^2(t)} + \left[\frac{T}{M}\right]\sin u(t)\right] \\ &+ p_4(t)\left[\frac{-x_3(t)x_4(t)}{x_1(t)} + \left[\frac{T}{M}\right]\cos u(t)\right].^{\dagger}\end{aligned} \tag{5.1-82}$$

The costate equations are, from (5.1-17b),

$$\begin{aligned}\dot{p}_1^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_1} = \frac{p_2^*(t)x_4^*(t)}{x_1^{*2}(t)} + p_3^*(t)\left[\frac{x_4^{*2}(t)}{x_1^{*2}(t)} - \frac{2g_0R^2}{x_1^{*3}(t)}\right] - \frac{p_4^*(t)x_3^*(t)x_4^*(t)}{x_1^{*2}(t)} \\ \dot{p}_2^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_2} = 0 \\ \dot{p}_3^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_3} = -p_1^*(t) + \frac{p_4^*(t)x_4^*(t)}{x_1^*(t)} \\ \dot{p}_4^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_4} = -\frac{p_2^*(t)}{x_1^*(t)} - \frac{2p_3^*(t)x_4^*(t)}{x_1^*(t)} + \frac{p_4^*(t)x_3^*(t)}{x_1^*(t)}.\end{aligned} \tag{5.1-83}$$

The state equations

$$\dot{\mathbf{x}}^*(t) = \mathbf{a}(\mathbf{x}^*(t), u^*(t)) \tag{5.1-84}$$

must be satisfied by an optimal trajectory, and Eq. (5.1-17c) gives the algebraic relationship

$$0 = \frac{\partial \mathscr{H}}{\partial u} = \left[\frac{T}{M}\right][p_3^*(t)\cos u^*(t) - p_4^*(t)\sin u^*(t)]. \tag{5.1-85}$$

Solving Eq. (5.1-85) for $u^*(t)$ gives

$$u^*(t) = \tan^{-1}\frac{p_3^*(t)}{p_4^*(t)}, \tag{5.1-86}$$

or, equivalently,

$$\sin u^*(t) = \frac{p_3^*(t)}{\sqrt{p_3^{*2}(t) + p_4^{*2}(t)}} \tag{5.1-87a}$$

$$\cos u^*(t) = \frac{p_4^*(t)}{\sqrt{p_3^{*2}(t) + p_4^{*2}(t)}}. \tag{5.1-87b}$$

† Notice that $\mathscr{H}$ is not explicitly dependent on time; hence, the argument t is omitted.

By substituting (5.1-87a) and (5.1-87b) in the state equations, $u^*(t)$ can be eliminated; unfortunately, as is often the case, the resulting $2n$ first-order differential equations are nonlinear.

Next, let us determine the boundary conditions at the final time. There will be five relationships to be satisfied at $t = t_f$; hence, the initial and final boundary conditions will give nine equations involving the eight constants of integration and t_f. From the problem statement we know that $x_1^*(t_f)$ must equal $R + D$. In addition, to have a circular orbit, the centrifugal force must be exactly balanced by the gravitational force; therefore, $M[r^*(t)\dot{\alpha}^*(t)]^2/r^*(t) = Mg_0R^2/r^{*2}(t)$ for $t \geq t_f$. Evaluating this expression at $t = t_f$ and using the specified value of $x_1^*(t_f)$, we obtain $x_4^*(t_f) = \sqrt{g_0R^2/[R + D]}$. The radial velocity must be zero at $t = t_f$, so $x_3^*(t_f) = 0$. The final time is not related to the unspecified final state value $x_2^*(t_f)$, so in Eq. (5.1-18) the coefficients of δt_f and δx_{2f} must both be zero. To summarize, the required boundary condition relationships are

$$\begin{aligned} x_1^*(t_f) &= R + D \\ p_2^*(t_f) &= 0 \\ x_3^*(t_f) &= 0 \\ x_4^*(t_f) &= \sqrt{\frac{g_0R^2}{[R + D]}} \\ \mathscr{H}(\mathbf{x}^*(t_f), \mathbf{p}^*(t_f)) &= 0 \end{aligned} \tag{5.1-88}$$

In writing the last equation it has been assumed that $u^*(t)$ has been eliminated from the Hamiltonian by using Eqs. (5.1-87).

Mission b. In this mission, shown in Fig. 5-5(b), the space vehicle is to be launched from point q and is to rendezvous with another spacecraft that is in a fixed circular orbit D miles above the moon with a period of two hours. At $t = 0$ both spacecraft are on the reference axis. The rendezvous is to be accomplished in minimum time.

Only the boundary conditions are changed from Mission a. The final state values of the controlled vehicle must lie on the moving point

$$\boldsymbol{\theta}(t) = \begin{bmatrix} R + D \\ \text{modulo}_{2\pi}(\pi t) \\ 0 \\ \pi[R + D] \end{bmatrix}.$$

Modulo$_{2\pi}$ (πt) means that after each revolution 2π radians are subtracted from the angular displacement of the spacecraft. Only the final value of x_2 depends on t, so we have

$$\delta x_{2_f} = \left[\frac{d\theta_2}{dt}(t_f)\right] \delta t_f$$
$$= \pi\, \delta t_f. \tag{5.1-89}$$

Thus, from Eq. (5.1-18), or entry 6 of Table 5-1,

$$-\pi p_2^*(t_f) + \mathscr{H}(\mathbf{x}^*(t_f), \mathbf{p}^*(t_f)) = 0, \tag{5.1-90}$$

since $h = 0$. The remaining four boundary relationships are

$$\mathbf{x}^*(t_f) = \begin{bmatrix} R + D \\ \text{modulo } (\pi t_f) \\ 2\pi \\ 0 \\ \pi[R + D] \end{bmatrix} = \boldsymbol{\theta}(t_f). \tag{5.1-91}$$

Mission c. A satellite is in synchronous orbit E miles above the point z shown in Fig. 5-5(c). It is desired to investigate this satellite with a spacecraft as quickly as possible. The spacecraft transmits television pictures to the lunar base upon arriving at a distance of C miles from the satellite.

Again, the state and costate equations and Eq. (5.1-85) remain unchanged. For this mission, however, the final states must lie on the curve given by

$$m(\mathbf{x}(t)) = [r(t) \cos \alpha(t) - [R + E] \cos \gamma]^2$$
$$+ [r(t) \sin \alpha(t) - [R + E] \sin \gamma]^2 - C^2 = 0. \tag{5.1-92}$$

Since the curve $m(\mathbf{x}(t))$ does not depend explicitly on t, we have from entry 7 of Table 5-1 (putting $h = 0$),

$$-\mathbf{p}^*(t_f) = d\left[\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}^*(t_f))\right].$$

Performing the gradient operation, and simplifying, we obtain

$$\begin{aligned} -\mathbf{p}^*(t_f) &= d \begin{bmatrix} 2r^*(t_f) - 2[R + E] \cos (\alpha^*(t_f) - \gamma) \\ 2r^*(t_f)[R + E] \sin (\alpha^*(t_f) - \gamma) \\ 0 \\ 0 \end{bmatrix} \\ &= 2d \begin{bmatrix} x_1^*(t_f) - [R + E] \cos (x_2^*(t_f) - \gamma) \\ x_1^*(t_f)[R + E] \sin (x_2^*(t_f) - \gamma) \\ 0 \\ 0 \end{bmatrix}, \end{aligned} \tag{5.1-93}$$

where d is an unknown variable.

Thus, $p_3^*(t_f) = 0$ and $p_4^*(t_f) = 0$. The other boundary condition equations are

$$m(\mathbf{x}^*(t_f)) = [x_1^*(t_f)\cos x_2^*(t_f) - [R + E]\cos\gamma]^2 + [x_1^*(t_f)\sin x_2^*(t_f) - [R + E]\sin\gamma]^2 - C^2 = 0, \quad (5.1\text{-}94)$$

and

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{p}^*(t_f)) = 0. \quad (5.1\text{-}95)$$

Equations (5.1-93) through (5.1-95) and $\mathbf{x}^*(0) = [R \quad 0 \quad 0 \quad 0]^T$ give a total of ten equations involving the eight constants of integration, the variable d, and t_f.

Mission d. A lunar-based radar operator detects an approaching spacecraft at $t = 0$ in the position shown in Fig. 5-5(d), and at this time a reconnaissance spacecraft is dispatched from point q. The reconnaissance vehicle is to close to a distance of C miles of the approaching spacecraft as quickly as possible, and relay television pictures to the lunar base. From the radar data the position history of the approaching spacecraft is

$$m(\mathbf{x}(t), t) = [r(t)\cos\alpha(t) - 2.78Rt + 6.95Rt^2 - R]^2 + [r(t)\sin\alpha(t) - 1.85Rt + 0.32R]^2 - C^2 = 0. \quad (5.1\text{-}96)$$

It is to be assumed that this position history will not change.
From Table 5-1, entry 8, we have

$$-\mathbf{p}^*(t_f) = d\left[\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f)\right]. \quad (5.1\text{-}97)$$

Performing the gradient operation and simplifying, we obtain

$$\begin{aligned}
-p_1^*(t_f) &= 2d[x_1^*(t_f) + R\{[-2.78t_f + 6.95t_f^2 - 1]\cos x_2^*(t_f) \\
&\quad + [-1.85t_f + 0.32]\sin x_2^*(t_f)\}] \\
-p_2^*(t_f) &= -2d[Rx_1^*(t_f)\{[-2.78t_f + 6.95t_f^2 - 1]\sin x_2^*(t_f) \\
&\quad + [1.85t_f - 0.32]\cos x_2^*(t_f)\}] \\
-p_3^*(t_f) &= 0 \\
-p_4^*(t_f) &= 0.
\end{aligned} \quad (5.1\text{-}98)$$

In addition, the specified constraint

$$[x_1^*(t_f)\cos x_2^*(t_f) - 2.78Rt_f + 6.95Rt_f^2 - R]^2 + [x_1^*(t_f)\sin x_2^*(t_f) - 1.85Rt_f + 0.32R]^2 - C^2 = 0, \quad (5.1\text{-}99)$$

must be satisfied and, from Table 5-1,

$$\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{p}^*(t_f)) = 2\,dR\{[-2.78 + 13.9t_f][x_1^*(t_f)\cos x_2^*(t_f) - 2.78Rt_f + 6.95Rt_f^2 - R] - 1.85[x_1^*(t_f)\sin x_2^*(t_f) - 1.85Rt_f + 0.32R]\}. \quad (5.1\text{-}100)$$

With the specified initial conditions, we have ten equations in ten unknowns.

5.2 LINEAR REGULATOR PROBLEMS

In this section we shall consider an important class of optimal control problems—linear regulator systems. We shall show that for linear regulator problems the optimal control law can be found as a linear time-varying function of the system states. Under certain conditions, which we shall discuss, the optimal control law becomes time-invariant. The results presented here are primarily due to R. E. Kalman.†

The plant is described by the linear state equations

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \quad (5.2\text{-}1)$$

which may have time-varying coefficients. The performance measure to be minimized is

$$J = \tfrac{1}{2}\mathbf{x}^T(t_f)\mathbf{H}\mathbf{x}(t_f) + \tfrac{1}{2}\int_{t_0}^{t_f}[\mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t)]\,dt; \quad (5.2\text{-}2)$$

the final time t_f is fixed, $\mathbf{H}$ and $\mathbf{Q}$ are real symmetric positive semi-definite matrices, and $\mathbf{R}$ is a real symmetric positive definite matrix. It is assumed that the states and controls are not bounded, and $\mathbf{x}(t_f)$ is free. We attach the following physical interpretation to this performance measure: It is desired to maintain the state vector close to the origin without an excessive expenditure of control effort.

The Hamiltonian is

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) = \tfrac{1}{2}\mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \tfrac{1}{2}\mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t) + \mathbf{p}^T(t)\mathbf{A}(t)\mathbf{x}(t) + \mathbf{p}^T(t)\mathbf{B}(t)\mathbf{u}(t), \quad (5.2\text{-}3)$$

and necessary conditions for optimality are

$$\dot{\mathbf{x}}^*(t) = \mathbf{A}(t)\mathbf{x}^*(t) + \mathbf{B}(t)\mathbf{u}^*(t) \quad (5.2\text{-}4)$$

$$\dot{\mathbf{p}}^*(t) = -\frac{\partial\mathscr{H}}{\partial\mathbf{x}} = -\mathbf{Q}(t)\mathbf{x}^*(t) - \mathbf{A}^T(t)\mathbf{p}^*(t) \quad (5.2\text{-}5)$$

† See references [K-5], [K-6], and [K-7].

$$\mathbf{0} = \frac{\partial \mathscr{H}}{\partial \mathbf{u}} = \mathbf{R}(t)\mathbf{u}^*(t) + \mathbf{B}^T(t)\mathbf{p}^*(t). \tag{5.2-6}$$

Equation (5.2-6) can be solved for $\mathbf{u}^*(t)$ to give

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{p}^*(t); \tag{5.2-7}$$

the existence of $\mathbf{R}^{-1}$ is assured, since $\mathbf{R}$ is a positive definite matrix. Substituting (5.2-7) into (5.2-4) yields

$$\dot{\mathbf{x}}^*(t) = \mathbf{A}(t)\mathbf{x}^*(t) - \mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{p}^*(t); \tag{5.2-8}$$

thus, we have the set of $2n$ linear homogeneous differential equations

$$\left[\begin{array}{c} \dot{\mathbf{x}}^*(t) \\ \hline \dot{\mathbf{p}}^*(t) \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}(t) & -\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t) \\ \hline -\mathbf{Q}(t) & -\mathbf{A}^T(t) \end{array}\right] \left[\begin{array}{c} \mathbf{x}^*(t) \\ \hline \mathbf{p}^*(t) \end{array}\right]. \tag{5.2-9}$$

The solution to these equations has the form

$$\left[\begin{array}{c} \mathbf{x}^*(t_f) \\ \hline \mathbf{p}^*(t_f) \end{array}\right] = \varphi(t_f, t) \left[\begin{array}{c} \mathbf{x}^*(t) \\ \hline \mathbf{p}^*(t) \end{array}\right], \tag{5.2-10}$$

where φ is the transition matrix of the system (5.2-9). Partitioning the transition matrix, we have

$$\left[\begin{array}{c} \mathbf{x}^*(t_f) \\ \hline \mathbf{p}^*(t_f) \end{array}\right] = \left[\begin{array}{c|c} \varphi_{11}(t_f, t) & \varphi_{12}(t_f, t) \\ \hline \varphi_{21}(t_f, t) & \varphi_{22}(t_f, t) \end{array}\right] \left[\begin{array}{c} \mathbf{x}^*(t) \\ \hline \mathbf{p}^*(t) \end{array}\right], \tag{5.2-10a}$$

where φ_{11}, φ_{12}, φ_{21}, and φ_{22} are $n \times n$ matrices.

From the boundary-condition equations—entry 2 of Table 5-1—we find that

$$\mathbf{p}^*(t_f) = \mathbf{H}\mathbf{x}^*(t_f). \tag{5.2-11}$$

Substituting this for $\mathbf{p}^*(t_f)$ in (5.2-10a) gives

$$\begin{aligned} \mathbf{x}^*(t_f) &= \varphi_{11}(t_f, t)\mathbf{x}^*(t) + \varphi_{12}(t_f, t)\mathbf{p}^*(t) \\ \mathbf{H}\mathbf{x}^*(t_f) &= \varphi_{21}(t_f, t)\mathbf{x}^*(t) + \varphi_{22}(t_f, t)\mathbf{p}^*(t). \end{aligned} \tag{5.2-12}$$

Substituting the upper equation into the lower, we obtain

$$\begin{aligned} \mathbf{H}\varphi_{11}(t_f, t)\mathbf{x}^*(t) + \mathbf{H}\varphi_{12}(t_f, t)\mathbf{p}^*(t) &= \varphi_{21}(t_f, t)\mathbf{x}^*(t) \\ &+ \varphi_{22}(t_f, t)\mathbf{p}^*(t), \end{aligned} \tag{5.2-13}$$

which, when solved for $\mathbf{p}^*(t)$, yields

$$\mathbf{p}^*(t) = [\boldsymbol{\varphi}_{22}(t_f, t) - \mathbf{H}\boldsymbol{\varphi}_{12}(t_f, t)]^{-1}[\mathbf{H}\boldsymbol{\varphi}_{11}(t_f, t) - \boldsymbol{\varphi}_{21}(t_f, t)]\mathbf{x}^*(t). \quad (5.2\text{-}14)$$

Kalman [K-7] has shown that the required inverse exists for all $t \in [t_0, t_f]$. Equation (5.2-14) can also be written as

$$\mathbf{p}^*(t) \triangleq \mathbf{K}(t)\mathbf{x}^*(t), \quad (5.2\text{-}15)$$

which means that $\mathbf{p}^*(t)$ is a linear function of the states of the system; $\mathbf{K}$ is an $n \times n$ matrix. Actually, $\mathbf{K}$ depends on t_f also, but t_f is specified.

Substituting in (5.2-7), we obtain

$$\begin{aligned}\mathbf{u}^*(t) &= -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t)\mathbf{x}(t) \\ &\triangleq \mathbf{F}(t)\mathbf{x}(t),\dagger\end{aligned} \quad (5.2\text{-}16)$$

which indicates that the optimal control law is a linear, albeit time-varying, combination of the system states. Notice that even if the plant is fixed, the feedback gain matrix $\mathbf{F}$ is time-varying.‡ In addition, measurements of all of the state variables must be available to implement the optimal control law. Figure 5-6 shows the plant and its optimal controller.

To determine the feedback gain matrix $\mathbf{F}$, we need the transition matrix for the system given in (5.2-9). If all of the matrices involved $(\mathbf{A}, \mathbf{B}, \mathbf{R}, \mathbf{Q})$ are time-invariant, the required transition matrix can be found by evaluating the inverse Laplace transform of the matrix

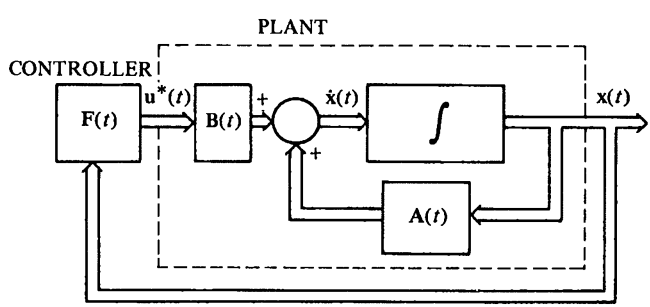


Figure 5-6 Plant and optimal feedback controller for linear regulator problems

† Here we drop the * notation because the optimal control law applies for all $\mathbf{x}(t)$.

‡ In certain cases it may be possible to implement a nonlinear, but time-invariant, optimal control law—see [J-1].

$$\left\{ s\mathbf{I} - \left[\begin{array}{c|c} \mathbf{A} & -\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T \\ \hline -\mathbf{Q} & -\mathbf{A}^T \end{array}\right] \right\}^{-1},$$

and substituting $(t_f - t)$ for t. Unfortunately, when the order of the system is large this becomes a tedious and time-consuming task. If any of the matrices in (5.2-9) is time-varying, we must generally resort to a numerical procedure for evaluating $\varphi(t_f, t)$.

There is an alternative approach, however; it can be shown (see Problem 5-9) that the matrix $\mathbf{K}$ satisfies the matrix differential equation

$$\dot{\mathbf{K}}(t) = -\mathbf{K}(t)\mathbf{A}(t) - \mathbf{A}^T(t)\mathbf{K}(t) - \mathbf{Q}(t) + \mathbf{K}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t), \tag{5.2-17}$$

with the boundary condition $\mathbf{K}(t_f) = \mathbf{H}$.

This matrix differential equation is of the Riccati type; in fact, we shall call (5.2-17) the *Riccati equation*.† Since $\mathbf{K}$ is an $n \times n$ matrix, Eq. (5.2-17) is a system of n^2 first-order differential equations. Actually, it can be shown (see Problem 5-9), that $\mathbf{K}$ is symmetric; hence, not n^2, but $n(n+1)/2$ first-order differential equations must be solved. These equations can be integrated numerically by using a digital computer. The integration is started at $t = t_f$ and proceeds backward in time to $t = t_0$; $\mathbf{K}(t)$ is stored, and the feedback gain matrix is determined from Eq. (5.2-16).

Let us illustrate these concepts with the following examples.

Example 5.2-1. Find the optimal control law for the system

$$\dot{x}(t) = ax(t) + u(t) \tag{5.2-18}$$

to minimize the performance measure

$$J(u) = \tfrac{1}{2}Hx^2(T) + \int_0^T \tfrac{1}{4}u^2(t)\,dt. \tag{5.2-19}$$

The admissible state and control values are unconstrained, the final time T is specified, $H > 0$, and $x(T)$ is free.

Equation (5.2-9) gives

$$\begin{bmatrix} \dot{x}^*(t) \\ \dot{p}^*(t) \end{bmatrix} = \begin{bmatrix} a & -2 \\ 0 & -a \end{bmatrix} \begin{bmatrix} x^*(t) \\ p^*(t) \end{bmatrix}, \tag{5.2-20}$$

† The Riccati equation is also derived in Section 3.12, where the Hamilton-Jacobi-Bellman equation is used.

which has the transition matrix

$$\varphi(t) = \left[\begin{array}{c|c} \epsilon^{at} & \frac{1}{a}\epsilon^{-at} - \frac{1}{a}\epsilon^{at} \\ \hline 0 & \epsilon^{-at} \end{array}\right];$$

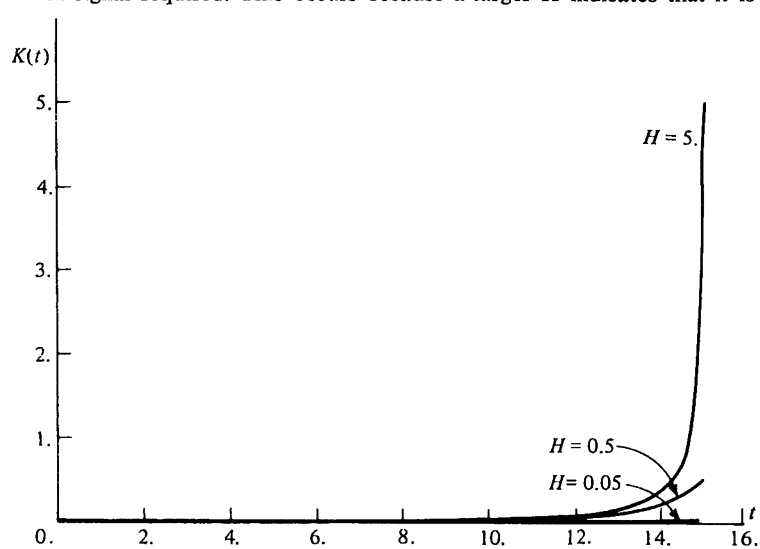
thus, from Eqs. (5.2-14) and (5.2-15) we have

$$K(t) = \left[\epsilon^{-a(T-t)} - \frac{H}{a}[\epsilon^{-a(T-t)} - \epsilon^{a(T-t)}]\right]^{-1}[H\epsilon^{a(T-t)}], \quad (5.2\text{-}21)$$

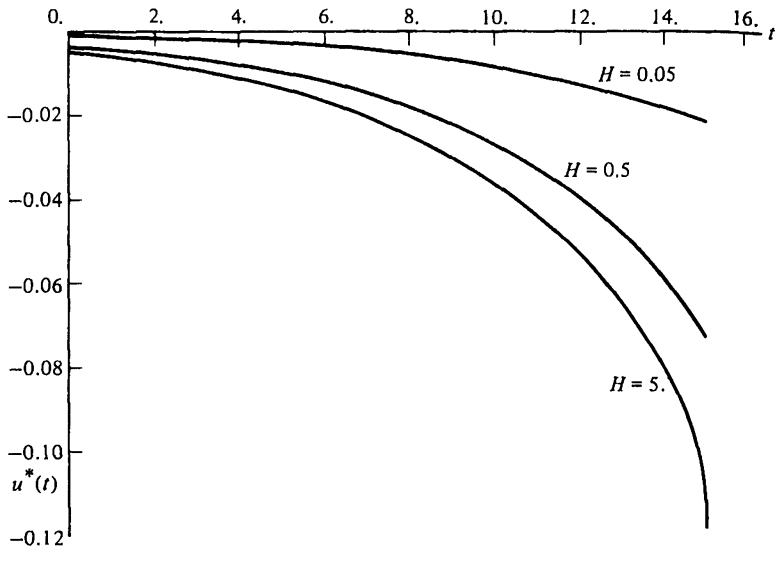
and from Eq. (5.2-16) the optimal control law is

$$u^*(t) = -2K(t)x(t) \quad (5.2\text{-}22)$$

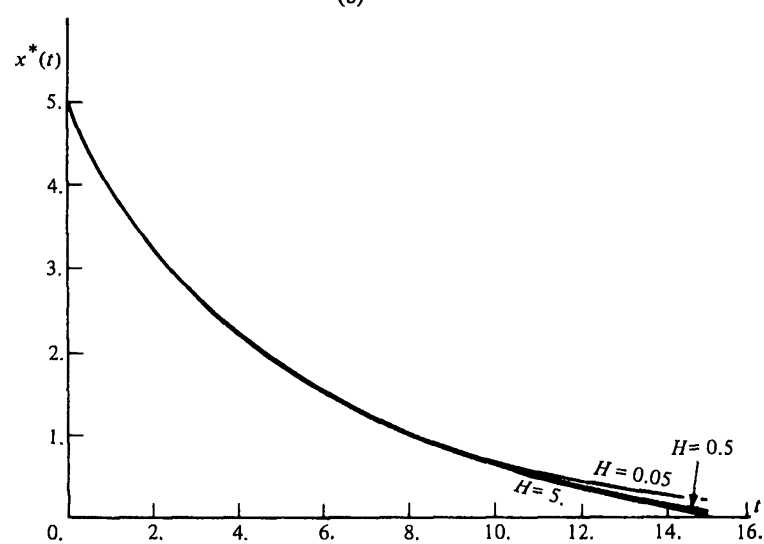
Figure 5-7(a) shows $K(t)$ as a function of time for $a = -0.2$ and $T = 15$, with $H = 5.0$, 0.5, and 0.05. The corresponding control histories and state trajectories for $x(0) = 5.0$ are shown in Fig. 5-7(b), (c). Notice that the state trajectories are almost identical and that the control signals are small in all three cases. These qualitative observations can be explained physically by noting that with $a = -0.2$ the plant is stable and tends toward zero—the desired state—even if no control is applied. Observe in Fig. 5-7(b) that the larger the value of H, the larger the control signal required. This occurs because a larger H indicates that it is



(a)



(b)

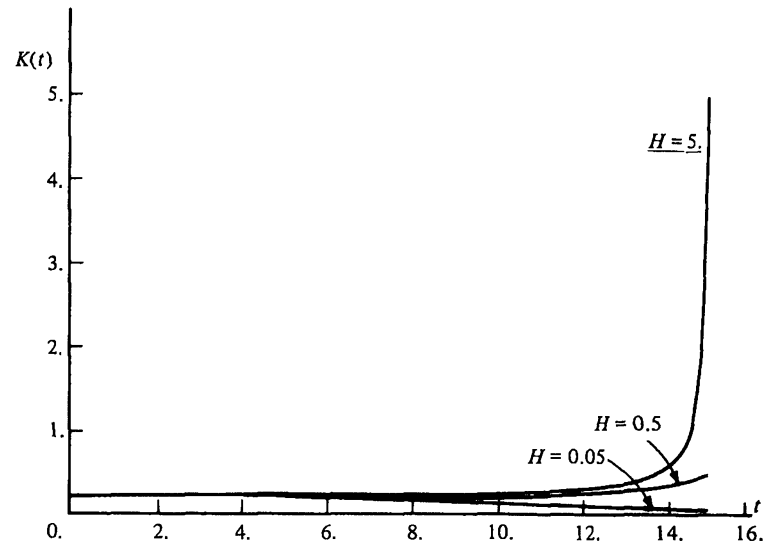


(c)

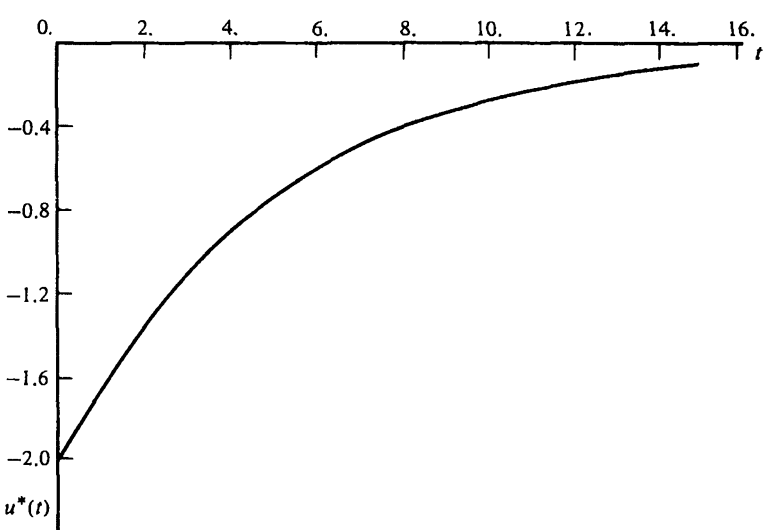
Figure 5-7 (a) Solution of the Riccati equation for $a = -0.2$, $H = 5, 0.5, 0.05$. (b) The optimal control histories for $a = -0.2$, $H = 5, 0.5, 0.05$. (c) The optimal trajectories for $a = -0.2$, $H = 5, 0.5,$ and 0.05.

desired to be closer to $x(15) = 0$ than with a smaller H—even if more control effort is required.

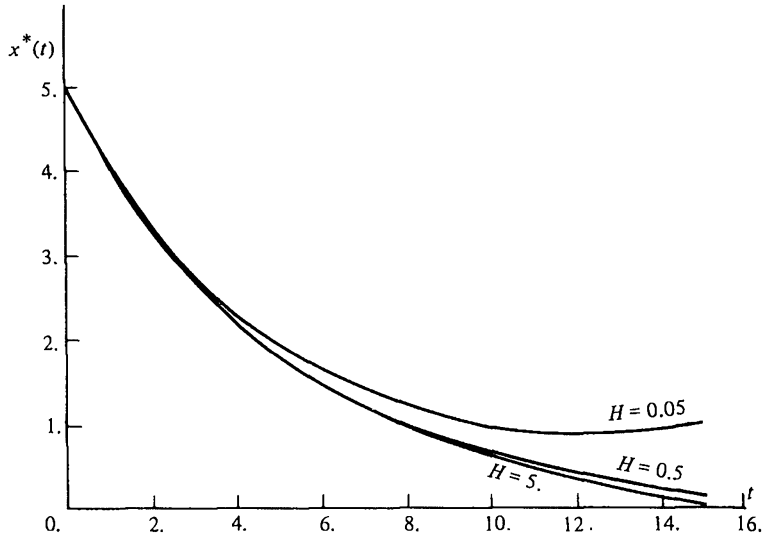
If $a = 0.2$ the results are as shown in Fig. 5-8. Notice that the control signals (which are essentially identical with one another) are much larger than when $a = -0.2$. This is expected, because the plant with $a = 0.2$ is unstable.



(a)



(b)



(c)

Figure 5-8 (a) Solution of the Riccati equation for $a = 0.2$, $H = 5, 0.5, 0.05$. (b) The optimal control histories for $a = 0.2$, $H = 5, 0.5, 0.05$. (c) The optimal trajectories for $a = 0.2$, $H = 5, 0.5, 0.05$.

Another point of interest is the period of time in the interval [0, 15] during which the control signals are largest in magnitude. For the stable plant ($a = -0.2$) the largest controls are applied as $t \longrightarrow 15$. This is the case because the controller "waits" for the system to approach zero on its own before applying control effort. On the other hand, if the controller were to wait for the unstable plant to move toward zero, the instability would cause the value of x to grow larger; hence, the largest control magnitudes are applied in the initial stages of the interval of operation.

Example 5.2-2. Consider the second-order system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= 2x_1(t) - x_2(t) + u(t), \end{aligned} \tag{5.2-23}$$

which is to be controlled to minimize

$$J(u) = \int_0^T [x_1^2(t) + \tfrac{1}{2}x_2^2(t) + \tfrac{1}{4}u^2(t)]\, dt. \tag{5.2-24}$$

Find the optimal control law.

By expanding the Riccati equation with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad R = \tfrac{1}{2},$$

we obtain

$$\begin{aligned} \dot{k}_{11}(t) &= 2[k_{12}^2(t) - 2k_{12}(t) - 1] \\ \dot{k}_{12}(t) &= 2k_{12}(t)k_{22}(t) - k_{11}(t) + k_{12}(t) - 2k_{22}(t) \\ \dot{k}_{22}(t) &= 2k_{22}^2(t) - 2k_{12}(t) + 2k_{22}(t) - 1. \end{aligned} \tag{5.2-25}$$

In arriving at (5.2-25) the symmetry of $\mathbf{K}$ has been used. The boundary conditions are $k_{11}(T) = k_{12}(T) = k_{22}(T) = 0$, and the optimal control law is

$$u^*(t) = -2[k_{12}(t) \quad k_{22}(t)]\mathbf{x}(t). \tag{5.2-26}$$

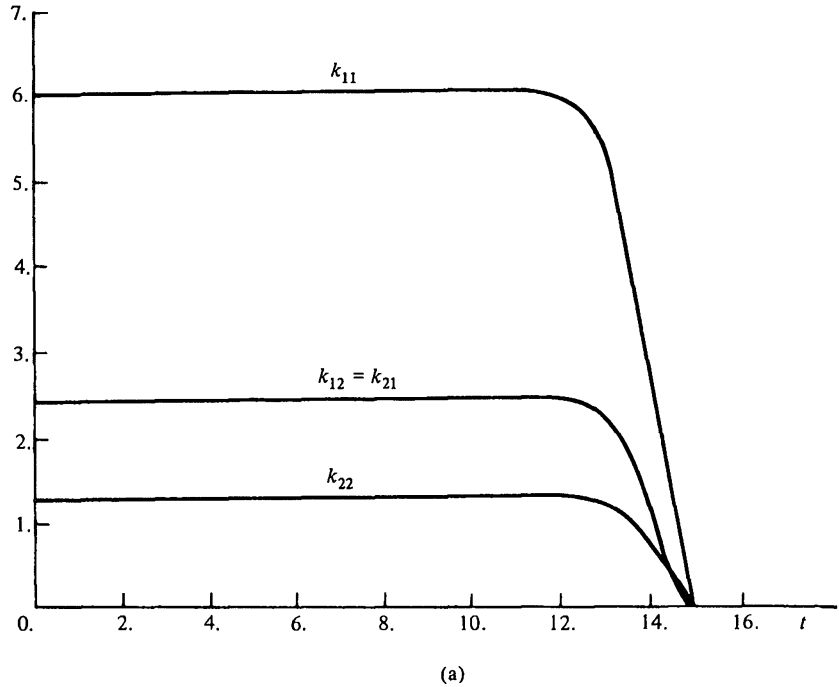
The solution of the Riccati equation and the optimal control and its trajectory are shown in Fig. 5-9 for $\mathbf{x}(0) = [-4 \quad 4]^T$.

The situation wherein the process is to be controlled for an interval of infinite duration merits special attention. Kalman [K-7] has shown that if (1) the plant is completely controllable, (2) $\mathbf{H} = \mathbf{0}$, and (3) $\mathbf{A}$, $\mathbf{B}$, $\mathbf{R}$, and $\mathbf{Q}$ are constant matrices, $\mathbf{K}(t) \longrightarrow \mathbf{K}$ (a constant matrix) as $t_f \longrightarrow \infty$. The engineering implications of this result are very important. If the above hypotheses are satisfied, then the optimal control law for an infinite-duration process is stationary. This means that the implementation of the optimal controller is as shown in Fig. 5-6, *except* that $\mathbf{F}(t)$ is constant; thus, the controller consists of m fixed summing amplifiers, each having n inputs. From a practical viewpoint, it may be feasible to use the fixed control law even for processes of finite duration. For instance, in Example 5.2-2 k_{11}, k_{12}, and k_{22} are essentially constants for $0 \leq t \leq 12$. Looking at the state trajectory in Fig. 5-9(b), we see that the states have both essentially reached zero when $t = 5$. This means that perhaps the constant values $k_{11} = 6.03$, $k_{12} = 2.41$, $k_{22} = 1.28$ can be used without significant performance degradation—the designer should compare system performance using the steady-state gains with performance using the time-varying optimal gains to decide which should be implemented.

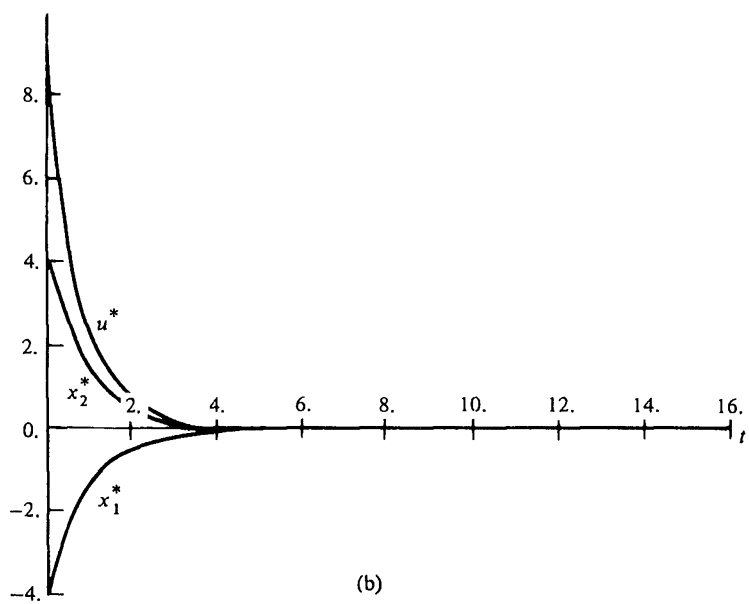
To determine the $\mathbf{K}$ matrix for an infinite-time process, we either integrate the Riccati equation backward in time until a steady-state solution is obtained [see Fig.5-9(a)] or solve the nonlinear algebraic equations

$$\mathbf{0} = -\mathbf{K}\mathbf{A} - \mathbf{A}^T\mathbf{K} - \mathbf{Q} + \mathbf{K}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{K}, \tag{5.2-27}$$

obtained by setting $\dot{\mathbf{K}}(t) = \mathbf{0}$ in Eq. (5.2-17).



(a)



(b)

Figure 5-9 (a) The solution of the Riccati equation. (b) The optimal control and its trajectory

Linear Tracking Problems

Next, let us generalize the results obtained for the linear regulator problem to the tracking problem; that is, the desired value of the state vector is not the origin.

The state equations are

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \tag{5.2-28}$$

and the performance measure to be minimized is

$$\begin{aligned} J &= \tfrac{1}{2}[\mathbf{x}(t_f) - \mathbf{r}(t_f)]^T\mathbf{H}[\mathbf{x}(t_f) - \mathbf{r}(t_f)] + \tfrac{1}{2}\int_{t_0}^{t_f} \{[\mathbf{x}(t) - \mathbf{r}(t)]^T\mathbf{Q}(t)[\mathbf{x}(t) - \mathbf{r}(t)] \\ &\quad + \mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t)\}\, dt \\ &\triangleq \tfrac{1}{2}\|\mathbf{x}(t_f) - \mathbf{r}(t_f)\|_{\mathbf{H}}^2 + \tfrac{1}{2}\int_{t_0}^{t_f} \{\|\mathbf{x}(t) - \mathbf{r}(t)\|_{\mathbf{Q}(t)}^2 + \|\mathbf{u}(t)\|_{\mathbf{R}(t)}^2\}\, dt, \end{aligned} \tag{5.2-29}$$

where $\mathbf{r}(t)$ is the desired or reference value of the state vector. The final time t_f is fixed, $\mathbf{x}(t_f)$ is free, and the states and controls are not bounded. $\mathbf{H}$ and $\mathbf{Q}$ are real symmetric positive semi-definite matrices, and $\mathbf{R}$ is real symmetric and positive definite.

The Hamiltonian is given by

$$\begin{aligned} \mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) &= \tfrac{1}{2}\|\mathbf{x}(t) - \mathbf{r}(t)\|_{\mathbf{Q}(t)}^2 + \tfrac{1}{2}\|\mathbf{u}(t)\|_{\mathbf{R}(t)}^2 \\ &\quad + \mathbf{p}^T(t)\mathbf{A}(t)\mathbf{x}(t) + \mathbf{p}^T(t)\mathbf{B}(t)\mathbf{u}(t). \end{aligned} \tag{5.2-30}$$

The costate equations are

$$\dot{\mathbf{p}}^*(t) = -\frac{\partial \mathscr{H}}{\partial \mathbf{x}} = -\mathbf{Q}(t)\mathbf{x}^*(t) - \mathbf{A}^T(t)\mathbf{p}^*(t) + \mathbf{Q}(t)\mathbf{r}(t), \tag{5.2-31}$$

and the algebraic relations that must be satisfied are given by

$$\mathbf{0} = \frac{\partial \mathscr{H}}{\partial \mathbf{u}} = \mathbf{R}(t)\mathbf{u}^*(t) + \mathbf{B}^T(t)\mathbf{p}^*(t); \tag{5.2-32}$$

therefore,

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{p}^*(t). \tag{5.2-33}$$

Substituting (5.2-33) in the state equations yields the state and costate equations

$$\left[\begin{array}{c} \dot{\mathbf{x}}^*(t) \\ \hline \dot{\mathbf{p}}^*(t) \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}(t) & -\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t) \\ \hline -\mathbf{Q}(t) & -\mathbf{A}^T(t) \end{array}\right] \left[\begin{array}{c} \mathbf{x}^*(t) \\ \hline \mathbf{p}^*(t) \end{array}\right] + \left[\begin{array}{c} \mathbf{0} \\ \hline \mathbf{Q}(t)\mathbf{r}(t) \end{array}\right] \tag{5.2-34}$$

Notice that the term $\mathbf{Q}(t)\overline{\mathbf{r}}(t)$ is a forcing function; these differential equations are linear and time-varying, but not homogeneous. The solution of (5.2-34) is

$$\begin{bmatrix} \mathbf{x}^*(t_f) \\ \hline \mathbf{p}^*(t_f) \end{bmatrix} = \varphi(t_f, t) \begin{bmatrix} \mathbf{x}^*(t) \\ \hline \mathbf{p}^*(t) \end{bmatrix} + \int_t^{t_f} \varphi(t_f, \tau) \begin{bmatrix} \mathbf{0} \\ \hline \mathbf{Q}(\tau)\mathbf{r}(\tau) \end{bmatrix} d\tau, \quad (5.2\text{-}35)$$

where φ is the transition matrix of the system (5.2-34). If φ is partitioned, and the integral replaced by the $2n \times 1$ vector

$$\begin{bmatrix} \mathbf{f}_1(t) \\ \hline \mathbf{f}_2(t) \end{bmatrix},$$

these equations can be written

$$\mathbf{x}^*(t_f) = \varphi_{11}(t_f, t)\mathbf{x}^*(t) + \varphi_{12}(t_f, t)\mathbf{p}^*(t) + \mathbf{f}_1(t) \quad (5.2\text{-}36a)$$

$$\mathbf{p}^*(t_f) = \varphi_{21}(t_f, t)\mathbf{x}^*(t) + \varphi_{22}(t_f, t)\mathbf{p}^*(t) + \mathbf{f}_2(t). \quad (5.2\text{-}36b)$$

The boundary conditions are

$$\mathbf{p}^*(t_f) = \mathbf{H}\mathbf{x}^*(t_f) - \mathbf{H}\mathbf{r}(t_f). \quad (5.2\text{-}37)$$

Replacing $\mathbf{p}^*(t_f)$ in (5.2-36b) by the right-hand side of (5.2-37) and then substituting $\mathbf{x}^*(t_f)$ from Eq. (5.2-36a) into (5.2-36b), we obtain

$$\mathbf{H}[\varphi_{11}(t_f, t)\mathbf{x}^*(t) + \varphi_{12}(t_f, t)\mathbf{p}^*(t) + \mathbf{f}_1(t)] - \mathbf{H}\mathbf{r}(t_f) = \varphi_{21}(t_f, t)\mathbf{x}^*(t) + \varphi_{22}(t_f, t)\mathbf{p}^*(t) + \mathbf{f}_2(t). \quad (5.2\text{-}38)$$

Solving for $\mathbf{p}^*(t)$ yields

$$\begin{aligned} \mathbf{p}^*(t) &= [\varphi_{22}(t_f, t) - \mathbf{H}\varphi_{12}(t_f, t)]^{-1} [\mathbf{H}\varphi_{11}(t_f, t) - \varphi_{21}(t_f, t)]\mathbf{x}^*(t) \\ &\quad + [\varphi_{22}(t_f, t) - \mathbf{H}\varphi_{12}(t_f, t)]^{-1} [\mathbf{H}\mathbf{f}_1(t) - \mathbf{H}\mathbf{r}(t_f) - \mathbf{f}_2(t)] \\ &\triangleq \mathbf{K}(t)\mathbf{x}^*(t) + \mathbf{s}(t). \end{aligned} \quad (5.2\text{-}39)$$

The definitions of $\mathbf{K}(t)$ and $\mathbf{s}(t)$ are apparent by inspection of Eq. (5.2-39); therefore, the optimal control law is

$$\begin{aligned} \mathbf{u}^*(t) &= -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t)\mathbf{x}(t) - \mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{s}(t) \\ &\triangleq \mathbf{F}(t)\mathbf{x}(t) + \mathbf{v}(t), \end{aligned} \quad (5.2\text{-}40)$$

where $\mathbf{F}(t)$ is the feedback gain matrix and $\mathbf{v}(t)$ is the command signal.† Notice that $\mathbf{v}(t)$ depends on the system parameters and on the reference signal $\mathbf{r}(t)$. In fact, $\mathbf{v}(t)$ depends on future values of the reference signal, so we might say that the optimal control has an anticipatory quality. This is reinforced by physical reasoning, which tells us that we must determine our present strategy on the basis of where we are now *and* where we intend to go. (Actually, this same sort of situation was present, though in a more subtle way, in regulator problems, where we utilized our desire to be at the origin.) A diagram of the plant and controller is shown in Fig. 5-10. Notice that, as in the regulator problem, we must be able to measure all of the states in order to synthesize the optimal control law.

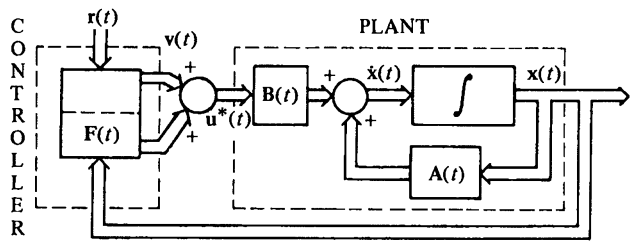


Figure 5-10 Plant and optimal feedback controller for linear tracking problems

Again we are confronted with the need to determine the transition matrix, but, as before, there is an easier computational route to travel. We begin with the equation

$$\mathbf{p}^*(t) = \mathbf{K}(t)\mathbf{x}^*(t) + \mathbf{s}(t). \tag{5.2-41}$$

Differentiating both sides with respect to t, we obtain

$$\dot{\mathbf{p}}^*(t) = \dot{\mathbf{K}}(t)\mathbf{x}^*(t) + \mathbf{K}(t)\dot{\mathbf{x}}^*(t) + \dot{\mathbf{s}}(t). \tag{5.2-42}$$

Substituting from (5.2-34) for $\dot{\mathbf{p}}^*(t)$ and $\dot{\mathbf{x}}^*(t)$, and using (5.2-41) to eliminate $\mathbf{p}^*(t)$, we obtain

$$\begin{aligned}[\dot{\mathbf{K}}(t) + \mathbf{Q}(t) + \mathbf{K}(t)\mathbf{A}(t) + \mathbf{A}^T(t)\mathbf{K}(t) - \mathbf{K}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t)]\mathbf{x}^*(t) \\ + [\dot{\mathbf{s}}(t) + \mathbf{A}^T(t)\mathbf{s}(t) - \mathbf{K}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{s}(t) - \mathbf{Q}(t)\mathbf{r}(t)] = \mathbf{0}.\end{aligned} \tag{5.2-43}$$

† Strictly speaking, we have not shown that this extremal control does minimize J. It turns out, however, that this extremal control is indeed the optimal control.

Because this must be satisfied for all $\mathbf{x}^*(t)$ and $\mathbf{r}(t)$, we conclude that

$$\dot{\mathbf{K}}(t) = -\mathbf{K}(t)\mathbf{A}(t) - \mathbf{A}^T(t)\mathbf{K}(t) - \mathbf{Q}(t) + \mathbf{K}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t) \tag{5.2-44}$$

and

$$\dot{\mathbf{s}}(t) = -[\mathbf{A}^T(t) - \mathbf{K}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)]\,\mathbf{s}(t) + \mathbf{Q}(t)\mathbf{r}(t). \tag{5.2-45}$$

Since $\mathbf{K}$ is symmetric and $\mathbf{s}$ is an $n \times 1$ vector, (5.2-44) and (5.2-45) are a set of $[n[n+1]/2] + n$ first-order differential equations. Notice that (5.2-44) is the same Riccati equation that we obtained for linear regulator problems. To obtain the boundary conditions we have, from (5.2-37) and (5.2-39),

$$\begin{aligned}\mathbf{p}^*(t_f) &= \mathbf{H}\mathbf{x}^*(t_f) - \mathbf{H}\mathbf{r}(t_f) \\ &= \mathbf{K}(t_f)\mathbf{x}^*(t_f) + \mathbf{s}(t_f).\end{aligned} \tag{5.2-46}$$

Since these equations must be satisfied for all $\mathbf{x}^*(t_f)$ and $\mathbf{r}(t_f)$, the boundary conditions are

$$\mathbf{K}(t_f) = \mathbf{H} \tag{5.2-47}$$

and

$$\mathbf{s}(t_f) = -\mathbf{H}\mathbf{r}(t_f). \tag{5.2-48}$$

To determine $\mathbf{F}(t)$ and $\mathbf{v}(t)$, we then integrate (5.2-44) and (5.2-45) *from* t_f *to* t_0 using the boundary conditions (5.2-47) and (5.2-48), and store the values for $\mathbf{K}(t)$ and $\mathbf{s}(t)$. $\mathbf{F}(t)$ and $\mathbf{v}(t)$ can then be determined by using (5.2-40). The procedure is illustrated by the following examples.

Example 5.2-3. The system

$$\begin{aligned}\dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= 2x_1(t) - x_2(t) + u(t)\end{aligned} \tag{5.2-49}$$

is to be controlled to minimize the performance measure

$$J(u) = [x_1(T) - 1]^2 + \int_0^T \{[x_1(t) - 1]^2 + 0.0025u^2(t)\}\,dt. \tag{5.2-50}$$

The final time T is specified, $\mathbf{x}(T)$ is free, and the admissible states and controls are not bounded. The optimal control law is to be found.

The performance measure indicates that the state x_1 is to be maintained close to 1.0 without excessive expenditure of control effort. In the nomenclature of linear tracking problems, we have

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \mathbf{Q} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{H},$$

$$R = 0.005 \quad \text{and} \quad \mathbf{r}(t) = \begin{bmatrix} 1.0 \\ 0 \end{bmatrix}. \dagger$$

The Riccati equation and the differential equations for s are found from Eqs. (5.2-44) and (5.2-45) with the result

$$\begin{aligned} \dot{k}_{11}(t) &= 2[100k_{12}^2(t) - 2k_{12}(t) - 1] \\ \dot{k}_{12}(t) &= 200k_{12}(t)k_{22}(t) - k_{11}(t) + k_{12}(t) - 2k_{22}(t) \\ \dot{k}_{22}(t) &= 200k_{22}^2(t) - 2k_{12}(t) + 2k_{22}(t) \end{aligned} \tag{5.2-51}$$

$$\begin{aligned} \dot{s}_1(t) &= 2[100k_{12}(t) - 1]s_2(t) + 2 \\ \dot{s}_2(t) &= -s_1(t) + [1 + 200k_{22}(t)]s_2(t), \end{aligned} \tag{5.2-52}$$

and, from Eqs. (5.2-47) and (5.2-48) the boundary conditions are

$$\mathbf{K}(T) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{s}(T) = \begin{bmatrix} -2 \\ 0 \end{bmatrix}.$$

The optimal control law, obtained from (5.2-40), is

$$u^*(t) = -200[k_{12}(t)x_1(t) + k_{22}(t)x_2(t) + s_2(t)]. \tag{5.2-53}$$

Figure 5-11(a) shows the optimal control and its trajectory for $T = 15$, and $\mathbf{x}(0) = \mathbf{0}$. The "tail" on the x_1^* curve as $t \longrightarrow t_f$ results because the controller anticipates that the final time is near and reduces the control to values near zero at the expense of deviations in x_1^*. When the control is made small, $x_1^*(t)$ begins to increase; this occurs because the plant (5.2-49) is unstable. The solutions of the Riccati equation and of (5.2-52) are shown in Fig. 5-11(b), (c).

Example 5.2-4. The plant to be controlled is the same as in Example 5.2-3, but the performance measure is

$$J(u) = \int_0^T \{[x_1(t) - 0.2t]^2 + 0.025u^2(t)\}\, dt. \tag{5.2-54}$$

T is specified, $\mathbf{x}(T)$ is free, and the admissible controls are not bounded. The optimal control law is to be determined.

In this problem the objective is to maintain the state x_1 close to the ramp function $r_1(t) = 0.2t$, without excessive expenditure of control effort. By substituting

† For the matrices **H** and **Q** given in this example, $r_2(t)$ does not affect the solution and hence can be selected arbitrarily.

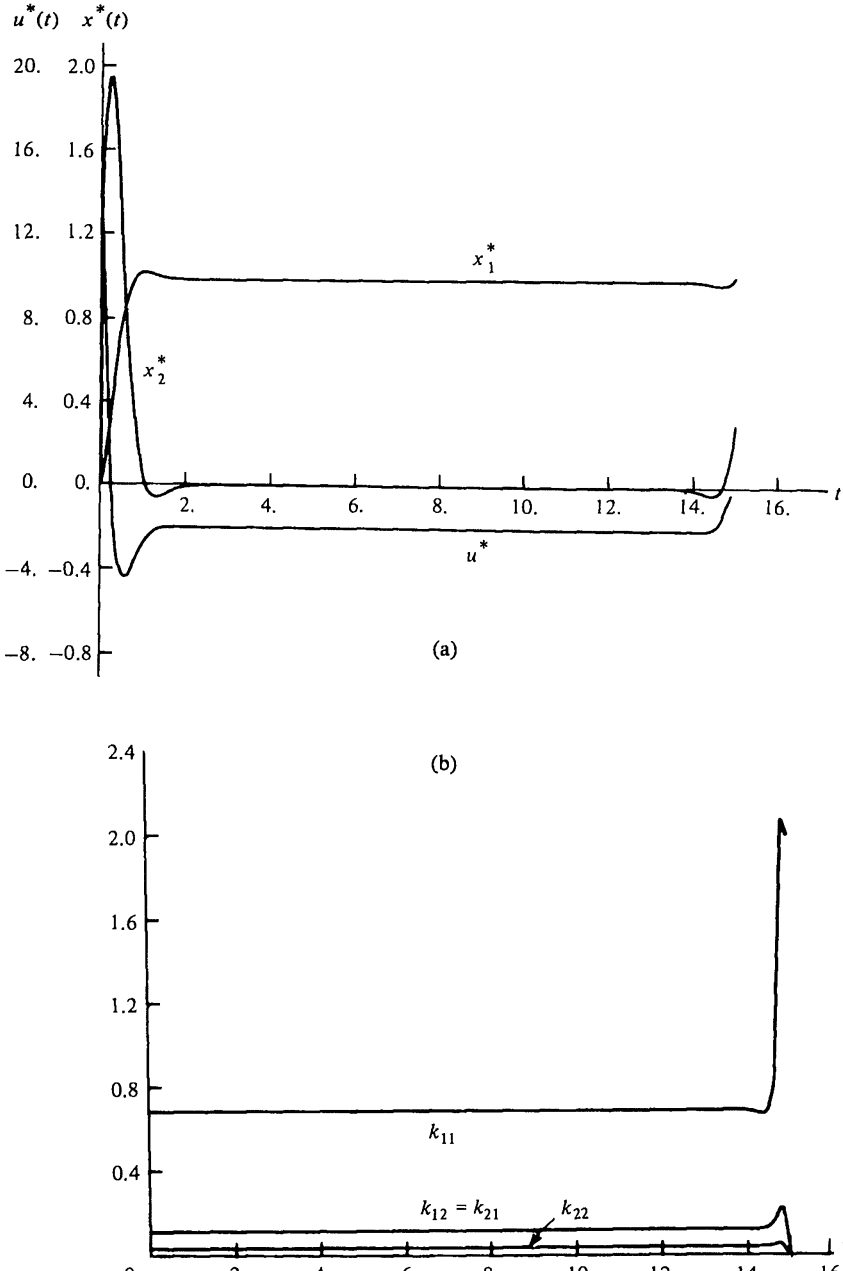


Figure 5-11 (a) The optimal control and trajectory for a linear tracking problem: $r_1(t) = 1.0$, $\mathbf{x}(0) = \mathbf{0}$. (b) Solution of the Riccati equation for Example 5.2-3. (c) s_1 and s_2 for Example 5.2-3

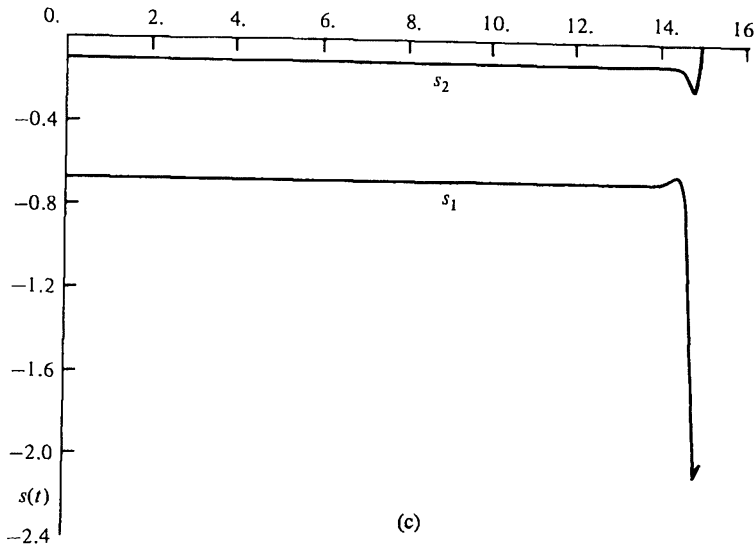


Figure 5-11 cont.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$R = 0.05, \quad \text{and} \quad \mathbf{r}(t) = \begin{bmatrix} 0.2t \\ 0 \end{bmatrix}$$

into (5.2-44) and (5.2-45), we obtain the differential equations

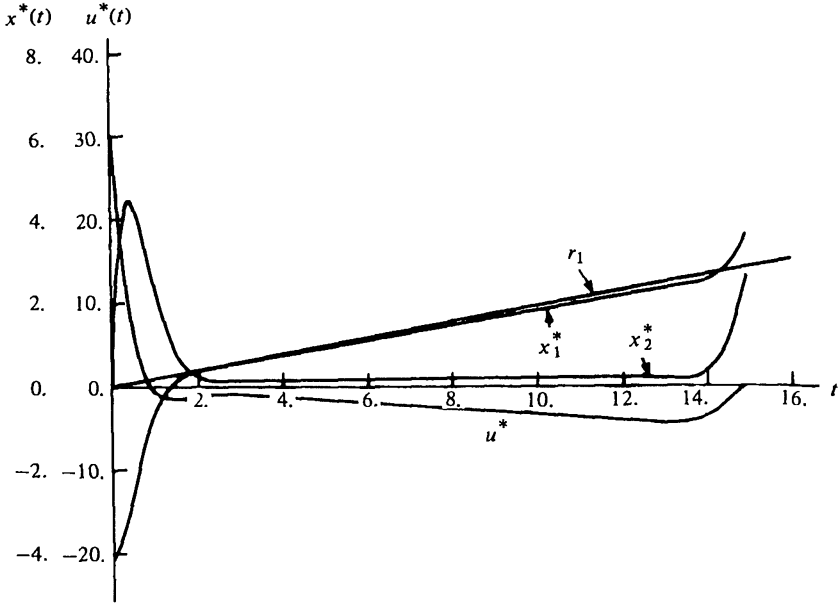
$$\begin{aligned} \dot{k}_{11}(t) &= 20k_{12}^2(t) - 4k_{12}(t) - 2 \\ \dot{k}_{12}(t) &= 20k_{12}(t)k_{22}(t) - k_{11}(t) + k_{12}(t) - 2k_{22}(t) \\ \dot{k}_{22}(t) &= 20k_{22}^2(t) - 2k_{12}(t) + 2k_{22}(t) \end{aligned} \tag{5.2-55}$$

$$\begin{aligned} \dot{s}_1(t) &= 2[10k_{12}(t) - 1]s_2(t) + 0.4t \\ \dot{s}_2(t) &= -s_1(t) + [20k_{22}(t) + 1]s_2(t). \end{aligned} \tag{5.2-56}$$

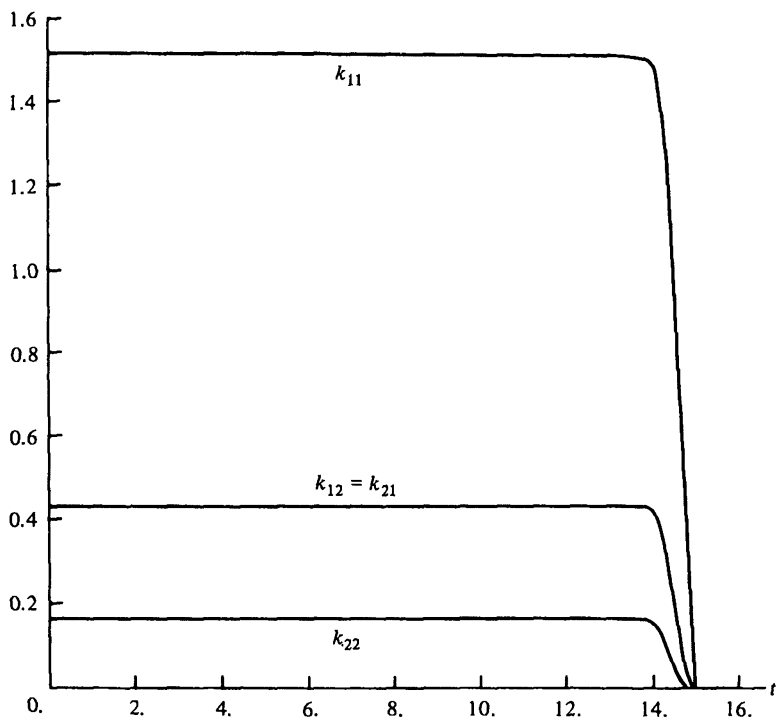
The boundary conditions for these five differential equations are $\mathbf{K}(T) = \mathbf{0}$, $\mathbf{s}(T) = \mathbf{0}$. Figures 5-12(b) and (c) show the solution of Eqs. (5.2-55) and (5.2-56) for $T = 15$. The optimal control law, obtained from Eq. (5.2-40), is

$$u^*(t) = -20[k_{12}(t)x_1(t) + k_{22}(t)x_2(t) + s_2(t)]. \tag{5.2-57}$$

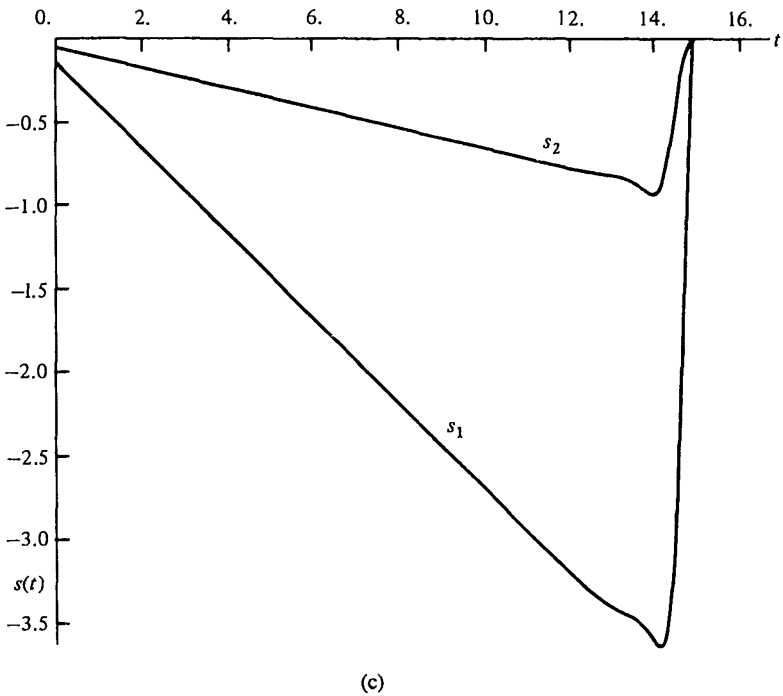
The optimal control and its trajectory for $\mathbf{x}(0) = [-4 \quad 0]^T$ are shown in Fig. 5-12(a). There is an initial transient period that is over at approximately $t = 2$. Thereafter, the difference between x_1^* and r_1 is small,


$x^*(t)$
$u^*(t)$
8.
6.
4.
2.
0.
−2.
−4.
40.
30.
20.
10.
0.
−10.
−20.
2.
4.
6.
8.
10.
12.
14.
16.
t
r_1
x_1^*
x_2^*
u^*

(a)


1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.
2.
4.
6.
8.
10.
12.
14.
16.
t
k_{11}
$k_{12} = k_{21}$
k_{22}

(b)



(c)

Figure 5-12 (a) The optimal control and trajectory for a linear tracking problem: $r_1(t) = 0.2t$, $\mathbf{x}(0) = [-4\ 0]^T$. (b) Solution of the Riccati equation for Example 5.2-4. (c) s_1 and s_2 for Example 5.2-4

although the deviation does grow larger with increasing time. This is attributed to the penalty in the performance measure on control-effort expenditure; as time increases, the magnitude of the control signal required for tracking grows larger, so the contribution of control effort to the performance measure becomes more significant. The "tail" present as $t \longrightarrow 15$ occurs because the control law anticipates the end of the control interval and, as a result, conserves control effort, allowing x_1^* to deviate from its desired values.

5.3 PONTRYAGIN'S MINIMUM PRINCIPLE AND STATE INEQUALITY CONSTRAINTS

So far, we have assumed that the admissible controls and states are not constrained by any boundaries; however, in realistic systems such constraints do commonly occur. Physically realizable controls generally have magnitude

limitations: the thrust of a rocket engine cannot exceed a certain value; motors, which provide torque, saturate; attitude control mass expulsion systems are capable of providing a limited torque. State constraints often arise because of safety, or structural restrictions: the current in an electric motor cannot exceed a certain value without damaging the windings; the turning radius of a maneuvering aircraft cannot be less than a specified minimum value; a spacecraft reentering the earth's atmosphere must satisfy certain attitude and velocity constraints to avoid burning up.

Let us first consider the effect of control constraints on the fundamental theorem derived in Section 4.1, and then show how the necessary conditions are modified.† This generalization of the fundamental theorem leads to Pontryagin's minimum principle.‡

Pontryagin's Minimum Principle

By definition, the control $\mathbf{u}^*$ causes the functional J to have a relative minimum if

$$J(\mathbf{u}) - J(\mathbf{u}^*) = \Delta J \geq 0 \tag{5.3-1}$$

for all admissible controls sufficiently close to $\mathbf{u}^*$. If we let $\mathbf{u} = \mathbf{u}^* + \delta\mathbf{u}$, the increment in J can be expressed as

$$\Delta J(\mathbf{u}^*, \delta\mathbf{u}) = \delta J(\mathbf{u}^*, \delta\mathbf{u}) + \text{higher-order terms}; \tag{5.3-2}$$

δJ is linear in $\delta\mathbf{u}$ and the higher-order terms approach zero as the norm of $\delta\mathbf{u}$ approaches zero. If we were to re-prove the fundamental theorem *for unbounded controls* using control system notation, the reasoning would be exactly as given in Section 4.1. That is, if the control were unbounded, we could use the linearity of δJ with respect to $\delta\mathbf{u}$, and the fact that $\delta\mathbf{u}$ can vary arbitrarily to show that a necessary condition for $\mathbf{u}^*$ to be an extremal control is that the variation $\delta J(\mathbf{u}^*, \delta\mathbf{u})$ must be zero for all admissible $\delta\mathbf{u}$ having a sufficiently small norm. Since we are no longer assuming that the admissible controls are not bounded, $\delta\mathbf{u}$ is arbitrary only if the extremal control is strictly within the boundary for all time in the interval $[t_0, t_f]$. In this case, the boundary has no effect on the problem solution. If, however, an extremal control lies on a boundary during at least one subinterval $[t_1, t_2]$ of the interval $[t_0, t_f]$, as shown in Fig. 5-13(a), then admissible control variations $\delta\hat{\mathbf{u}}$ exist whose negatives $(-\delta\hat{\mathbf{u}})$ are not admissible. One such control variation is shown in Fig. 5-13(b). If only these variations are considered, a necessary condition for $\mathbf{u}^*$ to minimize J is that $\delta J(\mathbf{u}^*, \delta\hat{\mathbf{u}}) \geq 0$. On the other hand, for variations

† The derivation given here is heuristic; for rigorous proofs see [P-1], [R-1], and [A-2].

‡ In Pontryagin's original work, [P-1], this result is referred to as the maximum principle because of a sign difference in the definition of the Hamiltonian.

$\delta\tilde{\mathbf{u}}$, which are nonzero only for t not in the interval $[t_1, t_2]$, as, for example, in Fig. 5-13(c), it is necessary that $\delta J(\mathbf{u}^*, \delta\tilde{\mathbf{u}}) = 0$; the reasoning used in proving the fundamental theorem applies. Considering all admissible variations with $\|\delta\mathbf{u}\|$ small enough so that the sign of ΔJ is determined by δJ, we see that a necessary condition for $\mathbf{u}^*$ to minimize J is

$$\delta J(\mathbf{u}^*, \delta\mathbf{u}) \geq 0. \tag{5.3-3}$$

It seems reasonable to ask if this result has an analog in calculus. To answer this question, refer to Fig. 4-4, where a function f, defined on a *closed*

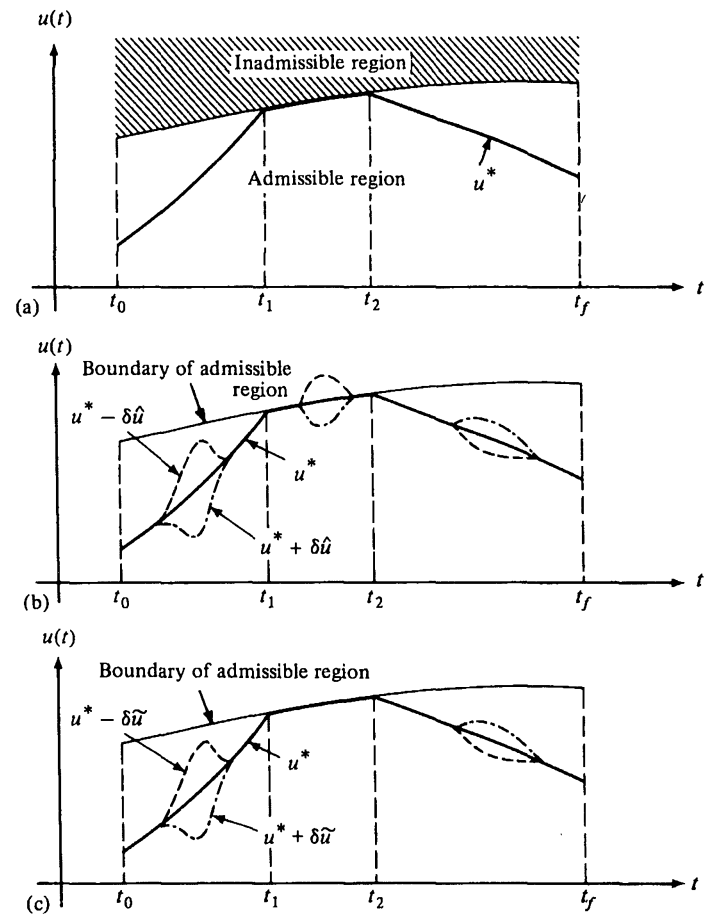


Figure 5-13 (a) An extremal control that is constrained by a boundary. (b) An admissible variation $\delta\hat{u}$ for which $-\delta\hat{u}$ is not admissible. (c) An admissible variation $\delta\tilde{u}$ for which $-\delta\tilde{u}$ is admissible

interval $[t_0, t_f]$, is shown. The differential df is the linear part of the increment Δf. Consider the end points t_0 and t_f of the interval, and admissible values of the time increment Δt, which are small enough so that the sign of Δf is determined by the sign of df. If t_0 is a point where f has a relative minimum, then $df(t_0, \Delta t)$ *must be* greater than or equal to zero. The same requirement applies for $f(t_f)$ to be a relative minimum. Thus, necessary conditions for the function f to have relative minima at the end points of the interval are

$$\begin{aligned} df(t_0, \Delta t) &\geq 0, \qquad \text{admissible } \Delta t \gtrsim 0 \\ df(t_f, \Delta t) &\geq 0, \qquad \text{admissible } \Delta t \lesssim 0, \end{aligned} \tag{5.3-4}$$

and a necessary condition for f to have a relative minimum at an interior point t, $t_0 < t < t_f$, is

$$df(t, \Delta t) = 0. \tag{5.3-5}$$

For the control problem the analogous necessary conditions are

$$\delta J(\mathbf{u}^*, \delta\mathbf{u}) \geq 0 \tag{5.3-6a}$$

if $\mathbf{u}^*$ lies on the boundary *during any portion of the time interval* $[t_0, t_f]$, and

$$\delta J(\mathbf{u}^*, \delta\mathbf{u}) = 0 \tag{5.3-6b}$$

if $\mathbf{u}^*$ lies within the boundary *during the entire time interval* $[t_0, t_f]$.

Next, let us see how this modification affects the necessary conditions, Eqs. (5.1-17) and (5.1-18), which were derived by using the assumption that the admissible control values were unconstrained. The increment of J is [if we use Eqs. (5.1-9), (5.1-13), and the definition of the Hamiltonian]

$$\begin{aligned} \Delta J(\mathbf{u}^*, \delta\mathbf{u}) = &\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f)\right]^T \delta\mathbf{x}_f \\ &+ \left[\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f)\right] \delta t_f \\ &+ \int_{t_0}^{t_f} \left\{\left[\dot{\mathbf{p}}^*(t) + \frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T \delta\mathbf{x}(t)\right. \\ &+ \left[\frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T \delta\mathbf{u}(t) \\ &\left. + \left[\frac{\partial \mathscr{H}}{\partial \mathbf{p}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) - \dot{\mathbf{x}}^*(t)\right]^T \delta\mathbf{p}(t)\right\} dt \\ &+ \text{higher-order terms.} \end{aligned} \tag{5.3-7}$$

If the state equations are satisfied, and $\mathbf{p}^*(t)$ is selected so that the coefficient of $\delta\mathbf{x}(t)$ in the integral is identically zero, and the boundary condition equation (5.1-18) is satisfied, we have

$$\Delta J(\mathbf{u}^*, \delta\mathbf{u}) = \int_{t_0}^{t_f} \left[\frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T \delta\mathbf{u}(t)\, dt + \text{higher-order terms.} \tag{5.3-8}$$

The integrand is the first-order approximation to the change in $\mathscr{H}$ caused by a change in $\mathbf{u}$ alone; that is,

$$\left[\frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)\right]^T \delta\mathbf{u}(t) \doteq \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t) + \delta\mathbf{u}(t), \mathbf{p}^*(t), t) - \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t); \tag{5.3-9}$$

therefore,

$$\Delta J(\mathbf{u}^*, \delta\mathbf{u}) = \int_{t_0}^{t_f} [\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t) + \delta\mathbf{u}(t), \mathbf{p}^*(t), t) - \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)]\, dt + \text{higher-order terms.} \tag{5.3-10}$$

If $\mathbf{u}^* + \delta\mathbf{u}$ is in a sufficiently small neighborhood of $\mathbf{u}^*$ ($\|\delta\mathbf{u}\| < \beta$) then the higher-order terms are small, and the integral in Eq. (5.3-10) dominates the expression for ΔJ. Thus, for $\mathbf{u}^*$ to be a minimizing control it is necessary that

$$\int_{t_0}^{t_f} [\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t) + \delta\mathbf{u}(t), \mathbf{p}^*(t), t) - \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)]\, dt \geq 0 \tag{5.3-11}$$

for all admissible $\delta\mathbf{u}$, such that $\|\delta\mathbf{u}\| < \beta$. We assert that in order for (5.3-11) to be satisfied for all admissible $\delta\mathbf{u}$ in the specified neighborhood, it is necessary that

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t) + \delta\mathbf{u}(t), \mathbf{p}^*(t), t) \geq \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \tag{5.3-12}$$

for all *admissible* $\delta\mathbf{u}(t)$ and for all $t \in [t_0, t_f]$. To show this, consider the control

$$\begin{aligned} \mathbf{u}(t) &= \mathbf{u}^*(t); \qquad t \notin [t_1, t_2] \\ \mathbf{u}(t) &= \mathbf{u}^*(t) + \delta\mathbf{u}(t); \qquad t \in [t_1, t_2], \end{aligned} \tag{5.3-13}$$

where $[t_1, t_2]$ is an arbitrarily small, but nonzero, time interval, and $\delta\mathbf{u}(t)$ is an admissible control variation that satisfies $\|\delta\mathbf{u}\| < \beta$.† Suppose that

† Let

$$\|\delta\mathbf{u}\| = \int_{t_0}^{t_f} \left[\sum_{i=1}^{m} |\delta u_i(t)|\right] dt.$$

Since $\mathbf{u}(t)$ is in a bounded region, each component of $\delta\mathbf{u}(t)$ is bounded and $\|\delta\mathbf{u}\|$ can be made less than β *for all admissible* $\delta\mathbf{u}(t)$ by making the interval $[t_1, t_2]$ small enough. Thus the control $\mathbf{u}(t)$ in Eq. (5.3-13) can be *any admissible control* in the interval $[t_1, t_2]$.

inequality (5.3-12) *is not satisfied* for the control described in Eq. (5.3-13); then in the interval $[t_1, t_2]$

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) < \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \tag{5.3-14}$$

and, therefore,

$$\int_{t_0}^{t_f} [\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) - \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)]\, dt$$
$$= \int_{t_1}^{t_2} [\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) - \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)]\, dt < 0. \tag{5.3-15}$$

Since the interval $[t_1, t_2]$ can be anywhere in the interval $[t_0, t_f]$, it is clear that if

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) < \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \tag{5.3-16}$$

for any $t \in [t_0, t_f]$, then it is always possible to construct an admissible control, as in Eq. (5.3-13), which makes $\Delta J < 0$, thus contradicting the optimality of the control $\mathbf{u}^*$. Our conclusion is, therefore, that a necessary condition for $\mathbf{u}^*$ to minimize the functional J is

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \leq \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) \tag{5.3-17}$$

for all $t \in [t_0, t_f]$ and for *all admissible controls.* Equation (5.3-17), which indicates that *an optimal control must minimize the Hamiltonian,* is called Pontryagin's minimum principle. Notice that we have established a necessary, but not (in general) sufficient, condition for optimality. An optimal control must satisfy Pontryagin's minimum principle; however, there may be controls that satisfy the minimum principle that are not optimal.

Let us now summarize the principal results of this section. A control $\mathbf{u}^* \in U$, which causes the system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{5.3-18}$$

to follow an admissible trajectory that minimizes the performance measure

$$J(\mathbf{u}) = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt, \tag{5.3-19}$$

is sought. In terms of the Hamiltonian

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) \triangleq g(\mathbf{x}(t), \mathbf{u}(t), t) + \mathbf{p}^T(t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)], \tag{5.3-20}$$

necessary conditions for $\mathbf{u}^*$ to be an optimal control are

$$\dot{\mathbf{x}}^*(t) = \frac{\partial \mathscr{H}}{\partial \mathbf{p}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \tag{5.3-21a}$$

$$\dot{\mathbf{p}}^*(t) = -\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \tag{5.3-21b}$$

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \leq \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) \quad \text{for all admissible } \mathbf{u}(t) \tag{5.3-21c}$$

(5.3-21a), (5.3-21b), (5.3-21c): for all $t \in [t_0, t_f]$

and

$$\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) - \mathbf{p}^*(t_f)\right]^T \delta \mathbf{x}_f + \left[\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) + \frac{\partial h}{\partial t}(\mathbf{x}^*(t_f), t_f)\right] \delta t_f = 0. \tag{5.3-22}$$

It should be emphasized that

1. $\mathbf{u}^*(t)$ is a control that causes $\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t)$ to assume its *global*, or absolute, minimum.
2. Equations (5.3-21) and (5.3-22) constitute a set of *necessary* conditions for optimality; these conditions are not, in general, sufficient.

In addition, the minimum principle, although derived for controls with values in a closed and bounded region, can also be applied to problems in which the admissible controls are not bounded. This can be done by viewing the unbounded control region as having arbitrarily large bounds, thus ensuring that the optimal control will not be constrained by the boundaries. In this case, for $\mathbf{u}^*(t)$ to minimize the Hamiltonian it is necessary (but not sufficient) that

$$\frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) = \mathbf{0}. \tag{5.3-23}$$

If Eq. (5.3-23) is satisfied, and the matrix

$$\frac{\partial^2 \mathscr{H}}{\partial \mathbf{u}^2}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)$$

is positive definite, this is sufficient to guarantee that $\mathbf{u}^*(t)$ causes $\mathscr{H}$ to be a *local* minimum; if the Hamiltonian can be expressed in the form

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) = f(\mathbf{x}(t), \mathbf{p}(t), t) + [\mathbf{c}(\mathbf{x}(t), \mathbf{p}(t), t)]^T \mathbf{u}(t) + \tfrac{1}{2}\mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t), \tag{5.3-24}$$

where $\mathbf{c}$ is an $m \times 1$ array that does not have any terms containing $\mathbf{u}(t)$, then satisfaction of (5.3-23) and $\partial^2\mathscr{H}/\partial\mathbf{u}^2 > 0$† are necessary and sufficient for $\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t)$ to be a *global* minimum.

For $\mathscr{H}$ of the form of (5.3-24),

$$\frac{\partial^2\mathscr{H}}{\partial\mathbf{u}^2}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) = \mathbf{R}(t); \tag{5.3-25}$$

thus, if $\mathbf{R}(t)$ is positive definite,

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{c}(\mathbf{x}^*(t), \mathbf{p}^*(t), t) \tag{5.3-26}$$

minimizes (globally) the Hamiltonian.

Example 5.3-1. Let us now illustrate the effect on the necessary conditions of constraining the admissible control values. Consider the system having the state equations

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -x_2(t) + u(t), \end{aligned} \tag{5.3-27}$$

with initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$. The performance measure to be minimized is

$$J(u) = \int_{t_0}^{t_f} \tfrac{1}{2}[x_1^2(t) + u^2(t)]\, dt; \tag{5.3-28}$$

t_f is specified, and the final state $\mathbf{x}(t_f)$ is free.

a. Find necessary conditions for an *unconstrained* control to minimize J. The Hamiltonian is

$$\begin{aligned} \mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = \tfrac{1}{2}x_1^2(t) + \tfrac{1}{2}u^2(t) + p_1(t)x_2(t) \\ - p_2(t)x_2(t) + p_2(t)u(t), \end{aligned} \tag{5.3-29}$$

from which the costate equations are

$$\begin{aligned} \dot{p}_1^*(t) &= -\frac{\partial\mathscr{H}}{\partial x_1} = -x_1^*(t) \\ \dot{p}_2^*(t) &= -\frac{\partial\mathscr{H}}{\partial x_2} = -p_1^*(t) + p_2^*(t). \end{aligned} \tag{5.3-30}$$

Since the control values are unconstrained, it is necessary that

$$\frac{\partial\mathscr{H}}{\partial u} = u^*(t) + p_2^*(t) = 0. \tag{5.3-31}$$

† The notation $\partial^2\mathscr{H}/\partial\mathbf{u}^2 > 0$ means that the $m \times m$ matrix $\partial^2\mathscr{H}/\partial\mathbf{u}^2$ is positive definite.

Notice that the Hamiltonian is of the form (5.3-24), and

$$\frac{\partial^2 \mathscr{H}}{\partial u^2} = 1; \tag{5.3-32}$$

therefore,

$$u^*(t) = -p_2^*(t) \tag{5.3-33}$$

does minimize the Hamiltonian. The boundary conditions are (see Table 5-1, entry 2)

$$\mathbf{p}^*(t_f) = \mathbf{0}. \tag{5.3-34}$$

b. Find necessary conditions for optimal control if

$$-1 \leq u(t) \leq +1 \qquad \text{for all } t \in [t_0, t_f]. \tag{5.3-35}$$

The state and costate equations and the boundary condition for $\mathbf{p}^*(t_f)$ remains unchanged; however, now u must be selected to minimize

$$\begin{aligned} \mathscr{H}(\mathbf{x}^*(t), u(t), \mathbf{p}^*(t)) = {} & \tfrac{1}{2}x_1^{*2}(t) + \tfrac{1}{2}u^2(t) + p_1^*(t)x_2^*(t) \\ & - p_2^*(t)x_2^*(t) + p_2^*(t)u(t) \end{aligned} \tag{5.3-36}$$

subject to the constraining relation in Eq. (5.3-35).

To determine the control that minimizes $\mathscr{H}$, we first separate all of the terms containing $u(t)$,

$$\tfrac{1}{2}u^2(t) + p_2^*(t)u(t), \tag{5.3-37}$$

from the Hamiltonian. For times when the optimal control is unsaturated, we have

$$u^*(t) = -p_2^*(t) \tag{5.3-38}$$

as in part a; clearly, this will occur when $|p_2^*(t)| \leq 1$. If, however, there are times when $|p_2^*(t)| > 1$, then from (5.3-37) the control that minimizes $\mathscr{H}$ is

$$u^*(t) = \begin{cases} -1, & \text{for } p_2^*(t) > 1 \\ +1, & \text{for } p_2^*(t) < -1. \end{cases} \tag{5.3-39}$$

Thus, $u^*(t)$ is the saturation function of $p_2^*(t)$ pictured in Fig. 5-14.

In summary, then, we have for the *unconstrained* control—part a,

$$u^*(t) = -p_2^*(t), \tag{5.3-33}$$

and, for the *constrained* control—part b,

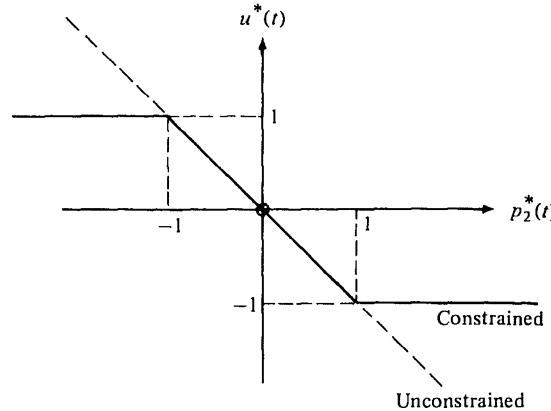


Figure 5-14 Constrained and unconstrained optimal controls for Example 5.3-1

$$u^*(t) = \begin{cases} -1, & \text{for } 1 < p_2^*(t) \\ -p_2^*(t), & \text{for } -1 \leq p_2^*(t) \leq 1 \\ +1, & \text{for } p_2^*(t) < -1. \end{cases} \tag{5.3-39a}$$

To determine $u^*(t)$ explicitly, the state and costate equations must be solved. Because of the differences in Eqs. (5.3-33) and (5.3-39a), the state-costate trajectories in the two cases will be the same only if the initial state values are such that the bounded control does not saturate. If this situation occurs, the control constraints do not affect the solution. It must be emphasized that the optimal control history for part b *cannot* be determined, in general, by calculating the optimal control history for part a and allowing it to saturate whenever the stipulated boundaries are violated.

Additional Necessary Conditions

Pontryagin and his co-workers have also derived other necessary conditions for optimality that we will find useful. We now state, without proof, two of these necessary conditions:

1. If the final time is fixed and the Hamiltonian does not depend explicitly on time, then the Hamiltonian must be a constant when evaluated on an extremal trajectory; that is,

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t)) = c_1 \qquad \text{for } t \in [t_0, t_f]. \tag{5.3-40}$$

2. If the final time is free, and the Hamiltonian does not explicitly depend on time, then the Hamiltonian must be identically zero when evaluated on an extremal trajectory; that is,

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t)) = 0 \qquad \text{for } t \in [t_0, t_f]. \tag{5.3-41}$$

State Variable Inequality Constraints

Let us now consider problems in which there may be inequality constraints that involve the state variables as well as the controls. It will be assumed that the state constraints are of the form

$$\mathbf{f}(\mathbf{x}(t), t) \geq \mathbf{0},\dagger \tag{5.3-42}$$

where $\mathbf{f}$ is an l-vector function ($l \leq m$) of the states and possibly time, which has continuous first and second partial derivatives with respect to $\mathbf{x}(t)$. It will also be assumed that the admissible control values lie in a closed and bounded region. Our approach will be to transform the l inequality constraints of (5.3-42) into a single equality constraint, and then to augment the performance measure with this equality constraint, as we have done previously with the state equations.

Let us define a new variable $\dot{x}_{n+1}(t)$ by

$$\begin{aligned}\dot{x}_{n+1}(t) \triangleq [f_1(\mathbf{x}(t), t)]^2 \mathbb{1}(-f_1) + [f_2(\mathbf{x}(t), t)]^2 \mathbb{1}(-f_2) \\ + \cdots + [f_l(\mathbf{x}(t), t)]^2 \mathbb{1}(-f_l),\end{aligned} \tag{5.3-43}$$

where $\mathbb{1}(-f_i)$ is a unit Heaviside step function defined by

$$\mathbb{1}(-f_i) = \begin{cases} 0, & \text{for } f_i(\mathbf{x}(t), t) \geq 0 \\ 1, & \text{for } f_i(\mathbf{x}(t), t) < 0, \end{cases} \tag{5.3-44}$$

for $i = 1, 2, \ldots, l$. Notice that $\dot{x}_{n+1}(t) \geq 0$ for all t, and that $\dot{x}_{n+1}(t) = 0$ only for times when *all* of the constraints (5.3-42) are satisfied. Now let us require that the variable $x_{n+1}(t)$, given by

$$x_{n+1}(t) = \int_{t_0}^{t} \dot{x}_{n+1}(t)\, dt + x_{n+1}(t_0), \tag{5.3-45}$$

satisfy the two boundary conditions $x_{n+1}(t_0) = 0$ and $x_{n+1}(t_f) = 0$. Since $\dot{x}_{n+1}(t) \geq 0$ for all t, satisfaction of these boundary conditions implies that $\dot{x}_{n+1}(t)$ must be zero throughout the interval $[t_0, t_f]$, but this occurs only if the constraints are satisfied for all $t \in [t_0, t_f]$.

Thus, to minimize the functional

$$J(\mathbf{u}) = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt \tag{5.3-46}$$

subject to the state equation constraints

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t), \tag{5.3-47}$$

† The notation $\mathbf{f}(\mathbf{x}(t), t) \geq \mathbf{0}$ means that each component of the vector $\mathbf{f}$ is ≥ 0.

admissibility constraints on the control variables, and state inequality constraints of the form

$$\mathbf{f}(\mathbf{x}(t), t) \geq \mathbf{0}, \tag{5.3-48}$$

first form the Hamiltonian

$$\begin{aligned}\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) &= g(\mathbf{x}(t), \mathbf{u}(t), t) + p_1(t)a_1(\mathbf{x}(t), \mathbf{u}(t), t) \\ &\quad + \cdots + p_n(t)a_n(\mathbf{x}(t), \mathbf{u}(t), t) \\ &\quad + p_{n+1}(t)\{[f_1(\mathbf{x}(t), t)]^2 \mathbb{1}(-f_1) + \cdots + [f_l(\mathbf{x}(t), t)]^2 \mathbb{1}(-f_l)\} \\ &\triangleq g(\mathbf{x}(t), \mathbf{u}(t), t) + \mathbf{p}^T(t)\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t),\end{aligned} \tag{5.3-49}$$

where $x_{n+1}(t)$ is given by (5.3-45), and

$$a_{n+1}(\mathbf{x}(t), t) \triangleq [f_1(\mathbf{x}(t), t)]^2 \mathbb{1}(-f_1) + \cdots + [f_l(\mathbf{x}(t), t)]^2 \mathbb{1}(-f_l). \tag{5.3-50}$$

Using the notation of (5.3-49) means that $\mathbf{p}(t)$ and $\mathbf{x}(t)$ are $n + 1$ vectors. Notice that the Hamiltonian does not contain $x_{n+1}(t)$ explicitly. We can now apply Eqs. (5.3-21) to obtain necessary conditions for optimality:

$$\left.\begin{aligned}
\dot{x}_1^*(t) &= a_1(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \\
&\vdots \\
\dot{x}_{n+1}^*(t) &= a_{n+1}(\mathbf{x}^*(t), t); \\
\dot{p}_1^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_1}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \\
&\vdots \\
\dot{p}_{n+1}^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_{n+1}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) = 0;
\end{aligned}\right\} \text{for all } t \in [t_0, t_f] \tag{5.3-51}$$

and

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \leq \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t)$$

for all admissible $\mathbf{u}(t)$.

$\dot{p}_{n+1}^*$ is zero because $x_{n+1}(t)$ does not appear explicitly in $\mathscr{H}$. The boundary conditions $\mathbf{x}^*(t_0)$ are specified [$x_{n+1}^*(t_0) = 0$ and $x_{n+1}^*(t_f) = 0$]; the remaining boundary conditions at $t = t_f$ can be determined by using the results obtained in Section 5.1.

Example 5.3-2. Let us now return to the problem discussed earlier in Example 5.3-1. The system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -x_2(t) + u(t) \tag{5.3-52}$$

is to be controlled to minimize the performance measure

$$J(u) = \int_{t_0}^{t_f} \tfrac{1}{2}[x_1^2(t) + u^2(t)]\, dt. \tag{5.3-53}$$

$\mathbf{x}(t_0)$ is specified, the final state $\mathbf{x}(t_f)$ is free, and t_f is given. The admissible control values are constrained by

$$-1 \leq u(t) \leq 1 \qquad \text{for } t \in [t_0, t_f]. \tag{5.3-54}$$

In addition, it is required that

$$-2 \leq x_2(t) \leq 2 \qquad \text{for } t \in [t_0, t_f]. \tag{5.3-55}$$

We must first express (5.3-55) in the form of (5.3-48). To do this, observe that (5.3-55) implies

$$[x_2(t) + 2] \geq 0, \tag{5.3-56a}$$

and

$$[2 - x_2(t)] \geq 0. \tag{5.3-56b}$$

Writing (5.3-55) as these two inequalities gives

$$f_1(\mathbf{x}(t)) = [x_2(t) + 2] \geq 0$$
$$f_2(\mathbf{x}(t)) = [2 - x_2(t)] \geq 0.\dagger \tag{5.3-57}$$

The Hamiltonian is given by

$$\begin{aligned}\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) &= \tfrac{1}{2}x_1^2(t) + \tfrac{1}{2}u^2(t) + p_1(t)x_2(t) \\ &\quad - p_2(t)x_2(t) + p_2(t)u(t) + p_3(t)\{[x_2(t) + 2]^2\, \mathbb{1}(-x_2(t) - 2) \\ &\quad + [2 - x_2(t)]^2\, \mathbb{1}(x_2(t) - 2)\}\end{aligned} \tag{5.3-58}$$

The necessary conditions for optimality, found from Eqs. (5.3-51), are

$$\begin{aligned}&\dot{x}_1^*(t) = x_2^*(t), \qquad x_1^*(t_0) = x_{1_0} \\ &\dot{x}_2^*(t) = -x_2^*(t) + u^*(t), \qquad x_2^*(t_0) = x_{2_0} \\ &\dot{x}_3^*(t) = [x_2^*(t) + 2]^2\, \mathbb{1}(-x_2^*(t) - 2) \\ &\qquad + [2 - x_2^*(t)]^2\, \mathbb{1}(x_2^*(t) - 2), \qquad x_3^*(t_0) = 0\end{aligned} \tag{5.3-59}$$

$$\dot{p}_1^*(t) = -\frac{\partial \mathscr{H}}{\partial x_1} = -x_1^*(t)$$

† We could also combine the inequalities (5.3-56) by writing $[x_2(t) + 2][2 - x_2(t)] \geq 0$.

$$\dot{p}_2^*(t) = -\frac{\partial \mathscr{H}}{\partial x_2} = -p_1^*(t) + p_2^*(t) - 2p_3^*(t)[x_2^*(t) + 2]\mathbb{1}(-x_2^*(t) - 2) + 2p_3^*(t)[2 - x_2^*(t)]\mathbb{1}(x_2^*(t) - 2)\dagger$$

$$\dot{p}_3^*(t) = -\frac{\partial \mathscr{H}}{\partial x_3} = 0 \Rightarrow p_3^*(t) = \text{a constant} \tag{5.3-60}$$

$$u^*(t) = \begin{cases} -1, & \text{for } 1 < p_2^*(t) \\ -p_2^*(t), & \text{for } -1 \leq p_2^*(t) \leq 1 \\ +1, & \text{for } p_2^*(t) < -1. \end{cases} \tag{5.3-61}$$

The boundary conditions at the final time are $x_3^*(t_f) = 0$ (specified), and $p_1^*(t_f) = p_2^*(t_f) = 0$—from Table 5-1, or Eq. (5.1-18).

Comparing these necessary conditions with the results obtained in Example 5.3-1b, we see that the expressions for the optimal controls in terms of the extremal costates are the same; however, the equations for $\dot{p}_2^*(t)$ are different because of the presence of the state inequality constraints; hence, the optimal trajectories and control histories will generally not be the same.

In our discussion of state and control inequality constraints we have not considered constraints that include both the states and controls, that is, constraints of the form

$$\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) \geq \mathbf{0}. \tag{5.3-62}$$

For an explanation of how to handle constraints of this form, as well as an alternative derivation of the minimum principle, the interested reader can refer to Chapter 4 of [S-3].

In the remainder of this chapter we shall consider several examples of the application of Pontryagin's minimum principle. These examples will illustrate both the utility and the limitations of the variational approach to optimal control problems.

5.4 MINIMUM-TIME PROBLEMS

In this section we shall consider problems in which the objective is to transfer a system from an arbitrary initial state to a specified target set in minimum time. The target set (which may be moving) will be denoted by

† Performing the differentiation $\partial \mathscr{H}/\partial x_2$ formally also results in the presence of two unit impulse functions, which occur at $x_2^*(t) = \pm 2$; however, these terms are such that either the impulse functions or their coefficients are zero for all $t \in [t_0, t_f]$, so the impulses do not affect the solution.

$S(t)$, and the minimum time required to reach the target set by t^*. Mathematically, then, our problem is to transfer a system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{5.4-1}$$

from an arbitrary initial state $\mathbf{x}_0$ to the target set $S(t)$ and minimize

$$J(\mathbf{u}) = \int_{t_0}^{t_f} dt = t_f - t_0. \tag{5.4-2}$$

Typically, the control variables may be constrained by requirements such as

$$|u_i(t)| \leq 1, \qquad i = 1, 2, \ldots, m, \qquad t \in [t_0, t^*]. \tag{5.4-3}$$

Our approach will be to use the minimum principle to determine the optimal control law.†

To introduce several important aspects of minimum-time problems, let us consider the following simplified intercept problem.

Example 5.4-1. Figure 5-15 shows an aircraft that is initially at the point $x = 0$, $y = 0$ pursuing a ballistic missile that is initially at the point $x = a > 0$, $y = 0$. The missile flies the trajectory

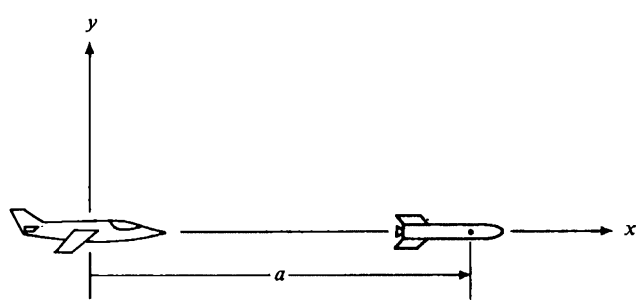


Figure 5-15 An intercept problem

$$\begin{aligned} x_M(t) &= a + 0.1t^3 \\ y_M(t) &= 0 \end{aligned} \tag{5.4-4}$$

for $t \geq 0$; thus, in this example the target set $S(t)$ is the position of the missile given by (5.4-4).

Neglecting gravitational and aerodynamic forces, let us model the aircraft as a point mass. Normalizing the mass to unity, we find that the motion of the aircraft in the x direction is described by

$$\ddot{x}(t) = u(t), \tag{5.4-5}$$

† For additional reading on time-optimal systems see [P-1] and [A-2].

or, in state form,

$$\dot{x}_1(t) = x_2(t) \tag{5.4-6}$$
$$\dot{x}_2(t) = u(t),$$

where $x_1(t) \triangleq x(t)$ and $x_2(t) \triangleq \dot{x}(t)$. The thrust $u(t)$ is constrained by the relationship

$$|u(t)| \leq 1.0. \tag{5.4-7}$$

By inspection of the geometry of the problem, it is clear that the optimal strategy for the pursuing aircraft is to accelerate with the maximum thrust possible in the positive x direction; therefore, $u^*(t)$ should be $+1.0$ for $t \in [0, t^*]$. To find t^*, we must determine the value(s) of t for which the x coordinate of the aircraft coincides with the target set $S(t)$; hence, assuming $\dot{x}(0) = 0$, we solve the equation

$$\tfrac{1}{2}[t^*]^2 = a + 0.1[t^*]^3 \tag{5.4-8}$$

for t^*. Common sense indicates that there may not be a positive real value of $t^* \geq 0$ for which Eq. (5.4-8) is satisfied—if the missile is far enough away initially he can escape. It can be shown that interception is impossible if a is greater than 1.85. If $a = 1.85$, interception occurs at $t^* = 3.33$; for $a < 1.85$ the minimum interception times are less than 3.33.

Although greatly simplified, the preceding example illustrates two important characteristics that are typical of minimum-time problems:

1. For certain values of the initial condition a, a time-optimal control does not exist.
2. The optimal control, if it exists, is maximum effort during the entire time interval of operation.

In the subsequent development we shall generalize these concepts; let us first consider the question of existence of an optimal control.

The Set of Reachable States

If a system can be transferred from some initial state to a target set by applying admissible control histories, then an optimal control exists and may be found by determining the admissible control that causes the system to reach the target set most quickly. A description of the target set is assumed to be known; thus, to investigate the existence of an optimal control it is useful to introduce the concept of reachable states.

DEFINITION 5-1

If a system with initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ is subjected to *all* admissible control histories for a time interval $[t_0, t]$, the collection of state values $\mathbf{x}(t)$ is called the set of states that are reachable (from $\mathbf{x}_0$) at time t, or simply *the set of reachable states.*

Although the set of reachable states depends on $\mathbf{x}_0$, t_0, and on t, we shall denote this set by $R(t)$. The following example illustrates the concept of reachable states.

Example 5.4-2. Find the set of reachable states for the system

$$\dot{x}(t) = u(t), \tag{5.4-9}$$

where the admissible controls satisfy

$$-1 \leq u(t) \leq 1. \tag{5.4-10}$$

The solution of Eq. (5.4-9) is

$$x(t) = x_0 + \int_{t_0}^{t} u(\tau)\, d\tau. \tag{5.4-11}$$

If only admissible control values are used, Eq. (5.4-11) implies that

$$x_0 - [t - t_0] \leq x(t) \leq x_0 + [t - t_0]. \tag{5.4-12}$$

Figure 5-16 shows the reachable sets for $t = t_1, t_2$, and t_3, where $t_1 < t_2 < t_3$.

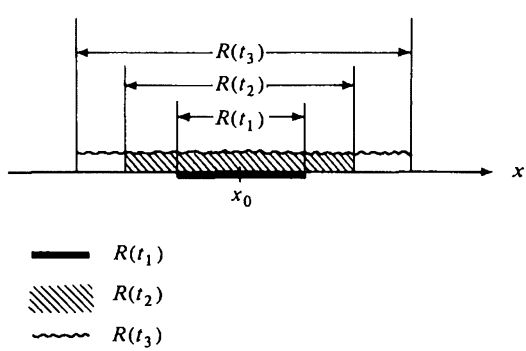


Figure 5-16 The reachable states for Example 5.4-2

The concept and properties of reachable sets are inextricably intertwined with the question of existence of time-optimal controls; if there is no value

of t for which the target set $S(t)$ has at least one point in common with the set $R(t)$, then a time-optimal control does not exist. Conversely, it is helpful to visualize the minimum-time problem as a matter of finding the earliest time t^* when $S(t)$ and $R(t)$ meet, as shown in Fig. 5-17 for a second-order system. The target set is a moving point, and the boundary of the set of reachable states at time t_i is denoted by $\partial R(t_i)$. The target set and the set of reachable states first intersect at point p, where $t^* = t_2$.

Unfortunately, although it is conceptually satisfying to think of minimum-time problems in this fashion, it is generally not feasible to determine solutions by finding the intersections of reachable sets with the target set except in very simple problems (like Example 5.4-1). General theorems concerning the existence of time-optimal controls are unavailable at this time; however, later in this section we shall state an existence theorem that applies to an important class of minimum-time problems.

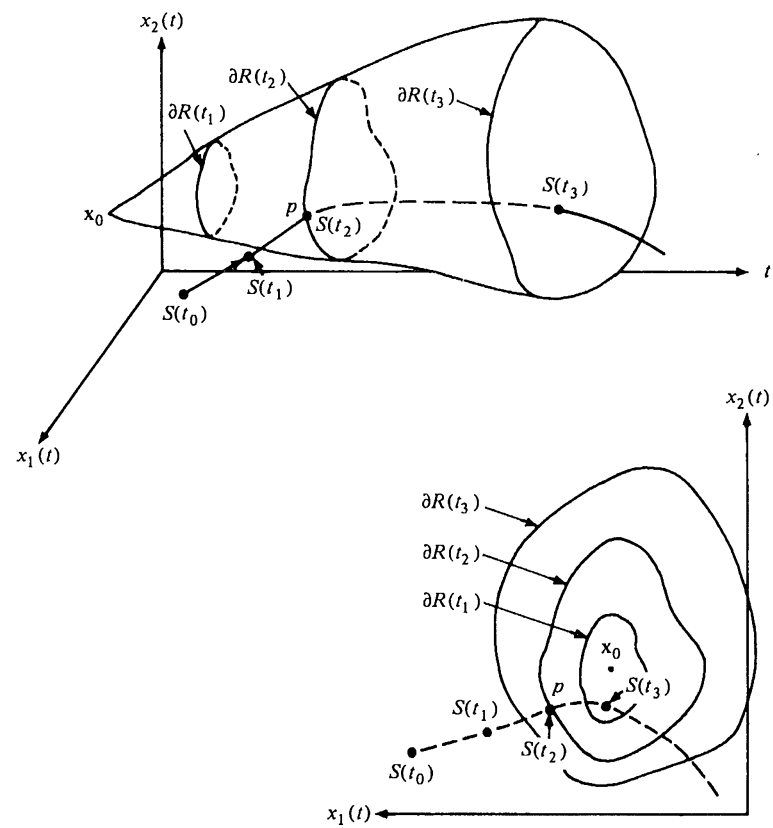


Figure 5-17 The minimum-time problem viewed as the intersection of a target set, $S(t)$, and the set of reachable states, $R(t)$

The Form of the Optimal Control for a Class of Minimum-Time Problems

Now let us determine the form of the optimal control for a particular class of systems by using the minimum principle. We shall assume that the state equations of the system are of the form

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), t) + \mathbf{B}(\mathbf{x}(t), t)\mathbf{u}(t), \tag{5.4-13}$$

where $\mathbf{B}$ is an $n \times m$ array that may be explicitly dependent on the states and time. It is specified that the admissible controls must satisfy the inequality constraints

$$M_{i-} \leq u_i(t) \leq M_{i+}, \qquad i = 1, 2, \ldots, m, \qquad t \in [t_0, t^*]; \tag{5.4-14}$$

M_{i+} and M_{i-} are known upper and lower bounds for the ith control component.

The Hamiltonian is

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) = 1 + \mathbf{p}^T(t)[\mathbf{a}(\mathbf{x}(t), t) + \mathbf{B}(\mathbf{x}(t), t)\mathbf{u}(t)]. \tag{5.4-15}$$

From the minimum principle, it is necessary that

$$\begin{aligned} &1 + \mathbf{p}^{*T}(t)[\mathbf{a}(\mathbf{x}^*(t), t) + \mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}^*(t)] \\ &\leq 1 + \mathbf{p}^{*T}(t)[\mathbf{a}(\mathbf{x}^*(t), t) + \mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}(t)] \end{aligned} \tag{5.4-16}$$

for all admissible $\mathbf{u}(t)$, and for all $t \in [t_0, t^*]$. Equation (5.4-16) implies that

$$\mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}^*(t) \leq \mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}(t); \tag{5.4-17}$$

hence, $\mathbf{u}^*(t)$ is the control that causes $\mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}(t)$ to assume its minimum value. If the array $\mathbf{B}$ is expressed as

$$\mathbf{B}(\mathbf{x}^*(t), t) = \left[\mathbf{b}_1(\mathbf{x}^*(t), t) \,\vdots\, \mathbf{b}_2(\mathbf{x}^*(t), t) \,\vdots\, \cdots \,\vdots\, \mathbf{b}_m(\mathbf{x}^*(t), t)\right], \tag{5.4-18}$$

where $\mathbf{b}_i(\mathbf{x}^*(t), t)$, $i = 1, \ldots, m$, is the ith column of the array, then the coefficient of the ith control component $u_i(t)$ in (5.4-17) is $\mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)$, and

$$\mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}(t) = \sum_{i=1}^{m} \mathbf{p}^{*T}(t)[\mathbf{b}_i(\mathbf{x}^*(t), t)]u_i(t). \tag{5.4-19}$$

Assuming that the control components are independent of one another, we then must minimize

$$\mathbf{p}^{*T}(t)[\mathbf{b}_i(\mathbf{x}^*(t), t)]u_i(t)$$

with respect to $u_i(t)$ for $i = 1, 2, \ldots, m$. If the coefficient of $u_i(t)$ is positive, $u_i^*(t)$ must be the smallest admissible control value M_{i-}. If the coefficient of $u_i(t)$ is negative, $u_i^*(t)$ must be the largest admissible control value M_{i+}; thus, the form of the optimal control is

$$u_i^*(t) = \begin{cases} M_{i+}, & \text{for } \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) < 0 \\ M_{i-}, & \text{for } \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) > 0 \\ \text{Undetermined}, & \text{for } \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) = 0. \end{cases} \qquad (5.4\text{-}20)$$

$$i = 1, 2, \ldots, m$$

If the extremal state and costate trajectories are such that the coefficient of $u_i(t)$ is as shown in Fig. 5-18(a), then the history of $u_i^*(t)$ will be as shown in Fig. 5-18(b).

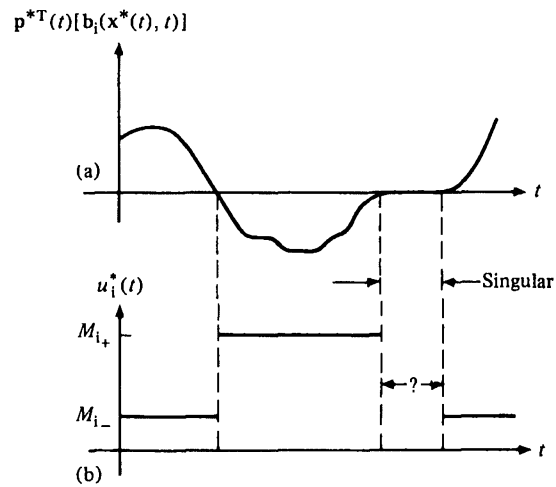


Figure 5-18 The relationship between a time-optimal control and its coefficient in the Hamiltonian

Notice that if $\mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)$ passes through zero, a switching of the control $u_i^*(t)$ is indicated. If $\mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)$ is zero for some finite time interval, then the coefficient of $u_i(t)$ in the Hamiltonian is zero, so the necessary condition that $u_i^*(t)$ minimize $\mathscr{H}$ provides no information about how to select $u_i^*(t)$; this signals the so-called *singular condition*, to be discussed in Section 5.6. Here we shall consider only problems in which the singular condition does not arise; such problems will be called *normal.*

Equation (5.4-20) is the mathematical statement of the well-known *bang-bang principle*, that is, if the state equations are of the form (5.4-13) and the admissible controls must satisfy constraints of the form (5.4-14),

then *the optimal control to obtain minimum-time response is maximum effort throughout the interval of operation.* The bang-bang concept is intuitively appealing as well. Certainly, the men who race automobiles come very close to bang-bang operation— they use the accelerator and brakes often; thus, their fuel consumption is large, tires and brakes do not last very long, and the cars are subjected to severe mechanical stresses, but barring accidents and mechanical failures, the drivers reach their destination quickly.

Before we move on to some problems that can be completely solved by using analytical methods, let us consider a nonlinear problem of the foregoing type.

Example 5.4-3.† Figure 5-19 shows a lunar rocket in the terminal phase of a minimum-time, soft landing on the surface of the moon. We shall make the following assumptions:

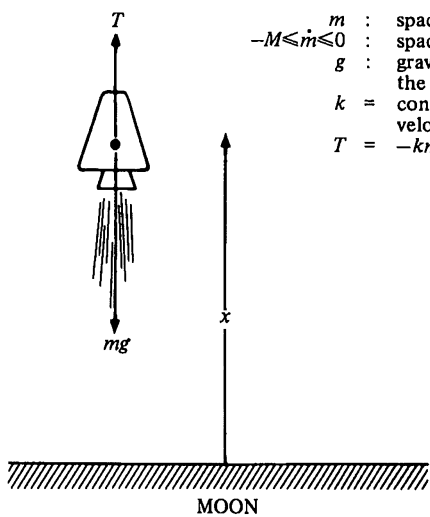


Figure 5-19 Lunar soft landing

a. Aerodynamic forces and gravitational forces of bodies other than the moon are negligible.
b. Lateral motion is ignored; thus, the descent trajectory is vertical and the thrust vector is tangent to the trajectory.
c. The acceleration of gravity is a constant, because of the nearness of the spacecraft to the moon.
d. The relative velocity of the exhaust gases with respect to the spacecraft is constant.

† See [M-2] and [M-3].

e. The mass rate is constrained by

$$-M \leq \dot{m} \leq 0. \tag{5.4-21}$$

The equation of motion is

$$\begin{aligned} m(t)\ddot{x}(t) &= -gm(t) + T(t) \\ &= -gm(t) - k\dot{m}(t). \end{aligned} \tag{5.4-22}$$

Defining the states of the system as $x_1 \triangleq x$, $x_2 \triangleq \dot{x}$, $x_3 \triangleq m$ and the control as $u \triangleq \dot{m}$ leads to the state equations

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -g - \frac{k}{x_3(t)}u(t) \\ \dot{x}_3(t) &= u(t). \end{aligned} \tag{5.4-23}$$

The Hamiltonian is

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = 1 + p_1(t)x_2(t) - gp_2(t) - \frac{kp_2(t)u(t)}{x_3(t)} + p_3(t)u(t), \tag{5.4-24}$$

and the optimal control must satisfy

$$\mathscr{H}(\mathbf{x}^*(t), u^*(t), \mathbf{p}^*(t)) \leq \mathscr{H}(\mathbf{x}^*(t), u(t), \mathbf{p}^*(t))$$

for all admissible $u(t)$, and for all $t \in [t_0, t_f]$; therefore,

$$u^*(t) = \begin{cases} 0, & \text{for } p_3^*(t) - \dfrac{kp_2^*(t)}{x_3^*(t)} < 0 \\ -M, & \text{for } p_3^*(t) - \dfrac{kp_2^*(t)}{x_3^*(t)} > 0 \\ \text{Undetermined}, & \text{for } p_3^*(t) - \dfrac{kp_2^*(t)}{x_3^*(t)} = 0. \end{cases} \tag{5.4-25}$$

To obtain an explicit solution for $u^*(t)$ we would have to solve a nonlinear two-point boundary-value problem (see Problem 5-31).

Minimum-Time Control of Time-Invariant Linear Systems

Armed with our knowledge about the form of time-optimal controls, for the remainder of this section we shall consider the following important class of problems: A linear, stationary system of order n having m controls is described by the state equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{5.4-26}$$

where **A** and **B** are constant $n \times n$ and $n \times m$ matrices, respectively. The components of the control vector are constrained by

$$|u_i(t)| \leq 1, \qquad i = 1, 2, \ldots, m. \tag{5.4-27}$$

Assuming that the system is completely controllable and normal (no singular intervals exist), find a control, if one exists, which transfers the system from an arbitrary initial state $\mathbf{x}_0$ at time $t = 0$ to the final state $\mathbf{x}(t_f) = \mathbf{0}$ in minimum time. We shall refer to this problem as the *stationary, linear regulator, minimum-time problem.*

From Eq. (5.4-20) we know that the optimal control, if it exists, is bang-bang. Let us now state without proof some important theorems due to Pontryagin et al. [P-1] which apply to stationary, linear regulator, minimum-time problems.

THEOREM 5.4-1 (EXISTENCE)

If *all* of the eigenvalues of **A** have nonpositive real parts, then an optimal control exists that transfers any initial state $\mathbf{x}_0$ to the origin.

THEOREM 5.4-2 (UNIQUENESS)

If an extremal control exists, then it is unique.†

Since an optimal control, if one exists, must be an extremal control, this theorem indicates that a control which satisfies the minimum principle and the required boundary conditions must be the optimal control. Thus, if an optimal control exists, satisfaction of the minimum principle is both necessary and sufficient for time-optimal control of stationary, linear regulator systems.

THEOREM 5.4-3 (NUMBER OF SWITCHINGS)

If the eigenvalues of **A** are all real, and a (unique) time-optimal control exists, then each control component can switch *at most* $(n - 1)$ times.

Thus, an nth-order system having all real, nonpositive eigenvalues has a unique time-optimal control with components that each switch at most $(n - 1)$ times.

Example 5.4-4. Find the optimal control satisfying

$$|u(t)| \leq 1 \tag{5.4-28}$$

which transfers the system

† Recall that a control which satisfies the necessary conditions in Eqs. (5.3-21) and the required boundary conditions is called an extremal control.

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u(t) \end{aligned} \tag{5.4-29}$$

from any initial state $\mathbf{x}_0$ to the origin in minimum time. Here

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{5.4-30}$$

Since the eigenvalues of A are both zero, we know from Theorems 5.4-1 through 5.4-3 that an optimal control exists, is unique, and has at most one switching.

The Hamiltonian is

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = 1 + p_1(t)x_2(t) + p_2(t)u(t); \tag{5.4-31}$$

thus, the minimum principle indicates that the optimal control $u^*(t)$ must satisfy

$$p_2^*(t)u^*(t) \leq p_2^*(t)u(t) \tag{5.4-32}$$

for all admissible $u(t)$ and for all $t \in [t_0, t_f]$. It can be shown that a singular interval cannot exist (see Section 5.6); therefore, the optimal control found from (5.4-32) is

$$u^*(t) = \left\{ \begin{matrix} -1, & \text{for } p_2^*(t) > 0 \\ +1, & \text{for } p_2^*(t) < 0 \end{matrix} \right\} \triangleq -\operatorname{sgn}(p_2^*(t)). \tag{5.4-33}$$

From the Hamiltonian the costate equations are

$$\begin{aligned} \dot{p}_1^*(t) &= 0 \\ \dot{p}_2^*(t) &= -p_1^*(t). \end{aligned} \tag{5.4-34}$$

The costate solution is of the form

$$\begin{aligned} p_1^*(t) &= c_1 \\ p_2^*(t) &= -c_1 t + c_2, \end{aligned} \tag{5.4-35}$$

where c_1 and c_2 are constants of integration. Equation (5.4-35) indicates that p_2^*, and therefore u^*, can change sign at most once (this result also follows from Theorem 5.4-3).

Since there can be *at most* one switching, the optimal control for a specified initial state must be one of the forms:

$$u^*(t) = \begin{cases} +1, & \text{for all } t \in [t_0, t^*], \text{ or} \\ -1, & \text{for all } t \in [t_0, t^*], \text{ or} \\ +1, & \text{for } t \in [t_0, t_1),\dagger \text{ and } -1, \text{ for } t \in [t_1, t^*], \text{ or} \\ -1, & \text{for } t \in [t_0, t_1), \text{ and } +1, \text{ for } t \in [t_1, t^*]. \end{cases} \tag{5.4-36}$$

† The notation $t \in [t_0, t_1)$ means $t_0 \leq t < t_1$.

Thus, segments of optimal trajectories can be found by integrating the state equations with $u = \pm 1$ to obtain

$$x_2(t) = \pm t + c_3 \tag{5.4-37}$$

$$x_1(t) = \pm \tfrac{1}{2}t^2 + c_3 t + c_4, \tag{5.4-38}$$

where c_3 and c_4 are constants of integration, and the upper sign corresponds to $u = +1$. Time can be eliminated from these equations by squaring the first equation, multiplying the result by $\frac{1}{2}$ and comparing with Eq. (5.4-38) to obtain

$$x_1(t) = \tfrac{1}{2}x_2^2(t) + c_5, \qquad \text{for } u = +1 \tag{5.4-39}$$

and

$$x_1(t) = -\tfrac{1}{2}x_2^2(t) + c_6, \quad \text{for } u = -1; \tag{5.4-40}$$

c_5 and c_6 are constants. Equations (5.4-39) and (5.4-40) each define a family of parabolas that are shown in Fig. 5-20(a) and (b)—the arrows indicate the direction of increasing time.

Now, let us consider each of the alternatives for the optimal control. From Fig. 5-20 we see that the controls given by Eq. (5.4-36) correspond to the following situations:

1. $u^*(t) = +1$ for $t \in [t_0, t^*]$. The initial state $\mathbf{x}_0$ must lie on segment A-0 in Fig. 5-20(a).
2. $u^*(t) = -1$ for $t \in [t_0, t^*]$. The initial state $\mathbf{x}_0$ must lie on segment B-0 in Fig. 5-20(b).
3. $u^*(t) = +1$ for $t \in [t_0, t_1)$, and $u^*(t) = -1$ for $t \in [t_1, t^*]$. Since the optimal control is -1 for $t \in [t_1, t^*]$, at time t_1 the system state must lie on segment B-0. This transfer has been accomplished by a control of $u^* = +1$; thus, the optimal trajectory consists of an initial segment like one of the trajectories in Fig. 5-20(a) followed by a switching of the control to -1 upon reaching B-0, and then on to the origin along B-0 with $u^* = -1$. Notice that B-0, in addition to being the terminal segment of the optimal trajectory, is the locus of state values where the control switches from $+1$ to -1; therefore, B-0 is referred to *as a switching curve*. Now, which initial states will have optimal trajectories as described above? Again referring to Fig. 5-20, we see that only the parabolic curves that have $c_5 < 0$ intersect B-0. In addition, only trajectories that begin below B-0 with $u^* = +1$ will ever intersect B-0. We conclude that for initial states lying below both A-0 and B-0 the optimal control will be $u^* = +1$ until B-0 is reached, followed by $u^* = -1$ thereafter.
4. $u^*(t) = -1$ for $t \in [t_0, t_1)$, and $u^*(t) = +1$ for $t \in [t_1, t^*]$. The same reasoning used in 3 leads to the conclusion that for states initially lying above A-0 and B-0 the optimal control will be $u^* = -1$ followed by $u^* = +1$; the switching occurs when the trajectory intersects A-0.

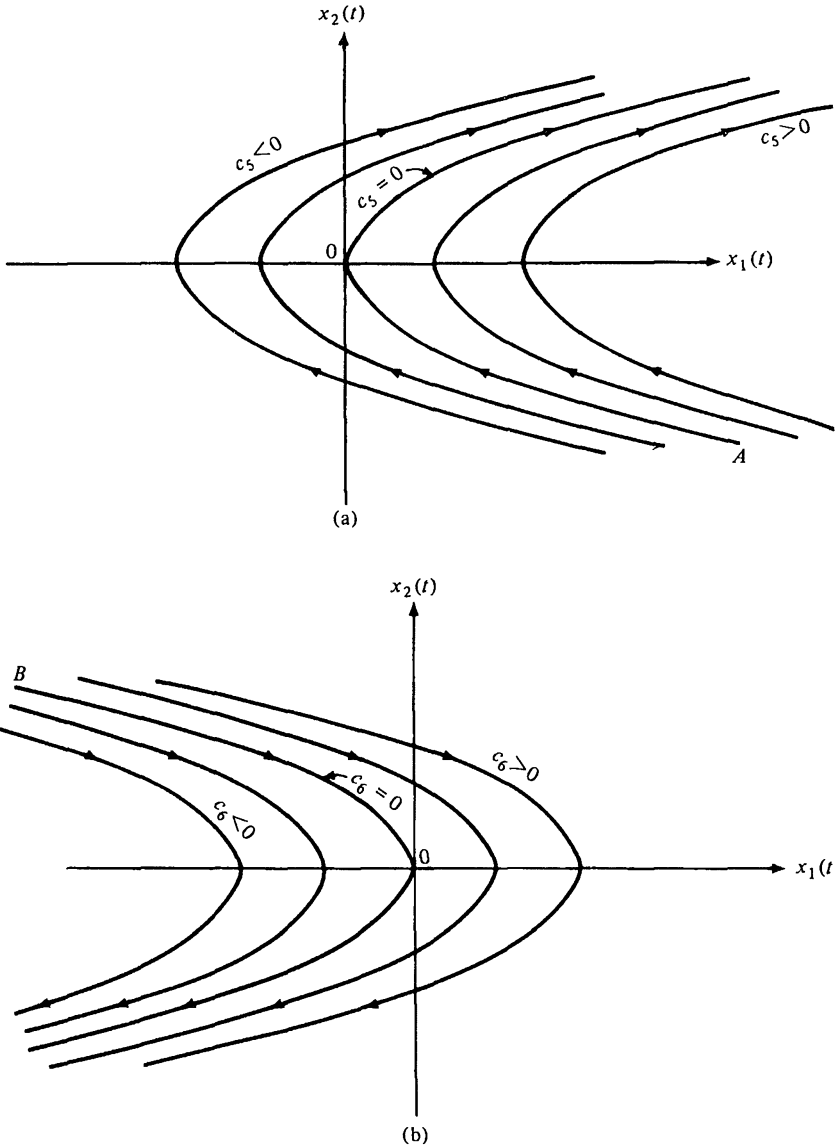


Figure 5-20 (a) Trajectories for $u = +1$. (b) Trajectories for $u = -1$

Thus, we see that A-0 and B-0, in addition to being terminal segments of optimal trajectories, together compose the switching curve A-0-B shown in Fig. 5-21(a). By putting $c_5 = c_6 = 0$ in Eqs. (5.4-39) and (5.4-40), we find the equation of this switching curve to be

$$x_1(t) = -\tfrac{1}{2}x_2(t)\,|x_2(t)|. \tag{5.4-41}$$

To summarize, for states above A-0-B the optimal control is $u^* = -1$ until the trajectory intersects A-0, where the optimal control switches to

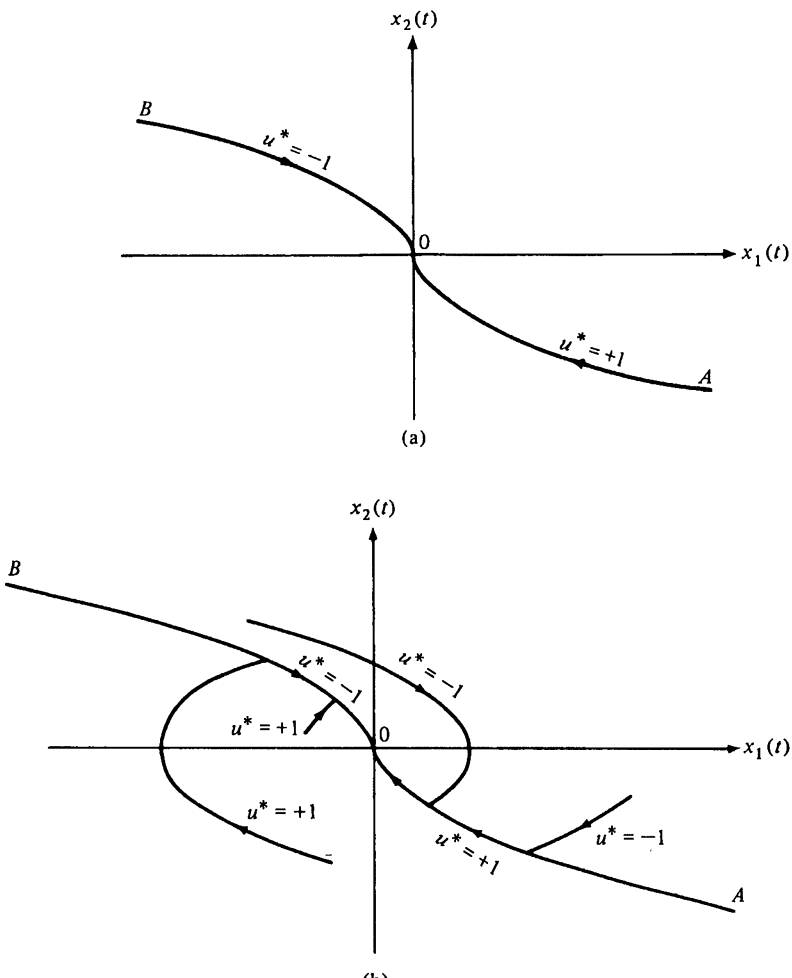


Figure 5-21 (a) The switching curve. (b) Optimal trajectories for several initial state values.

$u^* = +1$. The optimal control $u^* = +1$ is applied to transfer states below A-0-B to segment B-0, where the optimal control switches to $u^* = -1$. Once the system has reached the origin, it can be kept there by applying $u^*(t) = 0$ for $t > t^*$. Optimal trajectories for several initial state values are shown in Fig. 5-21(b).

It must be emphasized that we have succeeded in obtaining the optimal control *law*; that is, the optimal control at any time t is known as a function of the state value $\mathbf{x}(t)$. To express the optimal control law in a convenient form, let us define the *switching function* $s(\mathbf{x}(t))$, obtained from Eq. (5.4-41) as

$$s(\mathbf{x}(t)) \triangleq x_1(t) + \tfrac{1}{2}x_2(t)\,|x_2(t)|. \tag{5.4-42}$$

Notice that

$s(\mathbf{x}(t)) > 0$ implies $\mathbf{x}(t)$ lies above the switching curve A-0-B.

$s(\mathbf{x}(t)) < 0$ implies $\mathbf{x}(t)$ lies below the switching curve A-0-B.

$s(\mathbf{x}(t)) = 0$ implies $\mathbf{x}(t)$ lies on the switching curve A-0-B.

Thus, in terms of this switching function the optimal control law is

$$u^*(t) = \begin{cases} -1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) > 0 \\ +1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) < 0 \\ -1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) = 0 \text{ and } x_2(t) > 0 \\ +1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) = 0 \text{ and } x_2(t) < 0 \\ 0, & \text{for } \mathbf{x}(t) = 0. \end{cases} \tag{5.4-43}$$

An implementation of this optimal control law is shown in Fig. 5-22; the required hardware consists of a summing device, a sign changer, a nonlinear function generator, and an ideal relay.

The procedure used in solving the preceding example can be generalized to include nth-order, stationary, linear regulator systems controlled by one input. Let us assume that all of the eigenvalues of $\mathbf{A}$ are real and nonpositive; thus, for all initial states a unique time-optimal control exists and has at most $(n - 1)$ switchings. To obtain the optimal control law:

1. (a) We first determine the set of points from which the origin can be reached with $u = +1$ (call this set O_+), and the set of points from which the origin can be reached with $u = -1$ (call this set O_-). Let O_1 denote the set of points from which the origin can be reached with no control switchings; then

$$O_1 = O_+ \cup O_- \tag{5.4-44}$$

where $\cup$ denotes "the union of."†

† O_1 is the union of O_+ and O_-; this means that every element of O_1 is an element of either O_+, O_-, or both O_+ and O_-.

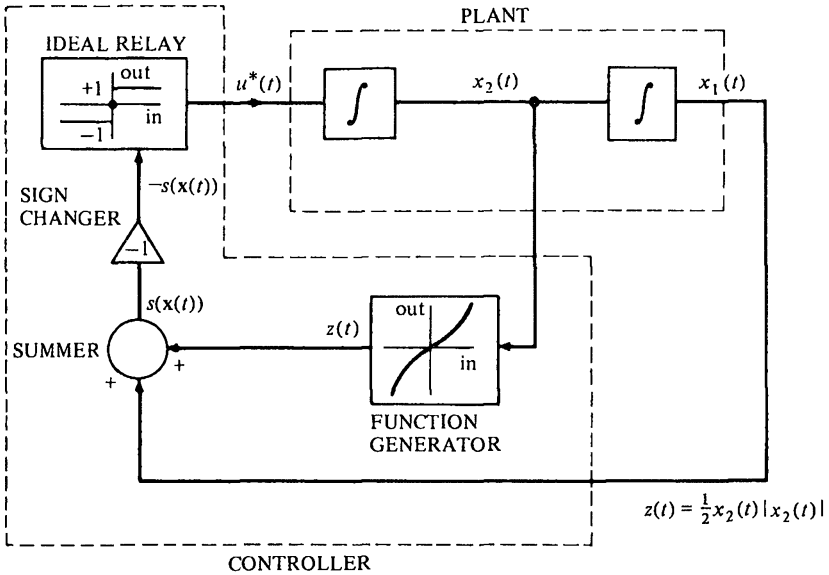


Figure 5-22 Implementation of the time-optimal control law for Example 5.4-4

(b) Next, we determine the set of points O_{-+} from which O_+ can be reached by applying $u = -1$; the origin can be reached from O_{-+} by applying $u = -1$ until reaching O_+, followed by $u = +1$. Similarly, we find the set of points O_{+-} from which O_- can be reached by applying $u = +1$. To reach the origin from O_{+-}, we apply $u = +1$ until reaching O_-, followed by $u = -1$. The set of points from which the origin can be reached with at most one switching (two control values) is given by

$$\begin{aligned} O_2 &= O_+ \cup O_- \cup O_{+-} \cup O_{-+} \\ &= O_1 \cup O_{+-} \cup O_{-+}\dagger \end{aligned} \tag{5.4-45}$$

(c) We continue until the set of points O_{n-1} from which the origin can be reached with at most $(n-2)$ switchings is determined. All points not in the set O_{n-1} require $(n-1)$ switchings to reach the origin. By eliminating time from the trajectory equations, express O_{n-1} in the form

$$s(\mathbf{x}(t)) = 0. \tag{5.4-46}$$

† $O_1 \cup O_{+-} \cup O_{-+}$ means the set of points which are in at least one of the sets O_1, O_{+-}, O_{-+}.

2. Next, we determine the optimal control to be applied at any point in the state space. The switching function $s(\mathbf{x}(t))$ defines a switching hypersurface that divides the state space into two half-spaces. From one half-space the control $u^* = +1$ is applied to drive the system to O_{n-1}, where the control switches to -1, until the system reaches O_{n-2}, where the control again switches to $+1$, etc., until the origin is reached. From the other half-space the control sequence is reversed; $u^* = -1$ is applied to transfer the system to O_{n-1}, where the control switches to $+1$, and so on, until reaching the origin.
3. Finally, we determine a combination of physical devices to implement the time-optimal control law.

Before concluding our consideration of time-optimal problems, let us solve another second-order example that illustrates the procedure we have just summarized.

Example 5.4-5. Find the control law for transferring the system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -ax_2(t) + u(t) \end{aligned} \tag{5.4-47}$$

from an arbitrary initial state $\mathbf{x}_0$ to the origin in minimum time. The admissible controls are constrained by

$$|u(t)| \leq 1, \tag{5.4-48}$$

and a is a positive real number.

The eigenvalues of this system are 0 and $-a$; thus, since both eigenvalues are real and nonpositive, the hypotheses of Theorems 5.4-1 through 5.4-3 are satisfied and we know that an optimal control exists that is unique and has at most $(n - 1)$ switchings.

We are again dealing with a second-order system; thus, we know that the optimal control law is determined by a switching *curve*—in higher-dimensional problems switchings occur on hypersurfaces in the state space.

From the minimum principle and Theorem 5.4-3, we find that the possible forms for the optimal control are the same as given for Example 5.4-4 in Eq. (5.4-36).

Next, let us proceed to find the sets O_+ and O_- (from which the origin can be reached by applying only $u = +1$, or $u = -1$) by solving the differential equations (5.4-47) with $u = \pm 1$. The solutions are

$$x_2(t) = c_1\epsilon^{-at} \pm \frac{1}{a}[1 - \epsilon^{-at}] \tag{5.4-49}$$

$$x_1(t) = -\frac{c_1}{a}\epsilon^{-at} \pm \frac{1}{a}t \pm \frac{1}{a^2}\epsilon^{-at} + c_2 \tag{5.4-50}$$

These equations define two families of curves; to determine the curves which pass through the origin, set $x_1(t) = x_2(t) = 0$ and $t = 0$ (since the system is time invariant, $t = 0$ is an arbitrary reference time), solve for c_1 and c_2, and substitute in (5.4-49) and (5.4-50) to obtain

$$x_2(t) = \pm \frac{1}{a}[1 - \epsilon^{-at}] \tag{5.4-51}$$

$$x_1(t) = \pm \frac{1}{a}t \pm \frac{1}{a^2}\epsilon^{-at} \mp \frac{1}{a^2}. \tag{5.4-52}$$

To determine O_+, use the upper sign (which corresponds to $u = +1$), solve (5.4-51) for t, and substitute in (5.4-52) to obtain the relationship

$$x_1(t) = -\frac{1}{a^2}\ln\left(-a\left[x_2(t) - \frac{1}{a}\right]\right) - \frac{1}{a}x_2(t).\dagger \tag{5.4-53}$$

The set of points in the x_1-x_2 plane for which this equation is satisfied is O_+. Similar reasoning yields as the expression for O_-

$$O_- = \left\{x_1(t), x_2(t): x_1(t) = \frac{1}{a^2}\ln\left(a\left[x_2(t) + \frac{1}{a}\right]\right) - \frac{1}{a}x_2(t)\right\}.\ddagger \tag{5.4-54}$$

Since Eq. (5.4-53) applies for $x_2(t) < 0$ and (5.4-54) applies for $x_2(t) > 0$, the expression for O_1 (the set of all points that are in either O_+ or O_-) is given by

$$O_1 = \left\{x_1(t), x_2(t): x_1(t) = \frac{x_2(t)}{|x_2(t)|}\frac{1}{a^2}\ln\left(a\left[|x_2(t)| + \frac{1}{a}\right]\right) - \frac{1}{a}x_2(t)\right\}. \tag{5.4-55}$$

The switching function is then

$$s(\mathbf{x}(t)) = x_1(t) - \frac{x_2(t)}{|x_2(t)|}\frac{1}{a^2}\ln\left(a\left[|x_2(t)| + \frac{1}{a}\right]\right) + \frac{1}{a}x_2(t). \tag{5.4-56}$$

The switching curves for $a = 0.5$, 1.0, and 2.0 are shown in Fig. 5-23, and some typical trajectories for $a = 0.5$ are shown in Fig. 5-24. It is left as an exercise for the reader to verify that for points *above* the switching curve the optimal control is $u^* = -1$ until reaching the switching curve, where u^* switches to $+1$, and remains at $+1$ until the origin is reached, at which time $u^* = 0$ is applied to keep the system at the origin. Similar reasoning gives the optimal control law for points below the switching curve. In summary, the optimal control law is

† ln denotes the natural logarithm, or $\log_\epsilon$.

‡ This notation means that O_- is the set of points that satisfy the equation

$$x_1(t) = \frac{1}{a^2}\ln\left(a\left[x_2(t) + \frac{1}{a}\right]\right) - \frac{1}{a}x_2(t).$$

$$u^*(t) = \begin{cases} -1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) > 0 \\ +1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) < 0 \\ -1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) = 0 \text{ and } x_2(t) > 0 \\ +1, & \text{for } \mathbf{x}(t) \text{ such that } s(\mathbf{x}(t)) = 0 \text{ and } x_2(t) < 0 \\ 0, & \text{for } \mathbf{x}(t) = \mathbf{0}. \end{cases} \quad (5.4\text{-}57)$$

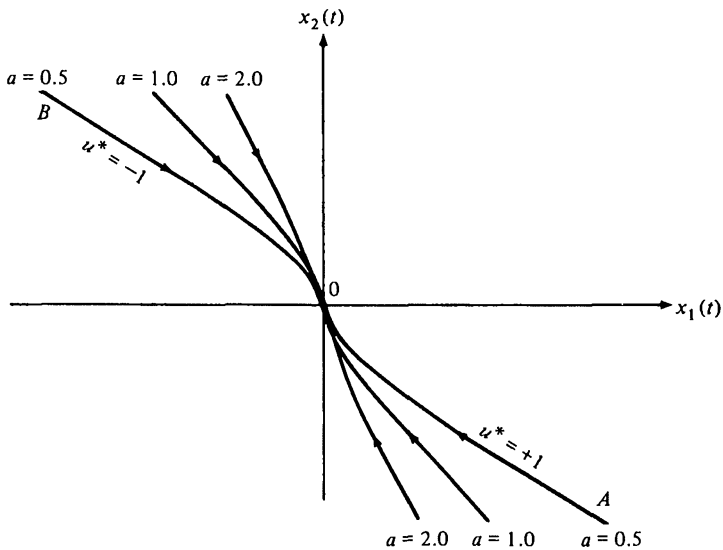


Figure 5-23 Time-optimal switching curves for Example 5.4-5 with $a = 0.5, 1.0, 2.0$

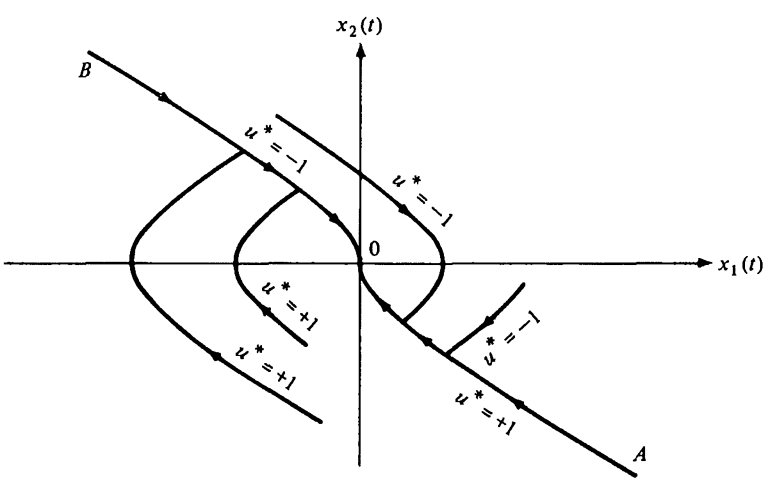


Figure 5-24 Several optimal trajectories for Example 5.4-5 with $a = 0.5$

Summary

In this section we have found that time-optimal controls for a rather general class of systems are "bang-bang"; that is, the optimal control switches between its maximum and minimum admissible values.

A procedure for finding time-optimal control laws for time-invariant, linear regulator systems was discussed and demonstrated for two second-order systems. Although this procedure is conceptually straightforward, it does have serious limitations:

1. For higher-order systems ($n \geq 3$) it is generally difficult, if not impossible, to obtain an analytical expression for the switching hypersurface.
2. Even in cases where an expression for the switching hypersurface can be found, physical implementation of the optimal control law may be quite complicated, indicating that a suboptimal, but easier-to-implement, control law may be preferable.
3. The procedure is generally not applicable to nonlinear systems, because of the difficulty of analytically integrating the differential equations.

5.5 MINIMUM CONTROL-EFFORT PROBLEMS

In the preceding section we considered problems in which the objective was to transfer a system from an arbitrary initial state to a specific target set as quickly as possible. Let us now consider problems in which control effort required, rather than elapsed time, is the criterion of optimality. Such problems arise frequently in aerospace applications, where often there are limited control resources available for achieving desired objectives.

The class of problems we will discuss is the following: Find a control $\mathbf{u}^*(t)$ satisfying constraints of the form

$$M_{i-} \leq u_i(t) \leq M_{i+}, \qquad i = 1, 2, \ldots, m, \tag{5.5-1}$$

which transfers a system described by

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{5.5-2}$$

from an arbitrary initial state $\mathbf{x}_0$ to a specified target set $S(t)$ with a minimum expenditure of control effort.

As measures of control effort we shall consider the two performance indices

$$J_1(\mathbf{u}) = \int_{t_0}^{t_f} \left[\sum_{i}^{m} \beta_i |u_i(t)| \right] dt \tag{5.5-3}$$

and

$$J_2(\mathbf{u}) = \int_{t_0}^{t_f} \left[\sum_{i=1}^{m} r_i u_i^2(t) \right] dt, \tag{5.5-4}$$

where β_i and r_i, $i = 1, \ldots, m$, are nonnegative weighting factors. As discussed in Chapter 2, the fuel consumed by a mass-expulsion thrusting system is often expressed by an integral of the form (5.5-3); thus, if a performance measure to be minimized has the form given by J_1, we shall refer to the problem as a *minimum-fuel problem.* The total electrical energy supplied to a network of resistors by several voltage and current sources is given by an integral of the form (5.5-4); hence, if a performance measure of this form is to be minimized, we shall say that we wish to solve a *minimum-energy problem.* The reader must be cautioned that in a particular problem (5.5-3) may not represent fuel expenditure, or control energy required may not be given by (5.5-4); therefore, the results obtained in this section will apply to the performance measure J_1 or J_2, not necessarily to the problems of minimizing fuel or energy consumption.

Our discussion will be primarily devoted to solving several example problems that are rather elementary, but nonetheless indicative of the characteristics of fuel and energy-optimal systems.†

Minimum-Fuel Problems

In our discussion of minimum-time problems in Section 5.4 the concept of reachable states was introduced. Recall that $R(t)$ was used to denote the set of states that can be reached at time t by starting from an initial state $\mathbf{x}_0$ at time t_0. Minimum-fuel problems may also be visualized in terms of reachable states; that is, the minimum-fuel solution is given by the intersection of the target set $S(t)$ with the set of reachable states $R(t)$, which requires the *smallest amount of consumed fuel.* To represent this idea geometrically we could use a state–time–consumed-fuel coordinate system and determine the intersections (if any) of $S(t)$ and $R(t)$. Unfortunately, although such a geometric representation is helpful as a conceptual device, it is of limited value in actually obtaining solutions. Instead of pursuing this avenue further, we shall approach minimum control-effort problems by starting with the necessary conditions provided by Pontryagin's minimum principle.

The Form of the Optimal Control for a Class of Minimum-Fuel Problems. Let us assume that the state equations of a system are of the form

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), t) + \mathbf{B}(\mathbf{x}(t), t)\mathbf{u}(t), \tag{5.5-5}$$

† For additional reading on fuel- and energy-optimal systems see [A-2], [L-3], and [L-4].

where $\mathbf{B}$ is an $n \times m$ array that may be explicitly dependent on the states and time. The performance measure to be minimized is

$$J(\mathbf{u}) = \int_{t_0}^{t_f} \left[\sum_{i=1}^{m} |u_i(t)| \right] dt, \tag{5.5-6}$$

and the admissible controls are to satisfy the constraints

$$-1 \leq u_i(t) \leq +1, \qquad i = 1, 2, \ldots, m, \qquad t \in [t_0, t_f].\dagger \tag{5.5-7}$$

The Hamiltonian is

$$\begin{aligned} \mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) &= \sum_{i=1}^{m} |u_i(t)| + \mathbf{p}^T(t)\mathbf{a}(\mathbf{x}(t), t) \\ &\quad + \mathbf{p}^T(t)\mathbf{B}(\mathbf{x}(t), t)\mathbf{u}(t), \end{aligned} \tag{5.5-8}$$

and the minimum principle requires that

$$\begin{aligned} &\sum_{i=1}^{m} |u_i^*(t)| + \mathbf{p}^{*T}(t)\mathbf{a}(\mathbf{x}^*(t), t) + \mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}^*(t) \\ &\leq \sum_{i=1}^{m} |u_i(t)| + \mathbf{p}^{*T}(t)\mathbf{a}(\mathbf{x}^*(t), t) + \mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}(t), \end{aligned} \tag{5.5-9}$$

or

$$\sum_{i=1}^{m} |u_i^*(t)| + \mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}^*(t) \leq \sum_{i=1}^{m} |u_i(t)| + \mathbf{p}^{*T}(t)\mathbf{B}(\mathbf{x}^*(t), t)\mathbf{u}(t) \tag{5.5-10}$$

for all admissible $\mathbf{u}(t)$, and for all $t \in [t_0, t_f]$. As in Section 5.4 let us express $\mathbf{B}$ in the form

$$\mathbf{B}(\mathbf{x}^*(t), t) = [\mathbf{b}_1(\mathbf{x}^*(t), t) \,\vdots\, \mathbf{b}_2(\mathbf{x}^*(t), t) \,\vdots\, \cdots \,\vdots\, \mathbf{b}_m(\mathbf{x}^*(t), t)],$$

where $\mathbf{b}_i(\mathbf{x}^*(t), t)$ is the ith column of the $n \times m$-dimensional $\mathbf{B}$ array. Assuming that the components of $\mathbf{u}$ are independent of one another, we have from (5.5-10) that

$$\begin{aligned} &|u_i^*(t)| + \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)u_i^*(t) \\ &\quad \leq |u_i(t)| + \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)u_i(t), \qquad i = 1, 2, \ldots, m. \end{aligned} \tag{5.5-11}$$

The definition of $|u_i(t)|$ is

† For simplicity we have assumed that $M_{i-} = -1$, $M_{i+} = +1$, and $\beta_i = 1$ for $i = 1, 2, \ldots, m$. The derivation is easily modified if these assumptions are not made.

$$|u_i(t)| \triangleq \begin{cases} u_i(t), & \text{for } u_i(t) \geq 0 \\ -u_i(t), & \text{for } u_i(t) \leq 0; \end{cases} \tag{5.5-12}$$

therefore,

$$\begin{aligned} &|u_i(t)| + \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)u_i(t) \\ &\quad = \begin{cases} [1 + \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)]u_i(t), & \text{for } u_i(t) \geq 0 \qquad \text{(5.5-13a)} \\ [-1 + \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)]u_i(t), & \text{for } u_i(t) \leq 0. \qquad \text{(5.5-13b)} \end{cases} \end{aligned}$$

If $\mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) > 1.0$, the minimum value of expression (5.5-13a) is 0, because $u_i(t) \geq 0$; the minimum value of (5.5-13b) is attained for $u_i(t) = -1$ and is equal to $[+1 - \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t)] < 0$.

If $\mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) = 1.0$, (5.5-13a) can be made equal to 0 by selecting $u_i(t) = 0$; on the other hand, (5.5-13b) will be 0 for all $u_i(t) \leq 0$; therefore, any nonpositive $u_i(t)$ will minimize (5.5-13).†

If $0 \leq \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) < 1.0$, the minimum values of both (5.5-13a) and (5.5-13b) are zero and are attained for $u_i(t) = 0$.

The same reasoning is used for $\mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) < 0$. In summary, the form of the optimal control is

$$u_i^*(t) = \begin{cases} 1.0, & \text{for } \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) < -1.0 \\ 0, & \text{for } -1.0 < \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) < 1.0 \\ -1.0, & \text{for } 1.0 < \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) \\ \text{an undetermined nonnegative value} & \text{if } \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) = -1.0 \\ \text{an undetermined nonpositive value} & \text{if } \mathbf{p}^{*T}(t)\mathbf{b}_i(\mathbf{x}^*(t), t) = +1.0. \end{cases} \tag{5.5-14}$$

Figure 5-25 illustrates the dependence of the optimal control on its coefficient in the Hamiltonian. Notice that whereas in minimum-time problems the optimal control is "bang-bang" (see Fig. 5-18) the minimum-fuel control may be described as "bang-off-bang" (if we assume no singular intervals).

In the remainder of this section we shall consider problems in which the plant dynamics are linear.

Free Final Time. Let us now consider some examples of linear minimum-fuel problems in which the final time t_f is not specified.

Example 5.5-1. The system

$$\dot{x}(t) = u(t) \tag{5.5-15}$$

is to be transferred from an arbitrary initial state x_0 to the origin. The performance measure to be minimized is

† If $\mathbf{p}^{*T}\mathbf{b}_i = \pm 1$ for a nonzero time interval, a singular solution exists; otherwise, a control switching is indicated.

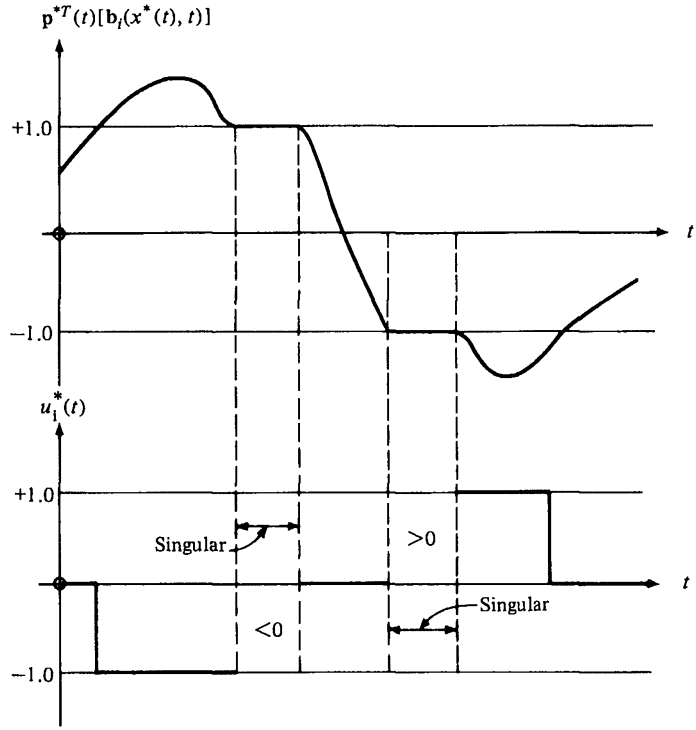


Figure 5-25 The relationship between a fuel-optimal control and its coefficient in the Hamiltonian

$$J(u) = \int_0^{t_f} |u(t)|\, dt, \tag{5.5-16}$$

where t_f is free, and the admissible controls satisfy

$$|u(t)| \leq 1.0. \tag{5.5-17}$$

It is desired to determine the optimal control law.

From (5.5-15) and (5.5-16) the Hamiltonian is

$$\mathscr{H}(x(t), u(t), p(t)) = |u(t)| + p(t)u(t). \tag{5.5-18}$$

The costate equation

$$\dot{p}^*(t) = -\frac{\partial \mathscr{H}}{\partial x} = 0 \tag{5.5-19}$$

has a solution of the form

$$p^*(t) = c_1, \tag{5.5-20}$$

where c_1 is a constant.

From Eq. (5.5-14) with $\mathbf{b}_i = \mathbf{B} = 1$, we have

$$u^*(t) = \begin{cases} 1.0, & \text{for } p^*(t) = c_1 < -1.0 \\ 0, & \text{for } -1.0 < c_1 < 1.0 \\ -1.0, & \text{for } 1.0 < c_1 \\ \text{an undetermined nonnegative value if } c_1 = -1.0 \\ \text{an undetermined nonpositive value if } c_1 = +1.0. \end{cases} \tag{5.5-21}$$

The solution of the state equation is

$$x(t) = x_0 + \int_0^t u(t)\, dt; \tag{5.5-22}$$

thus, for $x(t_f) = 0$

$$0 = x_0 + \int_0^{t_f} u(t)\, dt, \tag{5.5-23}$$

or

$$x_0 = -\int_0^{t_f} u(t)\, dt. \tag{5.5-24}$$

Clearly, from (5.5-24) the control $u(t) = 0$, $t \in [0, t_f]$ can be optimal only if $x_0 = 0$—a trivial case. Suppose that $x_0 = 5.0$; then each of the controls

$$\begin{aligned} u(t) &= -1, && t \in [0, 5] \\ u(t) &= -0.5 && t \in [0, 10] \\ u(t) &= -0.2, && t \in [0, 25] \\ u(t) &= -0.1, && t \in [0, 50] \\ u(t) &= \begin{cases} -1, & t \in [0, 2] \\ -0.5 & t \in (2, 8] \end{cases} \end{aligned} \tag{5.5-25}$$

satisfies (5.5-24) and each makes $J = 5.0$. Now suppose we calculate a lower limit on the fuel required to force this system from x_0 to the origin. From (5.5-24)

$$|x_0| = \left| \int_0^{t_f} u(t)\, dt \right| \leq \int_0^{t_f} |u(t)|\, dt = J. \tag{5.5-26}$$

But each of the controls of (5.5-25) satisfies $J = |x_0|$; therefore, each of these controls is optimal. In this example, the optimal controls are nonunique. Notice, however, that the optimal controls of Eq. (5.5-25) each require a different amount of time to transfer the system to the origin.

In the preceding example there were many optimal controls (an infinite number); let us now consider an example in which an optimal control does not exist.

Example 5.5-2. It is desired to transfer the system

$$\dot{x}(t) = -ax(t) + u(t) \tag{5.5-27}$$

from an arbitrary initial state x_0 to the origin with admissible controls satisfying

$$|u(t)| \leq 1, \tag{5.5-28}$$

and $a > 0$.

The performance measure to be minimized is

$$J(u) = \int_0^{t_f} |u(t)|\, dt, \tag{5.5-29}$$

where t_f is free.

Using the state equation and the performance measure, we find that the Hamiltonian is

$$\mathscr{H}(x(t), u(t), p(t)) = |u(t)| - p(t)ax(t) + p(t)u(t); \tag{5.5-30}$$

thus, the costate equation is

$$\dot{p}^*(t) = -\frac{\partial \mathscr{H}}{\partial x} = ap^*(t), \tag{5.5-31}$$

which implies that

$$p^*(t) = c_1 \epsilon^{at}, \tag{5.5-32}$$

where c_1 is a constant of integration.

From Eq. (5.5-14) with $\mathbf{b}_i = \mathbf{B} = 1$, the form of the optimal control is

$$u^*(t) = \begin{cases} +1.0, & \text{for } p^*(t) < -1.0 \\ 0, & \text{for } -1.0 < p^*(t) < 1.0 \\ -1.0, & \text{for } 1.0 < p^*(t). \end{cases} \tag{5.5-33}$$

Notice that when $|p^*(t)|$ passes through the value 1.0, a switching of the control is indicated. Another possibility is that $|p^*(t)|$ might remain equal to 1.0 for a finite time interval; however, since

$$p^*(t) = c_1 \epsilon^{at}$$

and $a > 0$, it is clear that this situation cannot occur. It should be emphasized that the foregoing development tacitly assumes that an optimal control exists; we shall test the validity of this assumption shortly.

$p^*(t)$ will be one of the five forms shown in Fig. 5-26, depending on the value of c_1. The optimal controls, given by Eq. (5.5-33), which correspond to Fig. 5-26 are

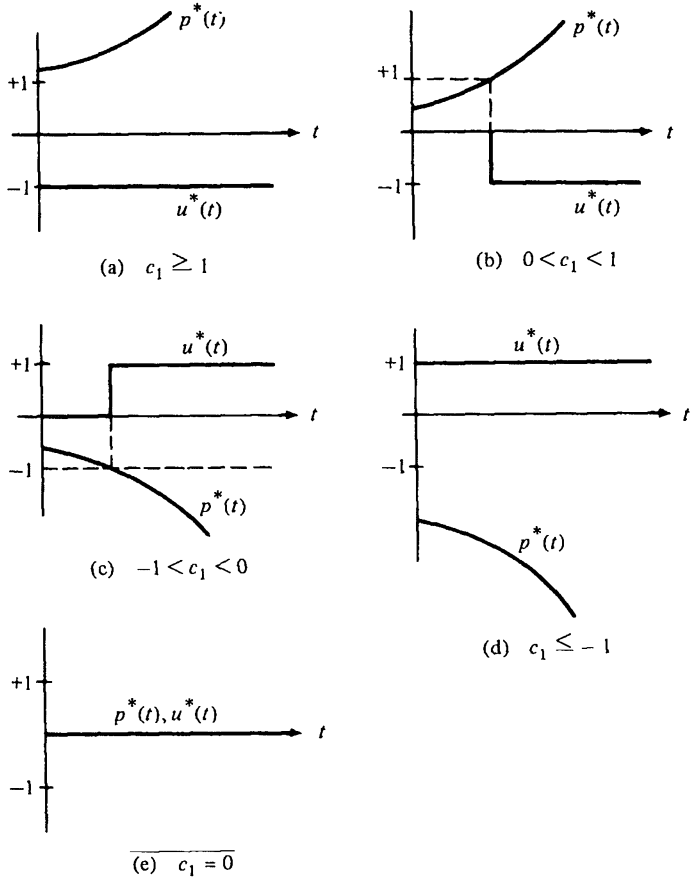


Figure 5-26 Possible forms for the costate and the corresponding fuel-optimal controls

$$
\begin{aligned}
u^*(t) &= -1, && t \in [0, t_f], && \text{for } 1 \leq c_1 \\
u^*(t) &= \begin{cases} 0, & t \in [0, t_1) \\ -1, & t \in [t_1, t_f], \end{cases} && && \text{for } 0 < c_1 < 1 \\
u^*(t) &= \begin{cases} 0, & t \in [0, t_1') \\ +1, & t \in [t_1', t_f], \end{cases} && && \text{for } -1 < c_1 < 0 \qquad (5.5\text{-}34) \\
u^*(t) &= +1, && t \in [0, t_f], && \text{for } c_1 \leq -1 \\
u^*(t) &= \ \ 0, && t \in [0, t_f], && \text{for } c_1 = 0.
\end{aligned}
$$

We shall denote these five forms by $u^* = \{-1\}$, $\{0, -1\}$, $\{0, +1\}$, $\{+1\}$, and $\{0\}$, respectively.

The solution of the state equation is

$$x(t) = \epsilon^{-at}x_0 + \epsilon^{-at}\int_0^t \epsilon^{a\tau}u(\tau)\,d\tau. \tag{5.5-35}$$

Notice that if the control is identically zero, then at $t = t_f$

$$x(t_f) = \epsilon^{-at_f}x_0. \tag{5.5-36}$$

Since the system is stable, it naturally moves toward zero when no control is applied. If we are willing to wait long enough, the system will come arbitrarily close to (but never precisely reach) zero—and without the expenditure of any control effort at all. However, the problem statement stipulated that $x(t_f) = 0$, not $|x(t_f)| < \eta$, where η is some arbitrarily small positive number. If $x_0 > 0$, then clearly $u^* = \{-1\}, \{0, -1\}$ are the only possible choices for the optimal control (why?). If $u(t) = -1$ for $t \in [0, t_f]$, it can be shown from (5.5-35) that $x(t_f) = 0$ implies

$$t_f = \frac{1}{a}\ln(ax_0 + 1); \tag{5.5-37}$$

thus, the fuel consumption using this control would be $[\ln(ax_0 + 1)]/a$.

Now, suppose $u(t) = 0$ is applied for $0 \leq t < t_1$ and $u(t) = -1$ for $t_1 \leq t \leq t_f$. From (5.5-35)

$$x(t_f) = \epsilon^{-at_f}x_0 + \epsilon^{-at_f}\int_{t_1}^{t_f} \epsilon^{a\tau}[-1]\,d\tau; \tag{5.5-38}$$

setting $x(t_f) = 0$ and performing the indicated integration, we obtain

$$0 = \epsilon^{-at_f}x_0 - \frac{1}{a}[1 - \epsilon^{-a[t_f - t_1]}]. \tag{5.5-39}$$

Solving for $t_f - t_1$ gives

$$t_f - t_1 = -\frac{1}{a}\ln(1 - ax_0\epsilon^{-at_f}), \tag{5.5-40}$$

but since t_f is free, $ax_0\epsilon^{-at_f}$ can be made arbitrarily small by letting $t_f \longrightarrow \infty$, so

$$[t_f - t_1] \longrightarrow 0 \quad \text{as} \quad t_f \longrightarrow \infty. \tag{5.5-41}$$

But $t_f - t_1$ is the interval during which $u = -1$ is applied, and by making t_f very large the consumed fuel can be made arbitrarily small (but not zero). Our conclusion is that if t_f is free an *optimal control does not exist*—given any candidate for an optimal control, it is always possible to find a control that transfers the system to the zero state with less fuel.

It is left as an exercise for the reader to verify that the same conclusions hold when $x_0 < 0$.

In the preceding example we have simply verified mathematically what common sense tells us; if elapsed time is not penalized, and the system moves toward the desired final state without consuming any fuel, the optimal strategy is to let the system drift as long as possible before any control is applied. At this point the reader might wonder: what did the minimum principle do for us? Could we not have deduced the same conclusions without using it at all? The answer to these questions is that quite likely the same conclusions could have been reached by intuitive reasoning alone, but the minimum principle, by specifying the possible forms of the optimal control, greatly reduced the number of control histories that had to be examined. In addition, we must remember that our interest is in solving problems that generally require more than physical reasoning and common sense.

Let us next discuss minimum-fuel problems with fixed final times.

Fixed Final Time. First, let us reconsider the preceding examples with the final time specified; that is, $t_f = T$. The value of T must be at least as large as t^*, the minimum time required to reach the specified target set from the initial state x_0.

In Example 5.5-1 we found that the optimal control was nonunique—there were an infinite number of controls that would transfer the system to $x(t_f) = 0$ with the minimum possible amount of fuel. The situation with $t_f = T$ is much the same unless $T = t^*$. In this case, the minimum-fuel and minimum-time controls are the same and unique. If, however, $T > t^*$, there are again an infinite number of controls that are optimal; it is left as an exercise for the reader to verify that this is the case. Fixing the final time does not alter the nonuniqueness of the optimal controls for the system of Example 5.5-1.

Let us now see if fixing the final time has any effect on the existence of fuel-optimal controls for the system of Example 5.5-2.

Example 5.5-3. The possible forms for optimal controls and the solution of the state equation are given in (5.5-34) and (5.5-35). If the fixed final time T is equal to the minimum time t^* required to reach the origin from the initial state x_0, then $u^*(t)$ is either $+1$ or -1 throughout the entire interval $[0, T]$, and

$$x(T) = 0 = \epsilon^{-aT}x_0 + \epsilon^{-aT}\int_0^T \epsilon^{a\tau}[\pm 1]\,d\tau, \tag{5.5-42}$$

or

$$x_0 = \mp\frac{1}{a}[\epsilon^{aT} - 1]. \tag{5.5-42a}$$

This expression defines the largest and smallest values of x_0 from which the origin can be reached in a (specified) time T. Initial states that satisfy

$$\frac{1}{a}[\epsilon^{aT} - 1] < |x_0| \tag{5.5-43}$$

cannot be transferred to the origin in time T; therefore, we shall assume in what follows that

$$|x_0| \leq \frac{1}{a}[\epsilon^{aT} - 1]. \tag{5.5-44}$$

If (5.5-44) is an equality, this means that $T = t^*$; otherwise, $T > t^*$, and the form of the optimal control must be as shown in Fig. 5-26(b) or (c). The optimal control must be nonzero during some part of the time interval, because we have previously shown that the system will not reach the origin in the absence of control.

If $x_0 > 0$, the optimal control must have the form $u^* = \{0, -1\}$ shown in Fig. 5-26(b). Substituting $u(t) = 0$, $t \in [0, t_1)$, $u(t) = -1$, $t \in [t_1, T]$, in (5.5-35) and performing the integration, we obtain

$$x(T) = 0 = \epsilon^{-aT}x_0 - \frac{1}{a}\epsilon^{-aT}[\epsilon^{aT} - \epsilon^{at_1}]. \tag{5.5-45}$$

Solving this equation for t_1, the time when the control switches from 0 to -1, gives

$$t_1 = \frac{1}{a}\ln(\epsilon^{aT} - ax_0). \tag{5.5-46}$$

Similarly, if $x_0 < 0$, the optimal control has the form $u^* = \{0, +1\}$ shown in Fig. 5-26(c), and

$$x(T) = 0 = \epsilon^{-aT}x_0 + \frac{1}{a}\epsilon^{-aT}[\epsilon^{aT} - \epsilon^{at_1'}]. \tag{5.5-47}$$

Solving for the switching time t_1' yields

$$t_1' = \frac{1}{a}\ln(\epsilon^{aT} + ax_0). \tag{5.5-48}$$

From (5.5-46) and (5.5-48) the optimal control is

$$u^*(t) = \begin{cases} 0, & \text{for } x_0 > 0 \text{ and } t < \frac{1}{a}\ln(\epsilon^{aT} - ax_0) \\ -1, & \text{for } x_0 > 0 \text{ and } \frac{1}{a}\ln(\epsilon^{aT} - ax_0) \leq t \leq T \\ 0, & \text{for } x_0 < 0 \text{ and } t < \frac{1}{a}\ln(\epsilon^{aT} + ax_0) \\ +1, & \text{for } x_0 < 0 \text{ and } \frac{1}{a}\ln(\epsilon^{aT} + ax_0) \leq t \leq T. \end{cases} \tag{5.5-49}$$

Notice that the optimal control expressed by (5.5-49) is in open-loop form, because $u^*(t)$ has been expressed in terms of x_0 and t; that is,

$$u^*(t) = e(x_0, t). \tag{5.5-50}$$

From an engineering point of view we would prefer to have the optimal control in feedback form; that is,

$$u^*(t) = f(x(t), t). \tag{5.5-51}$$

To obtain the optimal control law, we observe that

$$x(T) = \epsilon^{-a[T-t]}x(t) + \epsilon^{-aT}\int_t^T \epsilon^{a\tau}u(\tau)\,d\tau \tag{5.5-52}$$

for all t. We know that during the last part of the time interval the control is either $+1$ or -1, depending on whether $x(t)$ is less than zero or greater than zero; thus, assuming $x(t) > 0$, we have

$$x(T) = 0 = \epsilon^{-a[T-t]}x(t) - \epsilon^{-aT}\int_t^T \epsilon^{a\tau}\,d\tau, \qquad t \geq t_1. \tag{5.5-53}$$

Performing the indicated integration and solving for $x(t)$ gives

$$x(t) = \frac{1}{a}[\epsilon^{a[T-t]} - 1], \qquad t \geq t_1. \tag{5.5-54}$$

During the initial part of the time interval, the optimal control is zero; consequently,

$$x(t) = \epsilon^{-at}x_0, \qquad t < t_1. \tag{5.5-55}$$

The switching of the control from 0 to -1 occurs when the solution (5.5-55) for the *coasting interval* ($u = 0$) intersects the solution (5.5-54) for the *on*-negative *interval* ($u = -1$). Figure 5-27 shows these solutions. Defining

$$z(T-t) \triangleq \frac{1}{a}[\epsilon^{a[T-t]} - 1], \tag{5.5-56}$$

we observe that the control should switch from 0 to -1 when the state $x(t)$ is equal to $z(T-t)$. It is left as an exercise for the reader to verify that if $-[\epsilon^{aT} - 1]/a < x_0 < 0$ the optimal control switches from 0 to $+1$ when

$$x(t) = -z(T-t). \tag{5.5-57}$$

To summarize, the optimal control law is

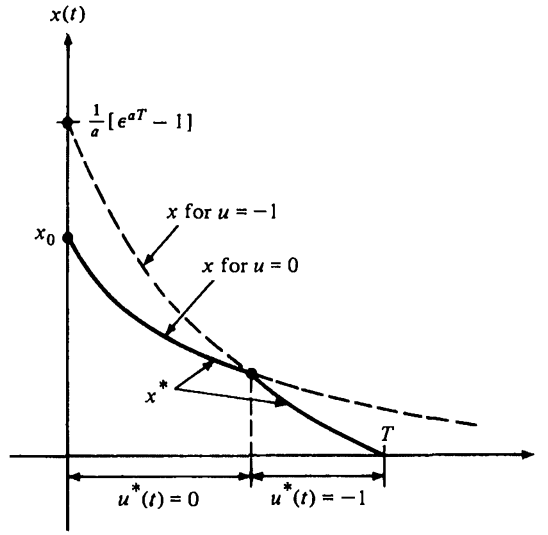


Figure 5-27 The two segments of a fuel-optimal trajectory $0 < x_0 < [e^{aT} - 1]/a$

$$u^*(t) = \begin{cases} -1, & \text{for } x(t) \geq z(T-t) \\ 0, & \text{for } |x(t)| < z(T-t) \\ +1, & \text{for } x(t) \leq -z(T-t) \end{cases} \tag{5.5-58}$$

or, more compactly,

$$u^*(t) = \begin{cases} 0, & \text{for } |x(t)| < z(T-t) \\ -\text{sgn}\,(x(t)), & \text{for } |x(t)| \geq z(T-t). \dagger \end{cases} \tag{5.5-58a}$$

An implementation of this optimal control law is shown in Fig. 5-28. The logic element shown controls the switch. Notice that the controller requires a clock to tell it the current value of the time—the control law is time-varying. Naturally, this complicates the implementation; a time-invariant control law would be preferable.

Selecting the Final Time. In the preceding example fixing the final time led to a time-varying control law. We next ask: "How is the final time specified?" To answer this question, let us see how the minimum fuel required depends on the value of the final time T. Equations (5.5-46) and (5.5-48) indicate that the control switches from 0 at

† Here we define sgn (0) $\triangleq$ 0.

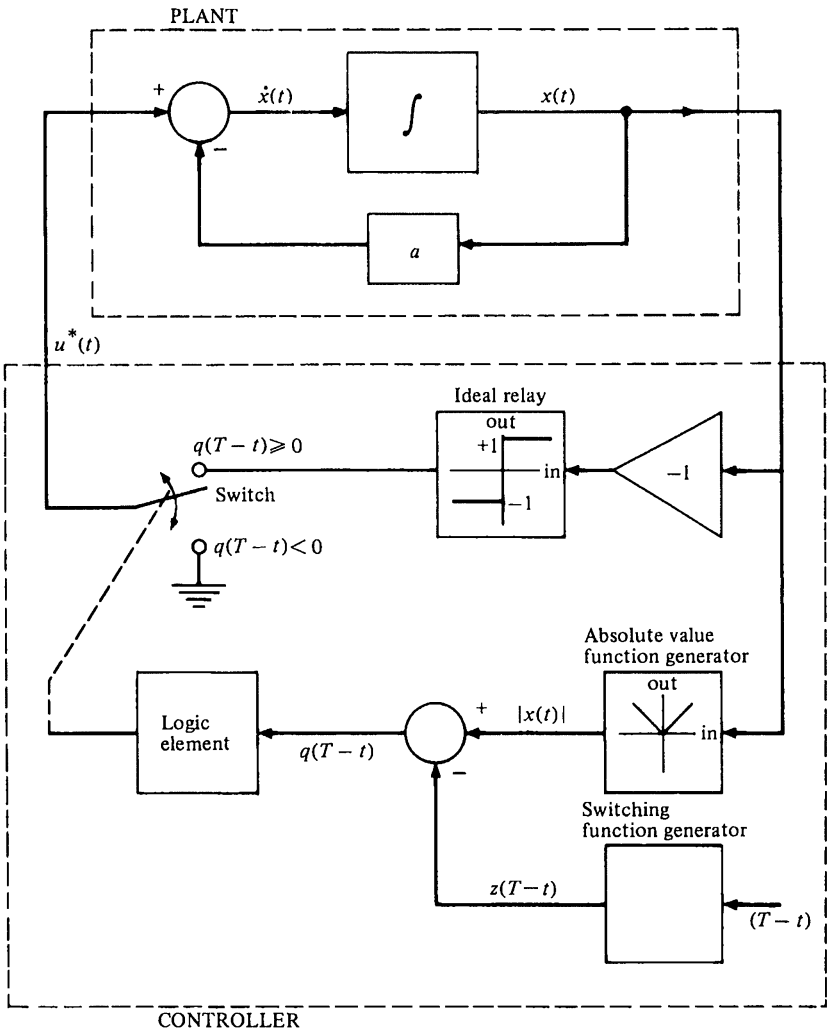


Figure 5-28 Implementation of a time-varying fuel-optimal control law.

$$t_1 = \frac{1}{a} \ln (\epsilon^{aT} - a\,|\,x_0\,|) \tag{5.5-59}$$

and remains at ± 1 until the final time is reached; thus, the fuel consumed is

$$T - t_1 = T - \frac{1}{a} \ln (\epsilon^{aT} - a\,|\,x_0\,|). \tag{5.5-60}$$

Using Eq. (5.5-60), the designer can obtain a plot of consumed fuel versus final time for several values of x_0 selected from the range of expected initial conditions; one such curve is shown in Fig. 5-29 for $|x_0| = 10.0$ and $a = 1.0$. The selection of T is then made by *subjectively* evaluating the information contained in these curves. Figure 5-29 indicates that in this particular example the value chosen for T will reflect the relative importance of consumed fuel and elapsed time.

The reader may have noticed that in Examples 5.5-1 through 5.5-3 a "trade-off" existed between fuel expenditure and elapsed time. The reason for this is that in each case the target set was the origin, and with no control applied the state of these systems either moved closer to the origin (Examples 5.5-2 and 5.5-3) or remained constant (Example 5.5-1). If the plants were of such a form that the states moved away from the target set with no control applied, the solutions obtained could have been quite different—see Problem 5-28.

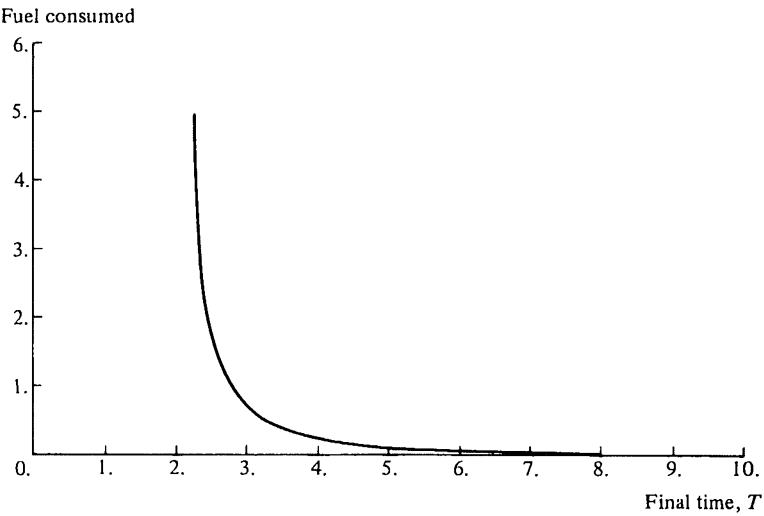


Figure 5-29 Dependence of consumed fuel on specified final time T, $|x_0| = 10$

A Weighted Combination of Elapsed Time and Consumed Fuel as the Performance Measure

The preceding examples in this section illustrated a trade-off between elapsed response time and consumed fuel; that is, the fuel expended to accomplish a specified state transfer was inversely proportional to the time required for the transfer. One technique for handling problems in which this

trade-off is present is to include both elapsed time and consumed fuel in the performance measure. For a system with one control, such a performance measure would have the form

$$J(u) = \int_{t_0}^{t_f} [\lambda + |u(t)|]\, dt. \tag{5.5-61}$$

The final time t_f is free, and $\lambda > 0$ is chosen to weight the relative importance of elapsed time and fuel expended. For $\lambda \to 0$ the optimal system will resemble a free-final-time, fuel-optimal system, whereas for $\lambda \to \infty$ the optimal solution will resemble a time-optimal solution. Let us now reconsider Example 5.5-2 with (5.5-61) as the performance measure.

Example 5.5-4. The state equation and control constraint are given in Eqs. (5.5-27) and (5.5-28). The Hamiltonian is

$$\mathscr{H}(x(t), u(t), p(t)) = \lambda + |u(t)| - p(t)ax(t) + p(t)u(t), \tag{5.5-62}$$

and the costate equation is (again)

$$\dot{p}^*(t) = ap^*(t); \tag{5.5-63}$$

thus,

$$p^*(t) = c_1 \epsilon^{at}, \tag{5.5-64}$$

where c_1 is a constant of integration. The requirement that $u^*(t)$ minimize the Hamiltonian on an extremal trajectory is unaffected by the presence of λ in the performance measure; therefore,

$$u^*(t) = \begin{cases} 1.0, & \text{for } p^*(t) < -1.0 \\ 0, & \text{for } -1.0 < p^*(t) < 1.0 \\ -1.0, & \text{for } p^*(t) > 1.0 \\ \text{undetermined, but nonnegative for } p^*(t) = -1.0\dagger \\ \text{undetermined, but nonpositive for } p^*(t) = +1.0.\dagger \end{cases} \tag{5.5-65}$$

If we recall that $a > 0$, Eq. (5.5-64) ensures that $p^*(t)$ cannot equal ± 1.0 for a nonzero time interval; hence, there are no intervals of singular control.

Equations (5.5-64) and (5.5-65) indicate that the optimal control must again be one of the forms shown in Fig. 5-26. Let us now examine the various alternatives.

Suppose that $t_0 = 0$ and $u^*(t) = 0$, $t \in [0, t_f]$; this implies that

$$\mathscr{H}(x^*(t), 0, p^*(t)) = \lambda - p^*(t)ax^*(t) \qquad \text{for all } t \in [0, t_f]. \tag{5.5-66}$$

† If $p^*(t) = \pm 1.0$ for a nonzero time interval, this signals the singular condition.

In this problem the final time is free and t does not appear explicitly in the Hamiltonian; therefore, from Eq. (5.3-41),

$$\mathscr{H}(x^*(t), u^*(t), p^*(t)) = 0 \qquad \text{for all } t \in [0, t_f]. \tag{5.5-67}$$

If Eq. (5.5-67) is to be satisfied, then Eq. (5.5-66) implies that

$$x^*(t) = \frac{\lambda}{ap^*(t)}, \tag{5.5-68}$$

or

$$x^*(t) = \frac{\lambda}{ac_1\epsilon^{at}} \qquad \text{for all } t \in [0, t_f]. \tag{5.5-68a}$$

Since $x^*(t_f) = 0$, Eq. (5.5-68a) can be satisfied for $\lambda > 0$ at t_f only if $t_f \longrightarrow \infty$, but this implies that the minimum cost approaches ∞. From our earlier discussion of this example, however, we know that controls can be found for which $J < \infty$; therefore, we conclude that $u(t) = 0$, $t \in [0, t_f]$, cannot be an optimal control.

If $u^* = \{0, -1\}$ is the form of the optimal control, $p^*(t)$ must pass through the value $+1.0$ at the time t_1, when the control switches [see Eq. (5.5-65)]. In addition, we know from Eq. (5.5-65) that $u^*(t_1)$ is some nonpositive value, so $|u^*(t_1)| = -u^*(t_1)$. The Hamiltonian must be zero for all t; thus, at time t_1

$$\mathscr{H}(x^*(t_1), u^*(t_1), p^*(t_1)) = \lambda - u^*(t_1) - ax^*(t_1) + u^*(t_1) = 0, \tag{5.5-69}$$

which implies that

$$x^*(t_1) = \frac{\lambda}{a}. \tag{5.5-70}$$

This equation is an important result, for it indicates that if there is a switching of control from 0 to -1 it occurs when $x^*(t)$ passes through the value λ/a. From Eq. (5.5-35)—the solution of the state equations—and Eq. (5.5-70) we obtain the family of optimal trajectories

$$x(t) = x_0\epsilon^{-at} \qquad \text{for } x(t) > \frac{\lambda}{a} \tag{5.5-71a}$$

$$x(t) = \frac{\lambda}{a}\epsilon^{-a[t-t_1]} - \frac{1}{a}[1 - \epsilon^{-a[t-t_1]}] \qquad \text{for } 0 < x(t) \leq \frac{\lambda}{a}. \tag{5.5-71b}$$

A control of the form $\{0, -1\}$ cannot transfer the system $\dot{x}(t) = -ax(t) + u(t)$ from a negative initial state value to the origin; hence, Eq. (5.5-71) applies for $x_0 > 0$.

Optimal trajectories for several different values of x_0 are shown in Fig. 5-30. Notice that if $0 < x_0 \leq \lambda/a$, the optimal strategy is to apply

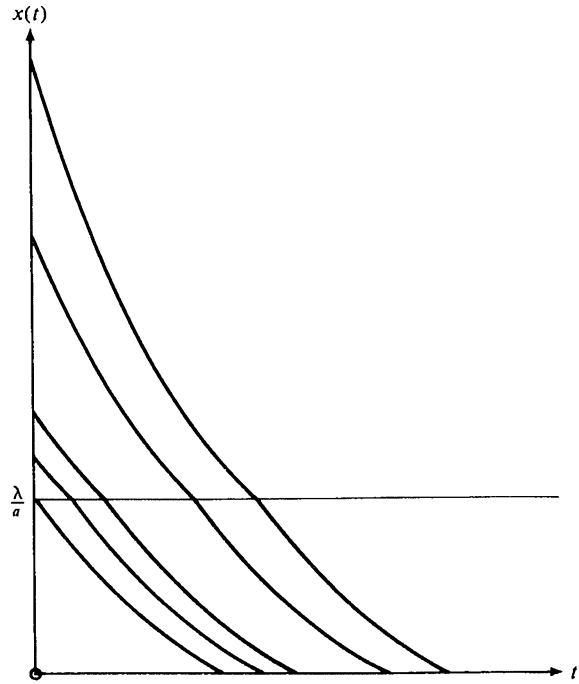


Figure 5-30 Several optimal trajectories for a weighted-time-fuel performance measure.

$u^*(t) = -1$ until the system reaches $x(t) = 0$. This is an intuitively reasonable result, because if $\lambda \longrightarrow \infty$, all trajectories begin with $x_0 \leq \lambda/a$ and will thus be minimum-time solutions. On the other hand, if $\lambda \longrightarrow 0$ the line λ/a moves very close to zero, and the optimal strategy approaches that indicated by Example 5.5-2 with free final time; let the system coast to as near the origin as possible before applying control.

The reader can show that for $x_0 < -\lambda/a$, the optimal strategy is to allow the system to coast [with $u^*(t) = 0$] until it reaches $x(t) = -\lambda/a$, where the optimal control switches to $u^*(t) = +1$.

The optimal control law—which is *time-invariant*—is summarized by

$$u^*(t) = \begin{cases} 0, & \text{for } \dfrac{\lambda}{a} < x(t) \\ -1.0, & \text{for } 0 < x(t) \leq \dfrac{\lambda}{a} \\ +1.0, & \text{for } -\dfrac{\lambda}{a} \leq x(t) < 0 \\ 0, & \text{for } x(t) < -\dfrac{\lambda}{a} \\ 0, & \text{for } x(t) = 0. \end{cases} \tag{5.5-72}$$

Figure 5-31 illustrates this optimal control law and its implementation. In solving this example the reader should note that we were able to determine the optimal control law using only the *form* of the costate solution—there was no need to solve for the constant of integration c_1. We also exploited the necessary condition that

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t)) = 0, \qquad t \in [0, t_f], \tag{5.5-73}$$

for t_f free and $\mathscr{H}$ not explicitly dependent on t, to determine the optimal control law and to show that the singular condition could not arise.

Let us now consider a somewhat less elementary example, which further illustrates the use of a weighted-time-fuel performance measure.

Example 5.5-5. Find the optimal control law for transferring the system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u(t) \end{aligned} \tag{5.5-74}$$

from an arbitrary initial state $\mathbf{x}(0) = \mathbf{x}_0 \neq \mathbf{0}$ to the final state $\mathbf{x}(t_f) = \mathbf{0}$ with a minimum value of the performance measure

$$J(u) = \int_0^{t_f} [\lambda + |u(t)|]\, dt. \tag{5.5-75}$$

The admissible controls are constrained by

$$|u(t)| \leq 1.0; \tag{5.5-76}$$

the final time t_f is free, and $\lambda > 0$.

The reader can easily verify that the presence of λ in the Hamiltonian

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = \lambda + |u(t)| + p_1(t)x_2(t) + p_2(t)u(t) \tag{5.5-77}$$

does not alter the form of the optimal control given by Eq. (5.5-14); therefore, we have

$$u^*(t) = \begin{cases} 1.0, & \text{for } p_2^*(t) < -1.0 \\ 0, & \text{for } -1.0 < p_2^*(t) < 1.0 \\ -1.0, & \text{for } 1.0 < p_2^*(t) \\ \text{undetermined, but } \geq 0 \text{ for } p_2^*(t) = -1.0 \\ \text{undetermined, but } \leq 0 \text{ for } p_2^*(t) = +1.0. \end{cases} \tag{5.5-78}$$

The costate equations

$$\begin{aligned} \dot{p}_1^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_1} = 0 \\ \dot{p}_2^*(t) &= -\frac{\partial \mathscr{H}}{\partial x_2} = -p_1^*(t) \end{aligned} \tag{5.5-79}$$

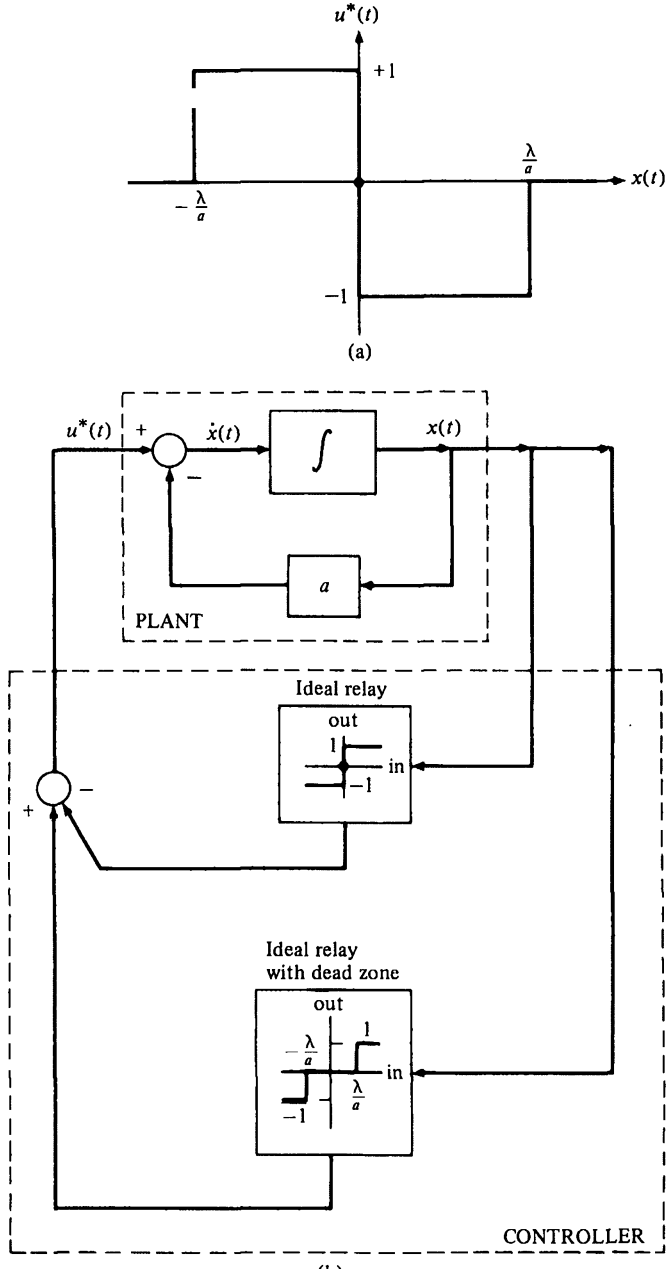


Figure 5-31 (a) The optimal control law for Example 5.5-4. (b) Implementation of the weighted-time-fuel optimal control law of Example 5.5-4

have solutions of the form

$$\begin{aligned} p_1^*(t) &= c_1 \\ p_2^*(t) &= -c_1 t + c_2. \end{aligned} \tag{5.5-80}$$

Clearly, p_2^* can change sign at most once, so the optimal control must have one of the forms (excluding singular solutions)

$$\begin{aligned} u^* = \{0\}, \{+1\}, \{-1\}, \{0, +1\}, \{0, -1\}, \{+1, 0\}, \\ \{-1, 0\}, \{+1, 0, -1\}, \{-1, 0, +1\}. \end{aligned} \tag{5.5-81}$$

First let us see whether or not there can be any singular solutions. For $p_2^*(t)$ to be equal to ± 1.0 during a finite time interval, it is necessary that $c_1 = 0$ and $c_2 = \pm 1.0$. Substituting $p_2^*(t) = \pm 1$ in (5.5-77), and using (5.5-78) and the definition of the absolute value function, we obtain

$$\mathscr{H}(\mathbf{x}^*(t), u^*(t), \mathbf{p}^*(t)) = \lambda > 0 \tag{5.5-82}$$

if the singular condition is to occur, but we know (since $\mathscr{H}$ is explicitly independent of time and t_f is free) that the Hamiltonian must be zero on an optimal trajectory. We conclude, then, that the singular condition cannot arise in this problem.

Let us now investigate the control alternatives given by Eq. (5.5-81). First, observe that none of the alternatives that ends with an interval of $u = 0$ can be optimal because the system (5.5-74) does not move to the origin with no control applied. Next, consider the optimal control candidates

$$u^* = \{-1\}, \{0, -1\}, \{+1, 0, -1\}. \tag{5.5-83}$$

To be optimal, the trajectories resulting from these three control forms must terminate at the origin with an interval of $u^* = -1$ control. The system differential equations are the same in this problem as in the minimum-time problem discussed in Example 5.4-4, so the terminal segments of these trajectories all lie on the curve B-0 in Fig. 5-20(b). Now, for any interval during which $u(t) = 0$ the state equations are

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= 0, \end{aligned} \tag{5.5-84}$$

which implies that

$$\begin{aligned} x_2(t) &= c_3 = \text{a constant} \\ x_1(t) &= c_3 t + c_4. \end{aligned} \tag{5.5-85}$$

Thus, as time increases, $x_1(t)$ increases or decreases, depending on whether $x_2(t)$ is greater or less than zero when the control switches to $u = 0$.

Several trajectories for $u = 0$ are shown in Fig. 5-32; the direction of increasing time is indicated by the arrows. Notice that if $x_2(t) = 0$ when the control switches to zero, the value of x_1 does not change until the control becomes nonzero.

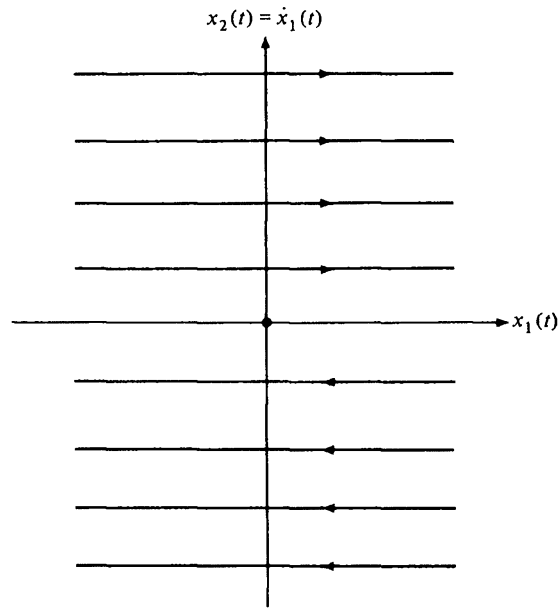


Figure 5-32 Trajectories for $u = 0$

Trajectory segments generated by $u = +1$ are the same as trajectories shown in Fig. 5-20(a).

To draw the candidates for an optimal trajectory we simply piece together segments of the trajectories shown in Figs. 5-20 and 5-32. The trajectories *C-D-E-0*, *C-F-G-0*, and *C-H-I-0* shown in Fig. 5-33 are three *candidates* for an optimal trajectory which has the initial state $\mathbf{x}_0$. Our task now is to determine the point on segment *C-K*, where the optimal control switches from $+1$ to 0. Once this point is known, we can easily determine the entire optimal trajectory.

Let t_1 be the time when the optimal control switches from $+1$ to 0, and let t_2 be the time when the optimal control switches from 0 to -1. Clearly, t_1 occurs somewhere on segment *C-K* and t_2 on segment *K-0*. We know from Eq. (5.4-40) that on *K-0*

$$x_1^*(t) = -\tfrac{1}{2}x_2^{*2}(t), \tag{5.5-86}$$

so

$$x_1^*(t_2) = -\tfrac{1}{2}x_2^{*2}(t_2). \tag{5.5-87}$$

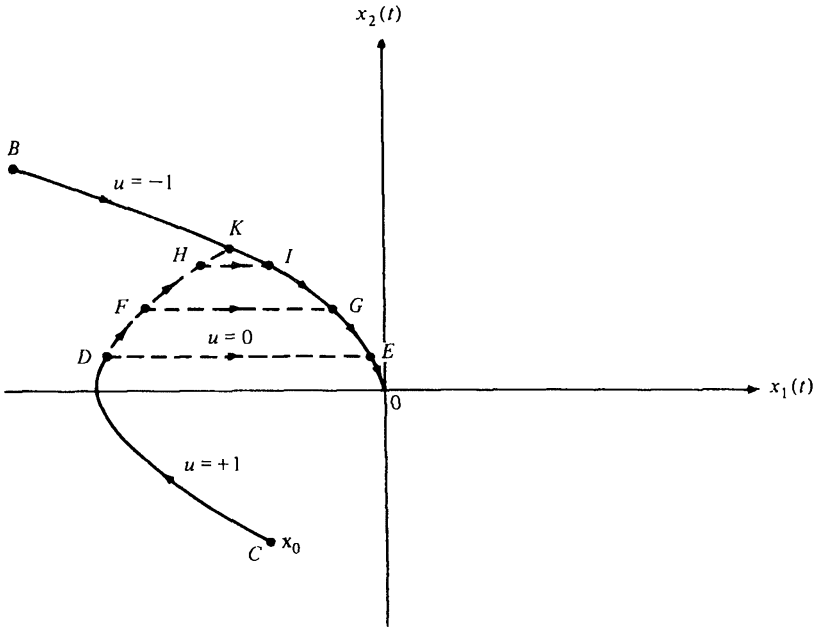


Figure 5-33 Three candidates for the optimal trajectory with initial state $\mathbf{x}_0$

In addition, integrating Eq. (5.5-84) gives

$$x_1^*(t_2) = x_1^*(t_1) + x_2^*(t_1)[t_2 - t_1], \tag{5.5-88}$$

and from Eqs. (5.5-80) and (5.5-78) we obtain

$$p_2^*(t_1) = -c_1 t_1 + c_2 = -1.0 \tag{5.5-89}$$

$$p_2^*(t_2) = -c_1 t_2 + c_2 = +1.0. \tag{5.5-90}$$

Because $p_2^*(t_1) = -1$ and $p_2^*(t_2) = +1$, the necessary condition that $\mathscr{H}$ be identically zero requires that

$$\lambda + c_1 x_2^*(t_1) = 0 \tag{5.5-91}$$

and

$$\lambda + c_1 x_2^*(t_2) = 0. \tag{5.5-92}$$

Let us now solve Eqs. (5.5-87) through (5.5-92) for $x_1^*(t_1)$.

First we observe that Eqs. (5.5-91) and (5.5-92) imply that $x_2^*(t_1) = x_2^*(t_2)$ and that

$$c_1 = \frac{-\lambda}{x_2^*(t_1)}. \tag{5.5-93}$$

Subtracting (5.5-90) from (5.5-89) gives

$$[t_2 - t_1] = -\frac{2}{c_1}, \tag{5.5-94}$$

which, if we use (5.5-93), becomes

$$[t_2 - t_1] = \frac{2x_2^*(t_1)}{\lambda}. \tag{5.5-95}$$

Putting this in (5.5-88) yields

$$x_1^*(t_2) = x_1^*(t_1) + \frac{2x_2^{*2}(t_1)}{\lambda}. \tag{5.5-96}$$

Substituting the right side of Eq. (5.5-87) for $x_1^*(t_2)$ and using the fact that $x_2^*(t_2) = x_2^*(t_1)$, yields

$$-\frac{1}{2}x_2^{*2}(t_1) = x_1^*(t_1) + \frac{2x_2^{*2}(t_1)}{\lambda}. \tag{5.5-97}$$

Collecting terms, we obtain

$$\boxed{x_1^*(t_1) = -\frac{\lambda + 4}{2\lambda}x_2^{*2}(t_1).} \tag{5.5-98a}$$

This is the sought-after result. The values of x_1 and x_2 that satisfy Eq. (5.5-98a) are the locus of points where the control switches from $+1$ to 0. It is left to the reader to show that for $u^* = \{-1, 0, +1\}$ the locus of points which defines the switching from $u^* = -1$ to $u^* = 0$ is given by

$$\boxed{x_1^*(t_1') = +\frac{\lambda + 4}{2\lambda}x_2^{*2}(t_1').} \tag{5.5-98b}$$

Notice particularly that Eq. (5.5-98) together with Eq. (5.5-87) and its counterpart for $u^*(t) = +1$ define the optimal control *law*. Furthermore, this optimal control law is time-invariant. The switching curves for $\lambda = 0.1$, 1.0, and 10.0 and several optimal trajectories for $\lambda = 1.0$ are shown in Fig. 5-34. Observe that if $\lambda \to \infty$ the switching curves merge together—the interval of $u^* = 0$ approaches zero, and trajectories

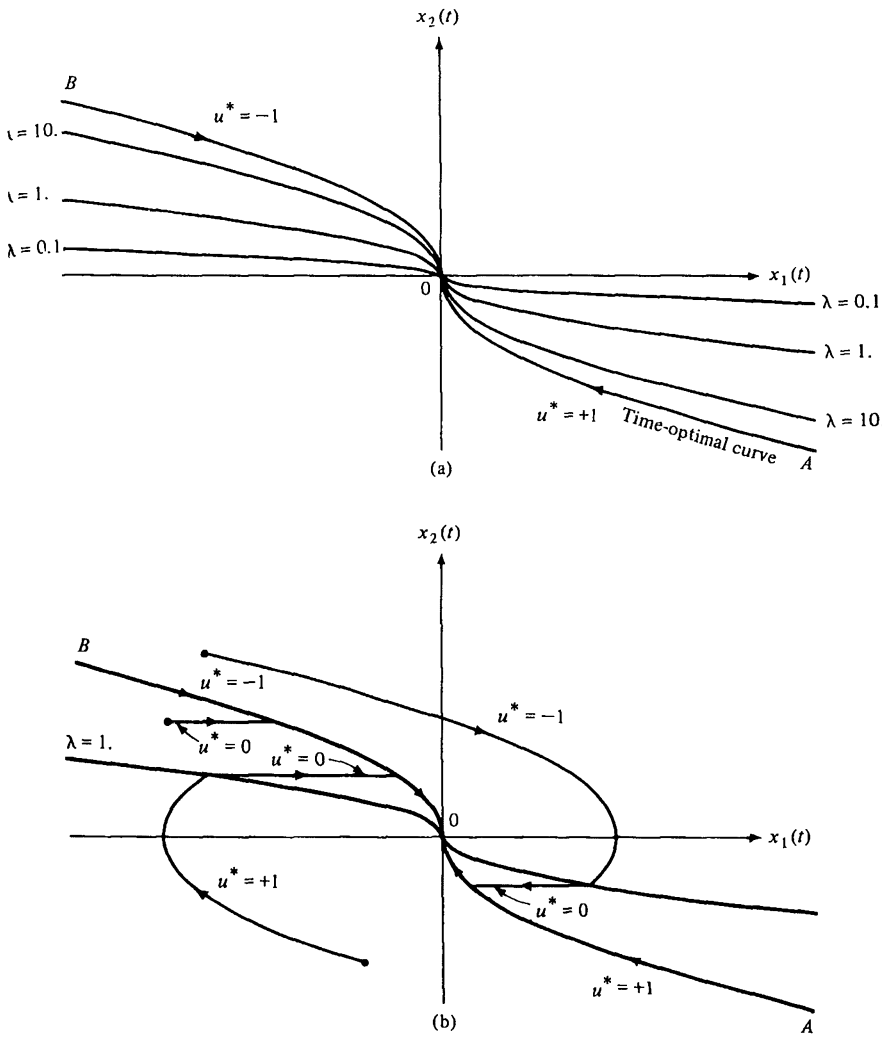


Figure 5-34 (a) Switching curves for weighted-time-fuel optimal performance. (b) Weighted-time-fuel optimal trajectories for three initial conditions ($\lambda = 1.0$)

approach the time-optimal trajectories of Example 5.4-4. On the other hand, if $\lambda \longrightarrow 0$, the interval of $u^* = 0$ approaches infinity, and the trajectories approach fuel-optimal trajectories.

The numerical value of λ must be decided upon subjectively by the designer. To help in making this decision, curves showing the dependence of elapsed time and consumed fuel on λ—such as Fig. 5-35—could be plotted for several initial conditions.

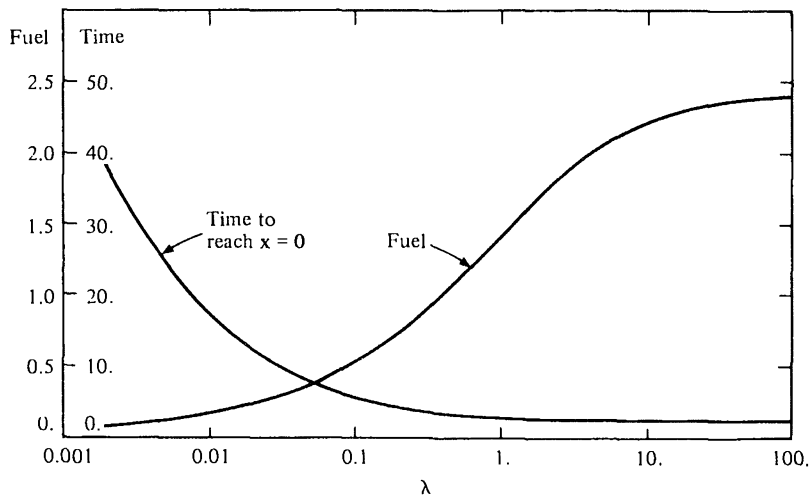


Figure 5-35 The dependence of elapsed time and consumed fuel on the weighting parameter λ, $\mathbf{x}(0) = \begin{bmatrix} -1.5 \\ 0.0 \end{bmatrix}$

Minimum-Energy Problems

The characteristics of fuel-optimal problems and energy-optimal problems are similar; therefore, the following discussion will be limited to one example, which illustrates some of the differences between these two types of systems.

Example 5.5-6. The plant of Examples 5.5-2 and 5.5-4

$$\dot{x}(t) = -ax(t) + u(t) \tag{5.5-99}$$

is to be transferred from an arbitrary initial state, $x(0) = x_0$, to the origin by a control that minimizes the performance measure

$$J(u) = \int_0^{t_f} [\lambda + u^2(t)]\, dt; \tag{5.5-100}$$

the admissible controls are constrained by

$$|u(t)| \leq 1. \tag{5.5-101}$$

The plant parameter a and the weighting factor λ are greater than zero, and the final time t_f is free. The objective is to find the optimal control law.

The first step, as usual, is to form the Hamiltonian,

$$\mathscr{H}(x(t), u(t), p(t)) = \lambda + u^2(t) - p(t)ax(t) + p(t)u(t). \tag{5.5-102}$$

The costate equation and its solution are

$$\dot{p}^*(t) = ap^*(t) \tag{5.5-103}$$

and

$$p^*(t) = c_1\epsilon^{at}. \tag{5.5-104}$$

For $|u(t)| < 1$, the control that minimizes $\mathscr{H}$ is the solution of the equation

$$\frac{\partial \mathscr{H}}{\partial u} = 2u^*(t) + p^*(t) = 0. \tag{5.5-105}$$

Notice that $\mathscr{H}$ is quadratic in $u(t)$ and

$$\frac{\partial^2 \mathscr{H}}{\partial u^2} = 2 > 0, \tag{5.5-106}$$

so

$$u^*(t) = -\tfrac{1}{2}p^*(t) \tag{5.5-107}$$

does globally minimize the Hamiltonian for $|u^*(t)| < 1$, or, equivalently, for

$$|p^*(t)| < 2. \tag{5.5-108}$$

If $|p^*(t)| \geq 2$, then the control that minimizes $\mathscr{H}$ is

$$u^*(t) = \begin{cases} +1.0, & \text{for } p^*(t) \leq -2.0 \\ -1.0, & \text{for } 2.0 \leq p^*(t). \end{cases} \tag{5.5-109}$$

Putting Eqs. (5.5-107) and (5.5-109) together, we obtain

$$u^*(t) = \begin{cases} 1.0, & \text{for } p^*(t) \leq -2.0 \\ -\tfrac{1}{2}p^*(t), & \text{for } -2.0 < p^*(t) < 2.0 \\ -1.0, & \text{for } 2.0 \leq p^*(t). \end{cases} \tag{5.5-110}$$

This relationship between an extremal control and an extremal costate is illustrated in Fig. 5-36. There is no possibility of singular solutions in this example, since there are no values of $p^*(t)$ for which the Hamiltonian is unaffected by $u(t)$.

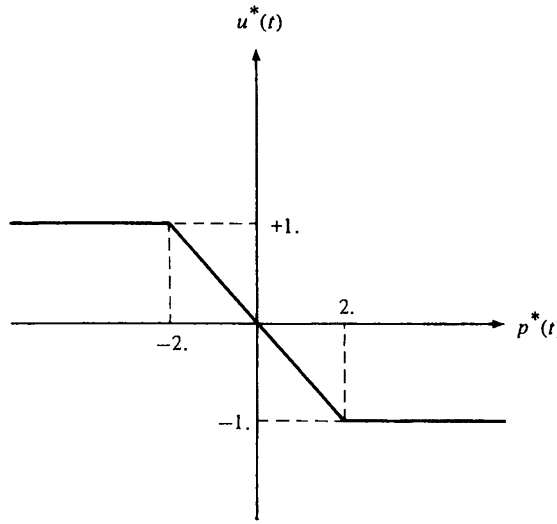


Figure 5-36 The relationship between an extremal control and costate

We rule out the possibility that $p^*(t) = 0$ for $t \in [0, t_f]$ {since this implies $u^*(t) = 0$ for $t \in [0, t_f]$ and the system would never reach the origin}; the possible forms for $p^*(t)$ are shown in Fig. 5-37. Corresponding to the costate curves labeled 1, 2, 3, and 4 are the optimal control possibilities:

1. $u^* = \{-\frac{1}{2}p^*\}$, or $\{-\frac{1}{2}p^*, -1.0\}$, (5.5-111a)

   depending on whether or not the system reaches the origin before p^* attains the value 2.0.

2. $u^* = \{-1.0\}$. (5.5-111b)

3. $u^* = \{-\frac{1}{2}p^*\}$, or $\{-\frac{1}{2}p^*, +1.0\}$, (5.5-111c)

   depending on whether or not the system reaches the origin before p^* attains the value -2.0.

4. $u^* = \{+1.0\}$. (5.5-111d)

The controls given by (5.5-111a) and (5.5-111b) are nonpositive for all $t \in [0, t_f]$ and correspond to positive state values. This can be seen from the solution of the state equation

$$x(t_f) = 0 = \epsilon^{-a[t_f - t]}x(t) + \epsilon^{-at_f}\int_t^{t_f} \epsilon^{a\tau}u(\tau)\,d\tau, \qquad (5.5\text{-}112)$$

which implies that

$$-\epsilon^{at}x(t) = \int_t^{t_f} \epsilon^{a\tau}u(\tau)\,d\tau. \qquad (5.5\text{-}113)$$

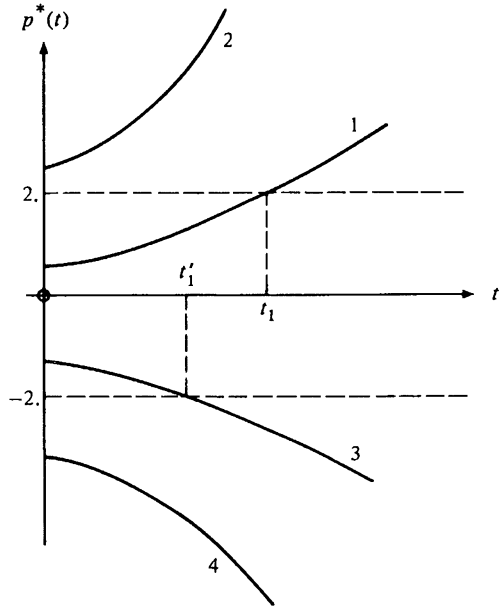


Figure 5-37 Possible forms for an extremal costate trajectory

For $u(\tau)$ nonpositive when $\tau \in [t, t_f]$, the integral is negative; therefore, $x(t)$ must be positive. Similarly, the nonnegative controls specified by Eqs. (5.5-111c) and (5.5-111d) correspond to negative values of $x(t)$.

Since t_f is free, and the Hamiltonian does not contain t explicitly, it is also necessary that

$$\mathscr{H}(x^*(t), u^*(t), p^*(t)) = 0, \qquad t \in [t_0, t_f]. \tag{5.5-114}$$

If the control saturates at the value -1 when $t = t_1$, then from (5.5-110), $p^*(t_1) = 2.0$; substituting $u^*(t_1) = -1$ and $p^*(t_1) = 2$ in $\mathscr{H}$, we obtain

$$\mathscr{H}(x^*(t_1), u^*(t_1), p^*(t_1)) = \lambda + 1 - 2ax^*(t_1) - 2 = 0, \tag{5.5-115}$$

which implies

$$x^*(t_1) = \frac{\lambda - 1}{2a}. \tag{5.5-116}$$

If the control saturates at -1 when $t = t_1$, then from (5.5-111) $u^*(t) = -1$ for $t \in [t_1, t_f]$, and $x^*(t) < x^*(t_1)$ for $t > t_1$; thus,

$$u^*(t) = -1 \qquad \text{for } 0 < x^*(t) < \frac{\lambda - 1}{2a}. \tag{5.5-117a}$$

Using similar reasoning, we can show that if the control saturates at the value $+1$ when $t = t_1'$, then

$$x^*(t_1') = \frac{\lambda - 1}{-2a} \tag{5.5-118}$$

and

$$u^*(t) = +1 \qquad \text{for} \quad \frac{\lambda - 1}{-2a} < x^*(t) < 0. \tag{5.5-117b}$$

Notice that if $\lambda \leq 1$, $x^*(t_1) \leq 0$ in (5.5-116), and $x^*(t_1') \geq 0$ in (5.5-118), but (5.5-116) applies for positive state values and (5.5-118) applies for negative state values; hence the optimal control does not saturate for $\lambda \leq 1$.

Let us now examine the unsaturated region where $u^*(t) = -\frac{1}{2}p^*(t)$. Again using the necessary condition of Eq. (5.5-114), by substituting $u^*(t) = -\frac{1}{2}p^*(t)$, we obtain

$$\begin{aligned}\mathscr{H}(x^*(t), u^*(t), p^*(t)) &= \lambda + \tfrac{1}{4}p^{*2}(t) - p^*(t)ax^*(t) \\ &\quad - \tfrac{1}{2}p^{*2}(t) = 0.\end{aligned} \tag{5.5-119}$$

Solving for $p^*(t)$ yields

$$p^*(t) = 2\left[-ax^*(t) \pm \sqrt{[ax^*(t)]^2 + \lambda}\right], \tag{5.5-120}$$

which implies

$$u^*(t) = \left[ax^*(t) \pm \sqrt{[ax^*(t)]^2 + \lambda}\right]. \tag{5.5-121}$$

If $x^*(t) > 0$, $u^*(t)$ must be negative, so the minus sign applies; for $x^*(t) < 0$ the positive sign applies. The optimal control law, if we put together Eqs. (5.5-121) and (5.5-117), is

$$u^*(t) = \begin{cases} \left[ax(t) - \sqrt{[ax(t)]^2 + \lambda}\right], & \text{for } 0 < \dfrac{\lambda - 1}{2a} < x(t) \\ -1.0, & \text{for } 0 < x(t) \leq \dfrac{\lambda - 1}{2a} \\ +1.0, & \text{for } -\dfrac{\lambda - 1}{2a} \leq x(t) < 0 \\ \left[ax(t) + \sqrt{[ax(t)]^2 + \lambda}\right], & \text{for } x(t) < -\dfrac{\lambda - 1}{2a} < 0 \\ 0, & \text{for } x(t) = 0. \end{cases} \tag{5.5-122}\dagger$$

Figure 5-38 illustrates this optimal control law and its implementation. Comparing Figs. 5-31(a) and 5-38(a), the reader will note that the weighted-time-fuel-optimal controls are either "on" (± 1) or "off" (0), whereas the weighted-time-energy-optimal controls can assume all values from -1 to $+1$.

† The optimal control law is valid for all state values, so we write $x(t)$ instead of $x^*(t)$.

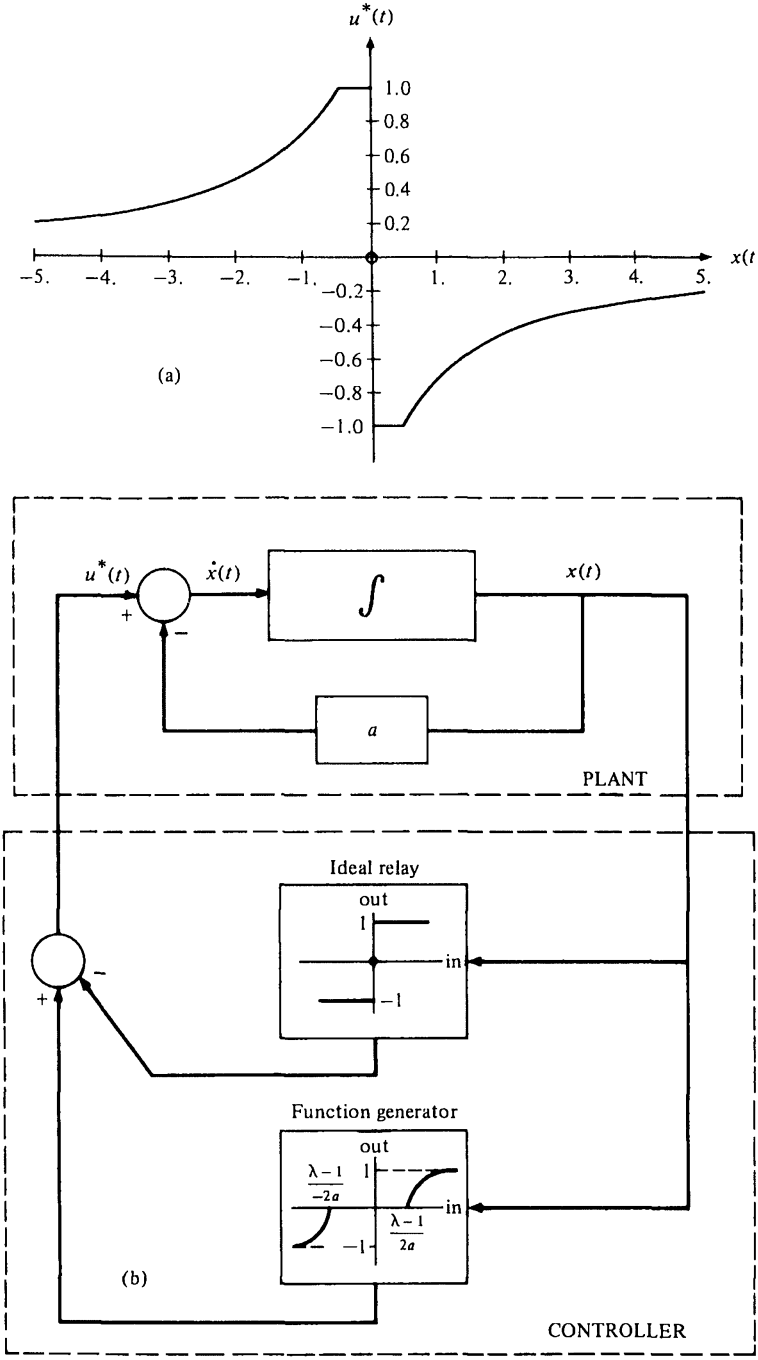


Figure 5-38 (a) The weighted-time-energy optimal control law for Example 5.5-6: $\lambda = 2.0, a = 1.0$. (b) Implementation of the weighted-time-energy optimal control law for Example 5.5-6

To provide an additional basis for comparing this energy-optimal system with the fuel-optimal system of Example 5.5-4, several optimal trajectories for each system are shown in Fig. 5-39. For the weighted-time-energy-optimal system λ was 2; the value of λ for the weighted-time-fuel-optimal system was adjusted for each initial condition to make the two systems require the same amount of time to reach the origin. The fuel and energy requirements for the two systems are summarized in Table 5-2.

Table 5-2 FUEL AND ENERGY REQUIREMENTS FOR THE SYSTEMS OF EXAMPLES 5.5-4 AND 5.5-6

Initial condition $x(0)$	*Time required to reach* $x(t_f) = 0$	*Fuel for time-fuel-optimal system*	*Fuel for time-energy-optimal system*	*Energy for time-energy-optimal system*	*Energy for time-fuel-optimal system*
1.5	0.982	0.8252	0.8434	0.7473	0.8252
2.0	1.205	0.9138	0.9559	0.8043	0.9138
2.5	1.393	0.9688	1.0326	0.8356	0.9688
3.0	1.555	1.0038	1.0883	0.8548	1.0038
5.0	2.034	1.0612	1.2090	0.8858	1.0612
7.0	2.361	1.0788	1.2636	0.8950	1.0788

Summary

In this section we have considered the optimization of systems whose control effort is to be conserved. Although our discussion was primarily concerned with the solution of several example problems, it was shown that the form of fuel-optimal controls for a class of nonlinear systems is "bang-off-bang"; it was left as an exercise for the reader (Problem 5-30) to show that the form of energy-optimal controls for the same class of nonlinear systems is a continuous, saturating function.

In all of the examples considered a trade-off existed between conservation of control effort and rapid action. It was found that such problems may be characterized by nonunique or nonexistent optimal controls when the final time is free, and that fixing the final time may still result in nonunique optimal controls or in a time-varying optimal control law. To circumvent these difficulties, a performance measure consisting of a weighted combination of elapsed-time and control-effort expended was introduced. In the problems solved, this form of performance measure resulted in time-invariant optimal control laws, and, in addition, reflected the trade-off between conservation of control effort and rapid action. It should be emphasized that there are alternative formulations of minimum-control-effort problems (see Problem 5-33)

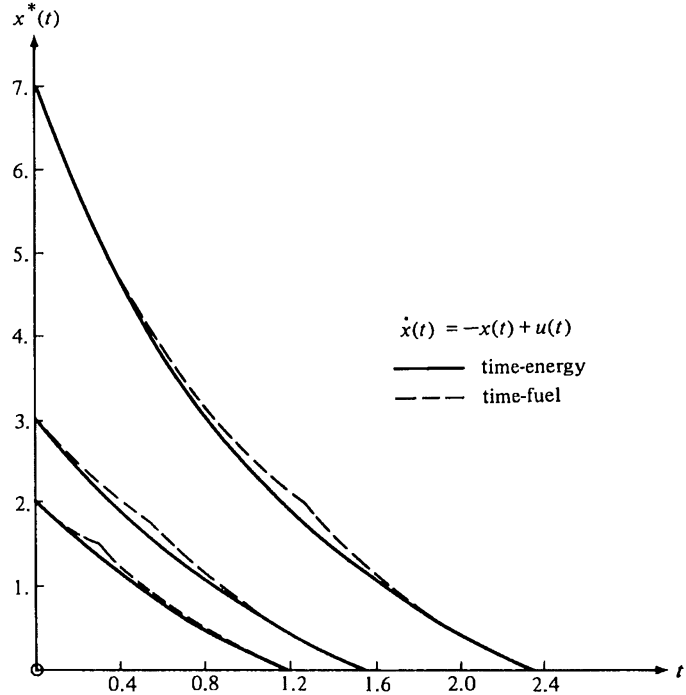


Figure 5-39 Weighted-time-fuel and weighted-time-energy optimal trajectories

and that conserving control effort and obtaining rapid action may not always be conflicting objectives (see Problems 5-28 and 5-31).

No attempt was made to generalize the results of the examples to a "design procedure." The reason for this omission is that unless the system is of low order, time-invariant, and linear, we have little hope of analytically determining the optimal control law. The difficulties mentioned at the end of Section 5.4 for time-optimal systems also apply to the energy- and fuel-optimal systems considered here—only more so. The primary virtue of the discussion in this section is that it provides insight into the form of the optimal control and furnishes a starting point for numerical determination of the optimal control law.

5.6 SINGULAR INTERVALS IN OPTIMAL CONTROL PROBLEMS

In discussing minimum-time and minimum-control-effort problems we have used Pontryagin's necessary condition,

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \leq \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) \tag{5.6-1}$$

for all $t \in [t_0, t_f]$ and for all admissible $\mathbf{u}(t)$, to determine $\mathbf{u}^*(t)$ in terms of the extremal states and costates. If, however, there is a time interval $[t_1, t_2]$ of finite duration during which the necessary condition (5.6-1) provides no information about the relationship between $\mathbf{u}^*(t)$, $\mathbf{x}^*(t)$, and $\mathbf{p}^*(t)$, then we say that the problem is singular. The interval $[t_1, t_2]$ is called an interval of singularity, or simply a *singular interval.*

We shall now investigate the conditions that allow singular intervals to occur, and the effects of singular intervals on optimal controls and trajectories. To begin our investigation, let us return to a minimum-time problem discussed in Section 5.4.

Example 5.6-1. In Example 5.4-4 we considered the problem of transferring the system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u(t) \end{aligned} \tag{5.6-2}$$

from an arbitrary initial state to the origin in minimum time. The admissible controls were required to satisfy the inequality

$$|u(t)| \leq 1.0. \tag{5.6-3}$$

In solving this problem we assumed that a singular interval did not exist; let us now verify that this assumption was correct.

The Hamiltonian is

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = 1 + p_1(t)\, x_2(t) + p_2(t)u(t), \tag{5.6-4}$$

and application of the minimum principle gives

$$1 + p_1^*(t)x_2^*(t) + p_2^*(t)u^*(t) \leq 1 + p_1^*(t)x_2^*(t) + p_2^*(t)u(t). \tag{5.6-5}$$

If there exists a time interval $[t_1, t_2]$ during which

$$p_2^*(t) = 0, \tag{5.6-6}$$

then (5.6-5) provides no information about the relationship between $u^*(t)$, $\mathbf{x}^*(t)$, and $\mathbf{p}^*(t)$. Therefore, if

$$p_2^*(t) = 0 \qquad \text{for } t \in [t_1, t_2], \tag{5.6-7}$$

then $[t_1, t_2]$ is a singular interval.† Let us investigate further to see if this condition can occur. The costate equations

† Isolated times when $p_2^*(t)$ *passes through* zero indicate a switching of the control, not a singular interval.

$$\begin{aligned} \dot{p}_1^*(t) &= 0 \\ \dot{p}_2^*(t) &= -p_1^*(t) \end{aligned} \tag{5.6-8}$$

have solutions of the form

$$\begin{aligned} p_1^*(t) &= c_1 \\ p_2^*(t) &= -c_1 t + c_2. \end{aligned} \tag{5.6-9}$$

But for $p_2^*(t) = 0$ for $t \in [t_1, t_2]$ it is necessary that

$$c_1 = 0 \tag{5.6-10a}$$

and

$$c_2 = 0. \tag{5.6-10b}$$

Substituting these values in the Hamiltonian gives

$$\mathscr{H}(\mathbf{x}^*(t), u^*(t), \mathbf{p}^*(t)) = 1 \qquad \text{for all } t \in [0, t_f], \tag{5.6-11}$$

but since the final time is free and $\mathscr{H}$ is explicitly independent of time, Eq. (5.6-11) violates the necessary condition that

$$\mathscr{H}(\mathbf{x}^*(t), u^*(t), \mathbf{p}^*(t)) = 0 \qquad \text{for all } t \in [0, t_f]. \tag{5.6-12}$$

We conclude that $p_2^*(t)$ cannot be zero during a finite time interval, and, thus, that a singular interval cannot exist.

Let us now discuss in more generality the possibility of singular intervals occurring in linear minimum-time problems.

Singular Intervals in Linear Time-Optimal Problems

Consider the minimum-time transfer of the linear, stationary system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \tag{5.6-13}$$

from an arbitrary initial state $\mathbf{x}_0$ at $t = 0$ to some target set $S(t)$. For simplicity we shall assume that the control is a scalar. The admissible controls satisfy the inequality

$$|u(t)| \leq 1.0. \tag{5.6-14}$$

Let us attempt to find conditions that are necessary for the existence of a singular interval.

The Hamiltonian is

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = 1 + \mathbf{p}^T(t)\mathbf{A}\mathbf{x}(t) + \mathbf{p}^T(t)\mathbf{b}u(t), \tag{5.6-15}$$

and from the minimum principle we know that if an optimal control u^* exists it must satisfy

$$1 + \mathbf{p}^{*T}(t)\mathbf{A}\mathbf{x}^*(t) + \mathbf{p}^{*T}(t)\mathbf{b}u^*(t) \leq 1 + \mathbf{p}^{*T}(t)\mathbf{A}\mathbf{x}^*(t) + \mathbf{p}^{*T}(t)\mathbf{b}u(t) \quad (5.6\text{-}16)$$

for all $t \in [0, t_f]$ and for all admissible $u(t)$. Since the final time is free and $\mathscr{H}$ does not contain t explicitly, we also know that

$$\mathscr{H}(\mathbf{x}^*(t), u^*(t), \mathbf{p}^*(t)) = 1 + \mathbf{p}^{*T}(t)\,\mathbf{A}\mathbf{x}^*(t) + \mathbf{p}^{*T}(t)\,\mathbf{b}u^*(t) = 0 \quad (5.6\text{-}17)$$

for all $t \in [0, t_f]$. From (5.6-16) we observe that $[t_1, t_2]$ is a singular interval if

$$\mathbf{p}^{*T}(t)\,\mathbf{b} = 0 \qquad \text{for all } t \in [t_1, t_2]. \quad (5.6\text{-}18)$$

Clearly, this condition occurs *if* $\mathbf{p}^*(t) = \mathbf{0}$ for $t \in [t_1, t_2]$. But this cannot happen, because substituting $\mathbf{p}^*(t) = \mathbf{0}$ in Eq. (5.6-17) leads to the contradiction $1 = 0$; therefore,

$$\mathbf{p}^*(t) \not\equiv \mathbf{0} \qquad \textit{for any } t \in [0, t_f]. \quad (5.6\text{-}19)$$

Equation (5.6-18) is also satisfied (for all t) if

$$\mathbf{b} = \mathbf{0}, \quad (5.6\text{-}20)$$

but this indicates that the control does not affect the system at all; we might say that the system is "completely uncontrollable." This is our first hint that perhaps controllability has something to do with the existence of singular intervals.

Having ruled out $\mathbf{p}^*(t) = \mathbf{0}$, or $\mathbf{b} = \mathbf{0}$ as possibilities, let us consider the remaining alternative, namely that the product $\mathbf{p}^{*T}(t)\,\mathbf{b} = 0$ for $t \in [t_1, t_2]$. If $\mathbf{p}^{*T}(t)\,\mathbf{b}$ is to be zero for a finite time interval, this implies that derivatives of all orders of $\mathbf{p}^{*T}(t)\,\mathbf{b}$ are zero during this interval; that is,

$$\begin{aligned} &\mathbf{p}^{*T}(t)\mathbf{b} = 0 \\ &\frac{d^k}{dt^k}[\mathbf{p}^{*T}(t)\mathbf{b}] = 0, \qquad k = 1, 2, \ldots \end{aligned} \quad (5.6\text{-}21)$$

Since $\mathbf{b}$ is an $n \times 1$ matrix of constants,

$$\begin{aligned} \frac{d^k}{dt^k}[\mathbf{p}^{*T}(t)\mathbf{b}] &= \frac{d^k}{dt^k}[\mathbf{p}^{*T}(t)]\mathbf{b} \\ &\triangleq \overset{(k)}{\mathbf{p}}{}^{*T}(t)\mathbf{b}. \end{aligned} \quad (5.6\text{-}22)$$

From the Hamiltonian the costate equation is

$$\dot{\mathbf{p}}^*(t) = -\mathbf{A}^T\mathbf{p}^*(t); \quad (5.6\text{-}23)$$

hence the costate solution is

$$\mathbf{p}^*(t) = \epsilon^{-\mathbf{A}^T t}\mathbf{c}, \tag{5.6-24}$$

where $\mathbf{c}$ is the vector of initial costate values.

Let us write out a few of the derivatives in Eq. (5.6-21); we have

$$\begin{aligned}
&\mathbf{p}^{*T}(t)\mathbf{b} = 0\\
&\dot{\mathbf{p}}^{*T}(t)\mathbf{b} = 0\\
&\ddot{\mathbf{p}}^{*T}(t)\mathbf{b} = 0\\
&\quad\cdot\\
&\quad\cdot\\
&\quad\cdot\\
&\overset{(k)}{\mathbf{p}}{}^{*T}(t)\mathbf{b} = 0 \qquad \text{for } t \in [t_1, t_2].
\end{aligned} \tag{5.6-25}$$

Now, using Eqs. (5.6-23) and (5.6-24), we have

$$\dot{\mathbf{p}}^{*T}(t)\mathbf{b} = -[\mathbf{A}^T\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{b} = 0. \tag{5.6-26}$$

By applying the matrix identity

$$[\mathbf{M}_1\mathbf{M}_2]^T = \mathbf{M}_2^T\mathbf{M}_1^T\dagger \tag{5.6-27}$$

Eq. (5.6-26) becomes

$$[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{A}\mathbf{b} = 0. \tag{5.6-28}$$

Similarly, differentiating Eq. (5.6-23) gives

$$\ddot{\mathbf{p}}^*(t) = -\mathbf{A}^T\dot{\mathbf{p}}^*(t), \tag{5.6-29}$$

so

$$\ddot{\mathbf{p}}^{*T}(t)\mathbf{b} = [[-\mathbf{A}^T][-\mathbf{A}^T]\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{b} = 0. \tag{5.6-30}$$

Using (5.6-27) twice on the term in brackets gives

$$[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{A}^2\mathbf{b} = 0. \tag{5.6-31}$$

The pattern is now clear; continuing to write out the terms of Eq. (5.6-21), using Eqs. (5.6-23), (5.6-24), and (5.6-27), we obtain for the kth derivative

$$\overset{(k)}{\mathbf{p}}{}^{*T}(t)\mathbf{b} = [-1]^k[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{A}^k\mathbf{b} = 0, \qquad k = 0, 1, 2, \ldots. \tag{5.6-32}$$

Cancelling the minus signs, we find that the first n equations are‡

† See Appendix 1.

‡ Recall that n is the order of the system.

$$
\begin{aligned}
&[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{b} = 0\\
&[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{A}\mathbf{b} = 0\\
&[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{A}^2\mathbf{b} = 0\\
&\quad\vdots\\
&[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\mathbf{A}^{n-1}\mathbf{b} = 0,
\end{aligned}
\tag{5.6-33}
$$

or, written together,

$$
[\epsilon^{-\mathbf{A}^T t}\mathbf{c}]^T\left[\mathbf{b}\,\middle|\,\mathbf{A}\mathbf{b}\,\middle|\,\mathbf{A}^2\mathbf{b}\,\middle|\,\cdots\,\middle|\,\mathbf{A}^{n-1}\mathbf{b}\right] = \mathbf{0}^T. \tag{5.6-33a}
$$

Taking the transpose of both sides and again using Eq. (5.6-27), we find that this becomes

$$
\left[\mathbf{b}\,\middle|\,\mathbf{A}\mathbf{b}\,\middle|\,\mathbf{A}^2\mathbf{b}\,\middle|\,\cdots\,\middle|\,\mathbf{A}^{n-1}\mathbf{b}\right]^T\epsilon^{-\mathbf{A}^T t}\mathbf{c} = \mathbf{0}. \tag{5.6-34}
$$

But

$$
\epsilon^{-\mathbf{A}^T t}\mathbf{c} = \mathbf{p}^*(t), \tag{5.6-24}
$$

and we have already shown [see Eq. (5.6-19)] that $\mathbf{p}^*(t) \not\equiv \mathbf{0}$ for any $t \in [0, t_f]$; therefore, if Eq. (5.6-34) is to be satisfied, the matrix

$$
\mathbf{E} \triangleq \left[\mathbf{b}\,\middle|\,\mathbf{A}\mathbf{b}\,\middle|\,\mathbf{A}^2\mathbf{b}\,\middle|\,\cdots\,\middle|\,\mathbf{A}^{n-1}\mathbf{b}\right]
$$

must be singular. From Section 1.2 we know that the matrix $\mathbf{E}$ is nonsingular if and only if the system (5.6-13) is completely controllable.

To summarize, we have found that in linear, stationary, minimum-time problems:

1. For a singular interval to exist, it is necessary that the system be uncontrollable.
2. Conversely, if the system is completely controllable, a singular interval cannot exist.

It can also be shown that if $\mathbf{E}$ is singular, a singular interval must exist.

In conclusion, the problem of transferring the system

$$
\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \tag{5.6-13}
$$

from an arbitrary initial state $\mathbf{x}_0$ to a specified target set in minimum time has a singular interval if and only if the system (5.6-13) is *not completely controllable*. This necessary and sufficient condition for the existence of an

interval of singularity can also be extended to the situation where the system has several inputs (see Problem 5-39).

Singular Intervals in Linear Fuel-Optimal Problems

Let us now investigate minimum-fuel systems to see whether or not singular intervals can exist. We begin by considering the fuel-optimal control of the system in Example 5.6-1.

Example 5.6-2. Determine whether the problem of transferring the system

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= u(t) \end{aligned} \tag{5.6-35}$$

from an arbitrary initial state $\mathbf{x}_0$ to a specified target set $S(t)$ with minimum fuel expenditure has any singular intervals. The final time is free.

The Hamiltonian is given by

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = |u(t)| + p_1(t)x_2(t) + p_2(t)u(t) \tag{5.6-36}$$

and from the minimum principle,

$$\begin{aligned} |u^*(t)| + p_1^*(t)x_2^*(t) + p_2^*(t)u^*(t) \leq |u(t)| \\ + p_1^*(t)x_2^*(t) + p_2^*(t)u(t). \end{aligned} \tag{5.6-37}$$

It is also necessary that on an extremal $\mathscr{H} \equiv 0$, so

$$|u^*(t)| + p_1^*(t)x_2^*(t) + p_2^*(t)u^*(t) = 0. \tag{5.6-38}$$

If $[t_1, t_2]$ is a singular interval, Eq. (5.6-37) indicates that either

$$p_2^*(t) = +1.0 \qquad \text{for all } t \in [t_1, t_2] \tag{5.6-39a}$$

or

$$p_2^*(t) = -1.0 \qquad \text{for all } t \in [t_1, t_2]. \tag{5.6-39b}$$

In either case, if (5.6-37) is satisfied, Eq. (5.6-38) reduces to

$$p_1^*(t)x_2^*(t) = 0 \qquad \text{for all } t \in [t_1, t_2]. \tag{5.6-40}$$

The costate solution, found earlier, is

$$\begin{aligned} p_1^*(t) &= c_1 \\ p_2^*(t) &= -c_1 t + c_2. \end{aligned} \tag{5.6-9}$$

In order that $p_2^*(t) = \pm 1$ for a finite time interval, c_1 must equal zero, and c_2 must equal ± 1.0. If $c_1 = 0$, $p_1^*(t) = 0$ for all t and Eq. (5.6-40)

will be satisfied. From this analysis, we have determined that a singular interval can exist, even though this system is completely controllable. Notice that if a singular interval occurs it will persist for all $t \in [0, t_f]$; thus, if the optimal control is singular at all, it is singular throughout the interval of operation of the system.

It is left as an exercise for the reader (Problem 5-36) to show that in this problem the existence of a singular interval signals the nonuniqueness of optimal controls for certain initial states and the nonexistence of optimal controls for the rest of the initial states.

Let us now consider linear fuel-optimal systems in more generality. We shall assume that the system has one control input and is described by state equations of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t). \tag{5.6-41}$$

The admissible controls must satisfy

$$|u(t)| \leq 1.0. \tag{5.6-42}$$

The system is to be transferred from an arbitrary initial state $\mathbf{x}_0$ to a specified target set $S(t)$ by a control that minimizes the performance measure

$$J(u) = \int_0^{t_f} |u(t)|\, dt \tag{5.6-43}$$

with t_f free. The Hamiltonian is

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = |u(t)| + \mathbf{p}^T(t)\mathbf{A}\mathbf{x}(t) + \mathbf{p}^T(t)\mathbf{b}u(t). \tag{5.6-44}$$

From the minimum principle

$$\begin{aligned} |u^*(t)| + \mathbf{p}^{*T}(t)\mathbf{A}\mathbf{x}^*(t) + \mathbf{p}^{*T}(t)\,\mathbf{b}u^*(t) \leq |u(t)| \\ + \mathbf{p}^{*T}(t)\mathbf{A}\mathbf{x}^*(t) + \mathbf{p}^{*T}(t)\,\mathbf{b}u(t). \end{aligned} \tag{5.6-45}$$

From Eq. (5.6-45), we see that for a singular interval to exist it is necessary that either

$$\mathbf{p}^{*T}(t)\,\mathbf{b} = +1.0 \qquad \text{for all } t \in [t_1, t_2] \tag{5.6-46a}$$

or

$$\mathbf{p}^{*T}(t)\,\mathbf{b} = -1.0 \qquad \text{for all } t \in [t_1, t_2]. \tag{5.6-46b}$$

If $u^*(t)$ minimizes the Hamiltonian, and either (5.6-46a) or (5.6-46b) is satisfied, then since $\mathscr{H}$ must be identically zero,

$$\mathbf{p}^{*T}(t)\,\mathbf{A}\mathbf{x}^*(t) = 0. \tag{5.6-47}$$

If $\mathbf{p}^{*T}(t)\,\mathbf{b}$ is to be either $+1$ or -1 during the entire time interval $[t_1, t_2]$, then this implies that

$$\frac{d^k}{dt^k}[\mathbf{p}^{*T}(t)\mathbf{b}] = 0, \qquad k = 1, 2, \ldots, \qquad t \in [t_1, t_2]. \tag{5.6-48}$$

Again using the necessary condition that $\mathscr{H}(\mathbf{x}^*(t), u^*(t), \mathbf{p}^*(t)) \equiv 0$, and following the same procedure as for minimum-time problems, we eventually obtain

$$\left[\epsilon^{-\mathbf{A}^T t}\mathbf{c}\right]^T\left[\mathbf{A}\mathbf{b} \mid \mathbf{A}^2\mathbf{b} \mid \cdots \mid \mathbf{A}^n\mathbf{b}\right] = \mathbf{0}^T \tag{5.6-49}$$

[compare this with Eq. (5.6-33a)]. This equation can also be written as

$$\left[\mathbf{b} \mid \mathbf{A}\mathbf{b} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{b}\right]^T\mathbf{A}^T\epsilon^{-\mathbf{A}^T t}\mathbf{c} = \mathbf{0}. \tag{5.6-50}$$

But

$$\epsilon^{-\mathbf{A}^T t}\mathbf{c} = \mathbf{p}^*(t) \neq \mathbf{0} \qquad \text{for } t \in [t_1, t_2] \tag{5.6-51}$$

because if $\mathbf{p}^*(t) = \mathbf{0}$, this would imply that $\mathbf{p}^{*T}(t)\,\mathbf{b} = 0$, which contradicts Eq. (5.6-46). Thus, if Eq. (5.6-50) is to be satisfied the matrix

$$\left[\mathbf{b} \mid \mathbf{A}\mathbf{b} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{b}\right]^T\mathbf{A}^T$$

must be singular. For this matrix to be singular either $\mathbf{A}$ or $[\mathbf{b} \mid \mathbf{A}\mathbf{b} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{b}]$, or both must be singular.† Notice that even if the system is completely controllable, in which case $[\mathbf{b} \mid \mathbf{A}\mathbf{b} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{b}]$ is nonsingular, an interval of singularity can still occur if the matrix $\mathbf{A}$ is singular. Thus a *necessary condition* for a singular interval to exist is that either the system (5.6-41) is not completely controllable, or $\mathbf{A}$ is singular.

Necessary Conditions for Singular Intervals

So far, we have concentrated on one aspect of singular solutions—necessary conditions for their existence. We have considered only linear, fixed, single-input systems, but the procedure followed applies as well to systems

† Because determinant $[\mathbf{M}_1\mathbf{M}_2]$ = determinant $\mathbf{M}_1$ · determinant $\mathbf{M}_2$ if $\mathbf{M}_1$ and $\mathbf{M}_2$ are square matrices; hence, determinant $[\mathbf{M}_1\mathbf{M}_2] = 0$ implies determinant $\mathbf{M}_1 = 0$, or determinant $\mathbf{M}_2 = 0$, or both.

that have several inputs or are nonlinear. The idea is quite straightforward: examine the Hamiltonian to determine whether there are situations in which the minimum principle does not yield sufficient information to determine the relationship between $\mathbf{u}^*(t)$, $\mathbf{x}^*(t)$, and $\mathbf{p}^*(t)$. If this situation occurs, use the fact that the Hamiltonian must be zero† (and that $\dot{\mathscr{H}}$, $\ddot{\mathscr{H}}$, ... equal zero) to determine other necessary conditions for the existence of singular intervals.

Effects of Singular Intervals on Problem Solution

Let us now consider an example that illustrates another facet of singular problems—the effects of singular intervals on problem solution.

Example 5.6-3. Find the control law that causes the response of the system

$$\dot{x}_1(t) = x_2(t) \tag{5.6-52a}$$

$$\dot{x}_2(t) = u(t) \tag{5.6-52b}$$

to minimize the performance measure

$$J = \tfrac{1}{2}\int_0^{t_f} [x_1^2(t) + x_2^2(t)]\, dt. \tag{5.6-53}$$

The final time t_f and the final states are free, and the controls are constrained by the inequality

$$|u(t)| \leq 1.0. \tag{5.6-54}$$

The Hamiltonian is given by

$$\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) = \tfrac{1}{2}x_1^2(t) + \tfrac{1}{2}x_2^2(t) + p_1(t)x_2(t) + p_2(t)u(t). \tag{5.6-55}$$

From the minimum principle and (5.6-55)

$$p_2^*(t)u^*(t) \leq p_2^*(t)u(t) \tag{5.6-56}$$

for all admissible $u(t)$ and for all $t \in [0, t_f]$. For $p_2^*(t) \neq 0$, Eq. (5.6-56) indicates that

$$u^*(t) = \begin{cases} -1.0, & \text{for } p_2^*(t) > 0 \\ +1.0, & \text{for } p_2^*(t) < 0. \end{cases} \tag{5.6-57}$$

Switchings of the optimal control occur at isolated instants when $p_2^*(t) = 0$. On the other hand, if there is a time interval $[t_1, t_2]$ during which

$$p_2^*(t) = 0 \qquad \text{for all } t \in [t_1, t_2], \tag{5.6-58}$$

† We assume free final time and $\mathscr{H}$ explicitly independent of time.

then $[t_1, t_2]$ is a singular interval; let us investigate this possibility.

Since the final time is free, and time does not appear explicitly in the Hamiltonian, it is necessary that

$$\tfrac{1}{2}x_1^{*2}(t) + \tfrac{1}{2}x_2^{*2}(t) + p_1^*(t)x_2^*(t) + p_2^*(t)u^*(t) = 0 \tag{5.6-59}$$

for $t \in [0, t_f]$. If $p_2^*(t) = 0$ for $t \in [t_1, t_2]$, then

$$p_2^*(t) = \dot{p}_2^*(t) = \ddot{p}_2^*(t) = \cdots = 0, \; t \in [t_1, t_2]. \tag{5.6-60}$$

In addition, from Eq. (5.6-59) we have

$$M \triangleq \tfrac{1}{2}x_1^{*2}(t) + \tfrac{1}{2}x_2^{*2}(t) + p_1^*(t)x_2^*(t) = 0 \tag{5.6-61}$$

for $t \in [t_1, t_2]$, and hence

$$M = \dot{M} = \ddot{M} = \cdots = 0, t \in [t_1, t_2], \tag{5.6-62}$$

if a singular interval is to exist.

The costate equations are

$$\dot{p}_1^*(t) = -x_1^*(t) \tag{5.6-63}$$

$$\dot{p}_2^*(t) = -x_2^*(t) - p_1^*(t). \tag{5.6-64}$$

During a singular interval, using Eqs. (5.6-60) and (5.6-64), we obtain

$$p_1^*(t) = -x_2^*(t). \tag{5.6-65}$$

Substituting this in (5.6-61) yields

$$x_1^{*2}(t) - x_2^{*2}(t) = 0 \tag{5.6-66}$$

or

$$[x_1^*(t) + x_2^*(t)][x_1^*(t) - x_2^*(t)] = 0, \qquad \text{for } t \in [t_1, t_2]. \tag{5.6-66a}$$

Equation (5.6-66) is satisfied if

$$x_1^*(t) + x_2^*(t) = 0 \tag{5.6-67a}$$

or if

$$x_1^*(t) - x_2^*(t) = 0, \qquad \text{for } t \in [t_1, t_2]. \tag{5.6-67b}$$

By differentiating Eq. (5.6-67a) and substituting in the state equation (5.6-52a) we find that

$$\dot{x}_1^*(t) = -\dot{x}_2^*(t) = x_2^*(t), \tag{5.6-68}$$

which with (5.6-52b) implies

$$u^*(t) = -x_2^*(t), \qquad \text{for } t \in [t_1, t_2]. \tag{5.6-69a}$$

Similarly, differentiating Eq. (5.6-67b) and substituting in the state equations, we obtain

$$u^*(t) = +x_2^*(t), \qquad \text{for } t \in [t_1, t_2]. \tag{5.6-69b}$$

Equations (5.6-67) define a locus of points in the state plane where singular controls may exist, and Eq. (5.6-69) gives an explicit expression for the singular control law. The singular lines, truncated at $|x_2(t)| = 1$, because $|u(t)| \leq 1$, are shown in Fig. 5-40. The arrows indicate the direction of increasing time.

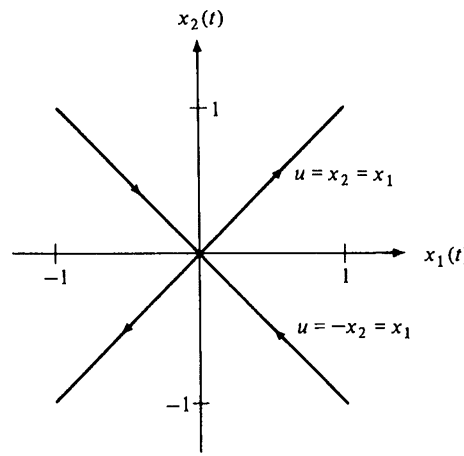


Figure 5-40 The singular lines for Example 5.6-3

We have determined two lines in the state plane where the control, states, and costates all satisfy the necessary conditions given by the minimum principle and the requirement that $\mathscr{H} \equiv 0$ on an extremal trajectory. Clearly, since the system moves away from the origin on the line $x_1 = x_2$, this segment cannot be part of an optimal trajectory. We still must determine the optimal control law for states not on the singular line, and also if the singular control law is optimal. Let us investigate some of the possibilities.

Suppose that at $t = 0$ the system is at state $\mathbf{x}_0$ shown in Fig. 5-41. The optimal control must be ± 1, because the system is not on the singular line. By examining the trajectories for this system with $u = \pm 1$, shown in Fig. 5-20, Section 5.4, it is clear that the optimal control should initially be $u^* = -1$. With this control the system trajectory is as shown in Fig. 5-41. We next ask the question: what happens when the trajectory intersects the singular line? Is the optimal control the one that keeps the system on the singular line, or should the control continue to be $u = -1$

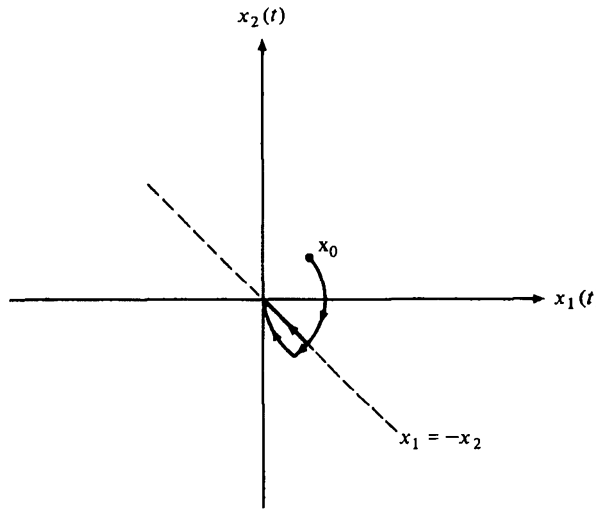


Figure 5-41 Optimal trajectory candidates for Example 5.6-3

until intersecting the curve from which the origin is reached by applying $u = +1$? To answer this question, consider what happens when a control switching is indicated. If u^* switches from $+1$ to -1 at some time t_1, then it follows that

$$\begin{aligned} p_2^*(t_1) &= 0 \\ \dot{p}_2^*(t_1) &> 0, \end{aligned} \tag{5.6-70}$$

or, if u^* switches from -1 to $+1$ at time t_1, then this implies

$$\begin{aligned} p_2^*(t_1) &= 0 \\ \dot{p}_2^*(t_1) &< 0. \end{aligned} \tag{5.6-71}$$

Now, since $\mathscr{H} \equiv 0$, $p_2^*(t_1) = 0$ implies that

$$p_1^*(t_1) = \frac{-\frac{1}{2}x_1^{*2}(t_1) - \frac{1}{2}x_2^{*2}(t_1)}{x_2^*(t_1)}. \tag{5.6-72}$$

Substituting this expression into the costate equation (5.6-64) gives

$$\dot{p}_2^*(t_1) = \frac{\frac{1}{2}[x_1^*(t_1) + x_2^*(t_1)][x_1^*(t_1) - x_2^*(t_1)]}{x_2^*(t_1)}. \tag{5.6-73}$$

By determining the sign of $\dot{p}_2^*(t_1)$ indicated by Eq. (5.6-73) for various regions in the state plane, we then know the allowable switchings that may occur. Table 5-3 shows how the sign of $\dot{p}_2^*(t_1)$ is determined for the regions of the state plane, and Fig. 5-42 illustrates these regions and the allowable switchings.

Table 5-3 DETERMINATION OF ALLOWABLE SWITCHINGS FOR REGIONS OF THE STATE PLANE

Region	Sign of $x_1(t_1)$	Sign of $x_2(t_1)$	Sign of $x_1(t_1) + x_2(t_1)$	Sign of $x_1(t_1) - x_2(t_1)$	Sign of $\dot{p}_2(t_1)$
R_1	+	+	+	−	−
R_2	+	+	+	+	+
R_3	+	−	+	+	−
R_4	+	−	−	+	+
R_5	−	−	−	+	+
R_6	−	−	−	−	−
R_7	−	+	−	−	+
R_8	−	+	+	−	−

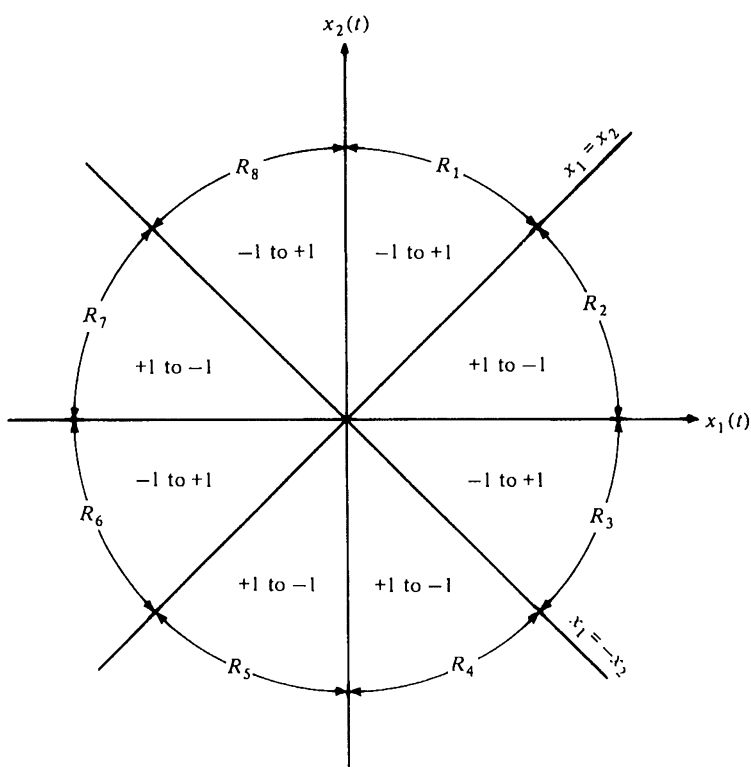


Figure 5-42 Allowable switchings in various regions of the state plane

Referring to Fig. 5-42, we see that if the trajectory in Fig. 5-41 is allowed to cross the singular line it is then in a region where switching from $u = -1$ to $u = +1$ violates the necessary condition that $\mathscr{H} \equiv 0$. We conclude then that the optimal trajectory beginning at this value of $\mathbf{x}_0$ must have its terminal segment on the singular line.

When an initial trajectory segment with $u^* = \pm 1$ does not intersect the singular line with $|x_2(t)| \leq 1$, then the optimal control will switch to $u^* = \mp 1$ and the optimal trajectory will eventually reach either the origin or the singular line. To determine where the switching occurs, let t_2 be the time when the trajectory reaches the singular line or the origin. Notice that the origin lies on the singular line, and from (5.6-60)

$$p_2(t_2) = 0. \tag{5.6-74}$$

Solving for the value of $p_1(t_2)$ on the line $x_1(t_2) = -x_2(t_2)$, which satisfies Eq. (5.6-59), gives

$$p_1(t_2) = -x_2(t_2). \tag{5.6-75}$$

Using the values of the costates given by (5.6-74) and (5.6-75) as initial conditions, and integrating the state and costate equations backward in time with $u = \pm 1$, we can determine the locations in the state plane where $p_2(t)$ again passes through zero. Doing this for several values of $x_2(t_2)$ (including zero) on the singular line, we obtain a locus of points that defines the switching curve *C*-*D*-0-*E*-*F* shown in Fig. 5-43. The optimal control law is given by

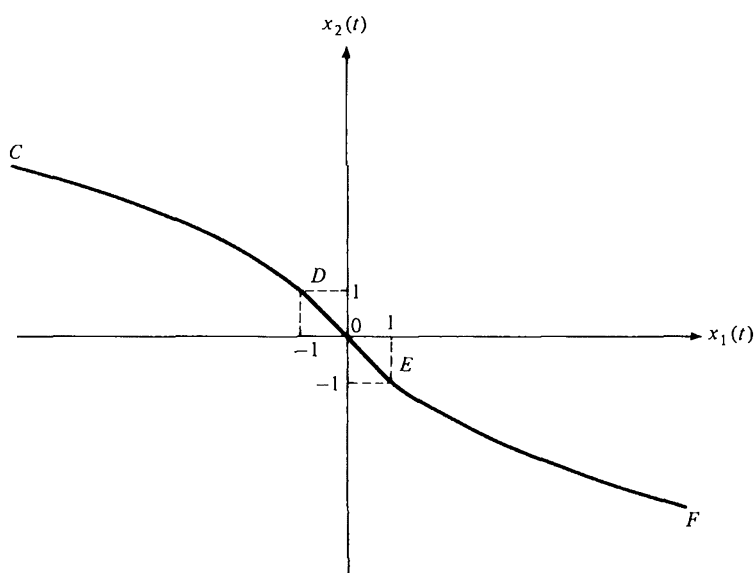


Figure 5-43 The optimal switching curve for Example 5.6-3

$$u^*(t) = \begin{cases} -1, & \text{for } \mathbf{x}(t) \text{ to the right of } C\text{-}0\text{-}F \\ +1, & \text{for } \mathbf{x}(t) \text{ to the left of } C\text{-}0\text{-}F \\ -1, & \text{for } \mathbf{x}(t) \text{ on segment } C\text{-}D \\ +1, & \text{for } \mathbf{x}(t) \text{ on segment } E\text{-}F \\ -x_2(t), & \text{for } \mathbf{x}(t) \text{ on segment } D\text{-}0\text{-}E. \end{cases} \tag{5.6-76}$$

Several optimal trajectories are pictured in Fig. 5-44; notice that *the switching curve is not a trajectory* except on the singular line D-0-E. As further illustration of this point, Fig. 5-45 shows the optimal switching curve, the curve $x_1 = \frac{1}{2}x_2^2$, which is the switching curve for bang-bang operation, the curve $x_1 = \frac{1}{2}x_2^2 + \frac{1}{2}$, which is the $u = +1$ trajectory that intersects the singular line at the point $(1, -1)$, and the line $x_1 = -x_2$. Observe that the optimal switching curve is above the line $x_1 = -x_2$ for all positive values of x_1; therefore, the switchings that occur on segment E-F do not violate the allowable switchings indicated in Fig. 5-42. Similarly, it can be verified that segment C-D of the switching curve lies entirely in region R_7 of the state plane, and so does not cause the allowable switching conditions to be violated.

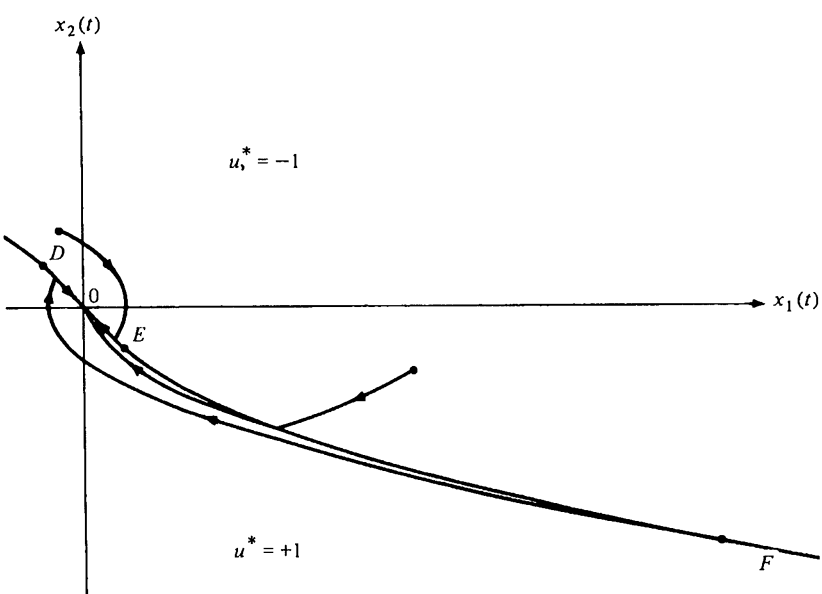


Figure 5-44 Some optimal trajectories for Example 5.6-3

Summary

The existence of singular intervals, although complicating the solution of optimal control problems, may turn out to be helpful in other respects.

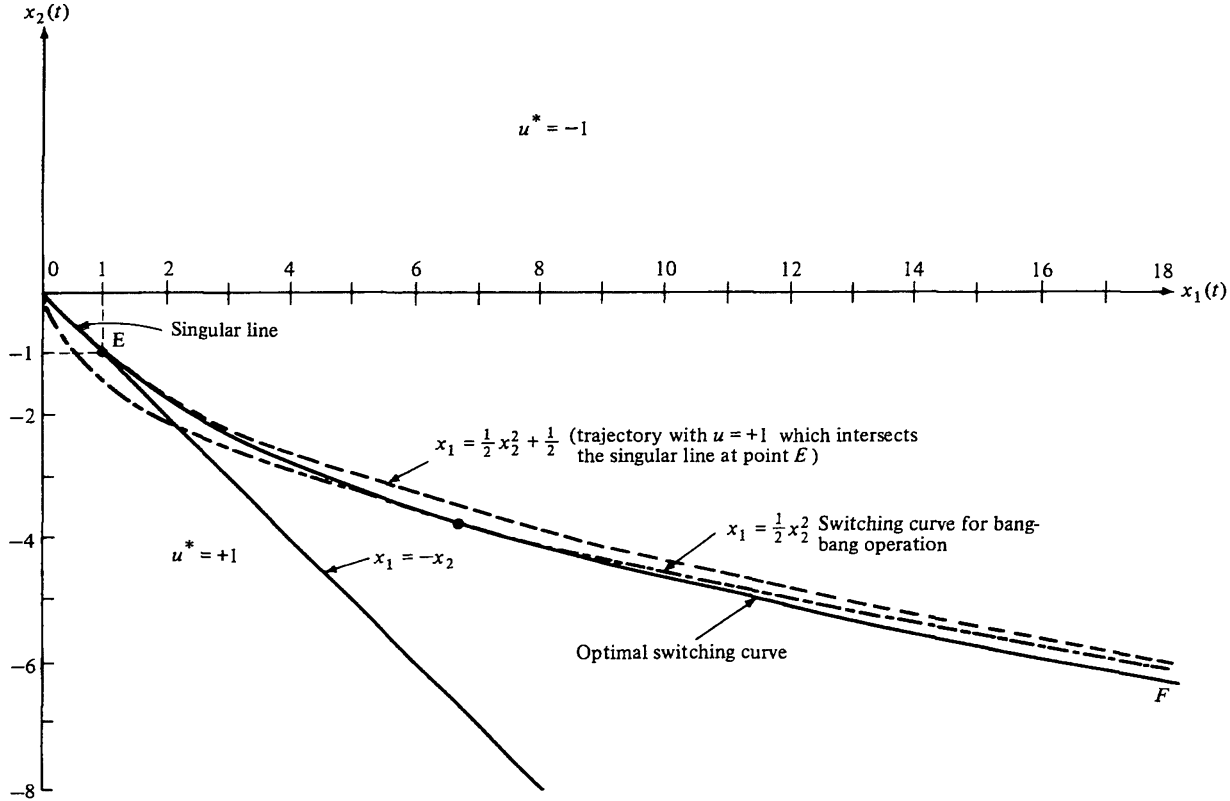


Figure 5-45 The optimal switching curve and three suboptimal alternatives

For example, a singular interval may indicate that an optimal control is nonunique; in this case we can select the optimal control which is easiest to implement, or which has other desirable features.

Our discussion has emphasized the following aspects of singular problems:

1. The determination of necessary conditions for the existence of singular intervals.
2. The use of these necessary conditions to find the regions in the state space where a singular control law exists.
3. The investigation of the singular control law to ascertain whether or not it is optimal.

The reader interested in additional material on singular intervals should refer to [A-2], [A-3], [J-1], [J-2], [R-2], [R-3], and [S-4].

5.7 SUMMARY AND CONCLUSIONS

In this chapter we have discussed the application of variational techniques to optimal control problems. The calculus of variations was used to derive a set of necessary conditions that must be satisfied by an optimal control and its associated state-costate trajectory. These necessary conditions for optimality lead to a (generally nonlinear) two-point boundary-value problem that must be solved to determine an explicit expression for the optimal control. In linear regulator problems, the resulting two-point boundary-value problem is linear and can be solved to obtain a linear, time-varying optimal control law.

Motivated by an interest in problems with bounded control or state variables, we then gave a heuristic derivation of Pontryagin's minimum principle and discussed a technique for dealing with state inequality constraints. The remainder of the chapter was concerned with applications of Pontryagin's minimum principle to problems with bounded admissible controls. Several examples of minimum-time and minimum-control-effort systems were discussed. These examples were elementary, but nonetheless indicative of procedures that are useful in obtaining optimal control laws. Finally, we investigated the occurrence of singular intervals during which the minimum principle fails to yield a relationship for the extremal control in terms of the extremal state-costate trajectory.

This chapter was not intended to be a handbook of solutions to optimal control problems. Indeed, the difficulties encountered should make the reader aware that no such handbook exists. We may regard the linear regulator problem as being solved; however, in the sections on minimum-time and minimum-control-effort problems we found that analytical solutions are

generally impossible for higher-order systems ($n \geq 3$) even if the systems are linear and time-invariant. For nonlinear systems it is even more difficult to obtain closed-form expressions for the optimal control laws.

Realistically, then, we must view the minimum principle as a starting point for obtaining numerical solutions to optimal control problems. From the minimum principle we obtain knowledge of the *form* of the optimal control (if it exists) and a statement of the two-point boundary-value problem, which, when solved, yields an explicit relationship for the optimal control.

REFERENCES

A-2 Athans, M., and P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications.* New York: McGraw-Hill Book Company, 1966.

A-3 Athans, M., "On the Uniqueness of the Extremal Controls for a Class of Minimum Fuel Problems," *IEEE Trans. Automatic Control* (1966), 660–668.

J-1 Johnson, C. D., and J. E. Gibson, "Optimal Control with Quadratic Performance Index and Fixed Terminal Time," *IEEE Trans. Automatic Control* (1964), 355–360.

J-2 Johnson, C. D., and J. E. Gibson, "Singular Solutions in Problems of Optimal Control," *IEEE Trans. Automatic Control* (1963), 4–15.

J-3 Johnson, C. D., "Singular Solutions in Problems of Optimal Control," *Advances in Control Systems: Theory and Applications*, Vol. 2, C. T. Leondes, ed. New York: Academic Press, Inc., 1965.

K-5 Kalman, R. E., "The Theory of Optimal Control and the Calculus of Variations," *Mathematical Optimization Techniques*, R. E. Bellman, ed. Santa Monica, Cal.: The RAND Corporation, 1963.

K-6 Kalman, R. E., "Mathematical Description of Linear Dynamical Systems," *J. SIAM Control*, series A (1963), 152–192.

K-7 Kalman, R. E., "Contributions to the Theory of Optimal Control," *Bol. Soc. Mat. Mex.* (1960), 102–119.

L-3 Leitmann, G., *An Introduction to Optimal Control.* New York: McGraw-Hill Book Company, 1966.

L-4 Leitmann, G., ed., *Optimization Techniques with Applications to Aerospace Systems.* New York: Academic Press, Inc., 1962.

M-2 Meditch, J. S., "On the Problem of Optimal Thrust Programming for a Lunar Soft Landing," *IEEE Trans. Automatic Control* (1964), 477–484.

M-3 Miele, A., "The Calculus of Variations in Applied Aerodynamic and Flight Mechanics," *Optimization Techniques with Applications to Aerospace Systems*, G. Leitmann, ed. New York: Academic Press, Inc., 1962.

P-1 Pontryagin, L. S., V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes.* New York: Interscience Publishers, Inc., 1962.

R-1 Rozonoer, L. I., "L. S. Pontryagin's Maximum Principle in the Theory of Optimum Systems I, II, III," *Automation and Remote Control* (1960), 1288–1302, 1405–1421, 1517–1532.

R-2 Rohrer, R. A., and M. Sobral, "Optimal Linear Switching for Singular Linear Systems," Report R-196, Coordinated Science Laboratory, University of Illinois, 1964.

R-3 Rohrer, R. A., and M. Sobral, "Optimal Singular Solutions for Linear Multi-Input Systems," *ASME Journal of Basic Engineering* (1966), 323–328.

S-3 Sage, A. P., *Optimum Systems Control.* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1968.

S-4 Sobral, M., "Linear Control Laws for Singular Linear Systems," Report R-188, Coordinated Science Laboratory, University of Illinois, 1964.

PROBLEMS

5-1. The boat shown in Fig. 5-P1 moves at a constant velocity v with respect to the water. The water moves in the positive y direction with known velocity $s(x)$ at the point x. The heading of the boat β is the control variable.

(a) Determine a set of state equations for the boat.

(b) Find necessary conditions for the boat to move from point 1 to point 2 in minimum time.

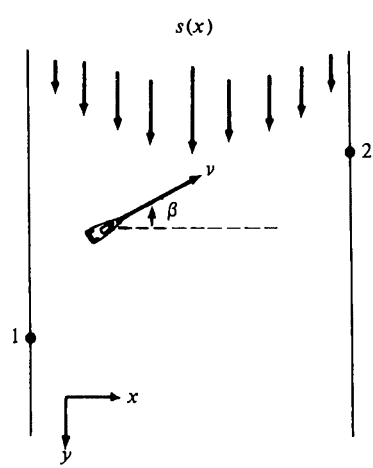


Figure 5-P1

(c) Suppose that the speed of the water is constant for all x and that point 2 and point 1 are in the relative locations shown. State a necessary condition for the existence of a minimum-time solution.
(d) If $s(x)$ is constant as in part (c), discuss the characteristics of the optimal steering angle β^*.

5-2. The system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -x_1(t) + [1 - x_1^2(t)]x_2(t) + u(t)$$

is to be controlled to minimize the performance measure

$$J = \int_0^1 \tfrac{1}{2}[2x_1^2(t) + x_2^2(t) + u^2(t)]\, dt.$$

The initial and final state values are specified.
(a) Determine the costate equations for the system.
(b) Determine the control that minimizes the Hamiltonian for:
(i) $u(t)$ not bounded.
(ii) $|u(t)| \leq 1.0$.

5-3. The system given in Problem 5-2 is to be transferred from the origin to the plane

$$15x_1(t) + 20x_2(t) + 12t = 60$$

while the performance measure

$$J = \tfrac{1}{2}\int_0^{t_f} u^2(t)\, dt$$

is minimized. The final time t_f is free.
(a) Determine the costate equations for the system.
(b) Find the control that minimizes $\mathscr{H}$ for
(i) $u(t)$ not bounded.
(ii) $-1.0 \leq u(t) \leq 2.0$.
(c) Determine the boundary conditions at $t = t_f$.

5-4. The system given in Problem 5-2 is to be transferred from the origin to the surface

$$[x_1(t) - 4]^2 + [x_2(t) - 5]^2 + [t - 2]^2 = 9$$

with minimum fuel expenditure. The final time is free, and $|u(t)| \leq 1.0$.
(a) Find the costate equations.
(b) Determine the control that minimizes the Hamiltonian.
(c) Determine the boundary conditions at $t = t_f$.

5-5. Assume that a nonlinear time-invariant system is to be transferred from a specified initial state $\mathbf{x}_0$ to a specified final state $\mathbf{x}_f$, and minimize the

performance measure $J = \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t))\, dt$. The admissible controls are not bounded and t_f is free. Show that the Hamiltonian is identically zero on an optimal trajectory.

Note. The assumption of unbounded admissible controls is not required, but the proof is more complicated for the bounded-control case.

5-6. A first-order system is described by the state equation

$$\dot{x}(t) = x(t) + u(t).$$

(a) Find the unconstrained control, in closed-loop form, which minimizes the functional

$$J = \int_0^T [1.5x^2(t) + 0.5u^2(t)]\, dt.$$

T is fixed, and $x(t_f)$ is free.

(b) Show that for $T \longrightarrow \infty$ the optimal control law is of the form

$$u^*(t) = Fx^*(t),$$

where F is a constant. Find F.

5-7. A linear first-order system is described by the differential equation

$$\dot{x}(t) = -ax(t) + u(t).$$

It is desired to bring the system from some arbitrary fixed initial state $x(0)$ to the origin in T seconds and minimize the performance measure

$$J = \int_0^T u^2(t)\, dt.$$

(a) Find the expression for the optimal trajectory in terms of $x(0)$, a, and T.

(b) Find the expression for the optimal control, in terms of $x(0)$, a, and T.

(c) The optimal control can be expressed in terms of $x(t)$ in the form

$$u^*(t) = F(t, T, a)x(t).$$

Find $F(t, T, a)$.

(d) Using physical reasoning and assuming $a > 0$, comment on the values of F and u^* as $t \longrightarrow T$. Also comment on F and u^* for the case where $T \longrightarrow \infty$.

5-8. For linear regulator problems discussed in Section 5.2, show that the control that is a solution of $\partial \mathscr{H}/\partial \mathbf{u} = \mathbf{0}$ satisfies a sufficient condition for minimization of the Hamiltonian.

5-9. (a) Show that for linear regulator problems discussed in Section 5.2 with $\mathbf{x}(t_f)$ free, the matrix $\mathbf{K}(t)$ satisfies the Riccati equation

$$\dot{\mathbf{K}}(t) = -\mathbf{K}(t)\mathbf{A}(t) - \mathbf{A}^T(t)\mathbf{K}(t) - \mathbf{Q}(t) + \mathbf{K}(t)\mathbf{B}(t)\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{K}(t)$$

with boundary conditions $\mathbf{K}(t_f) = \mathbf{H}$.

Hint. Differentiate $\mathbf{p}^*(t) = \mathbf{K}(t)\mathbf{x}^*(t)$.

(b) Show that the matrix $\mathbf{K}(t)$ is symmetric and hence only $n(n+1)/2$ differential equations need to be integrated to obtain $\mathbf{K}(t)$.

(c) What modifications are required in part (a) if $\mathbf{x}(t_f) = \mathbf{0}$?

5-10. Show that for linear regulator problems discussed in Section 5.2, if $\mathbf{x}(t_f) = \mathbf{0}$, the optimal control law is

$$\mathbf{u}^*(t) = \mathbf{R}^{-1}(t)\mathbf{B}^T(t)[\boldsymbol{\varphi}_{12}(t_f, t)]^{-1}\boldsymbol{\varphi}_{11}(t_f, t)\mathbf{x}(t).$$

5-11. (a) Determine the optimal control law for the system

$$\dot{x}(t) = -x(t) + u(t)$$

to be transferred to the origin from an arbitrary initial state. The performance measure is

$$J = \int_0^1 \tfrac{1}{2}[3x^2(t) + u^2(t)]\,dt.$$

The admissible controls are not bounded.

(b) Determine the optimal control law for the system and performance measure in part (a) with $x(1)$ free.

5-12. (a) Find the control that transfers the system given in Example 5.1-1 from $\mathbf{x}(0) = \mathbf{0}$ to the line $x_1(t) + 5x_2(t) = 15$ and minimizes

$$J = \tfrac{1}{2}[x_1(2) - 5]^2 + \tfrac{1}{2}[x_2(2) - 2]^2 + \tfrac{1}{2}\int_0^2 u^2(t)\,dt.$$

(b) Determine the cost of control

$$\tfrac{1}{2}\int_0^2 u^{*2}(t)\,dt$$

for part (a) above, and for parts (a) and (b) of Example 5.1-1. Compare the control costs and discuss the comparison qualitatively.

5-13. A differential equation that describes the leaky reservoir shown in Fig. 5-P13 is

$$\dot{x}(t) = -0.1x(t) + u(t),$$

where $x(t)$ is the height of the water, and $u(t)$ is the net inflow rate of water at time t. Assume that $0 \leq u(t) \leq M$.

(a) Find the optimal control law if it is desired to minimize

$$J = \int_0^{100} -x(t)\,dt.$$

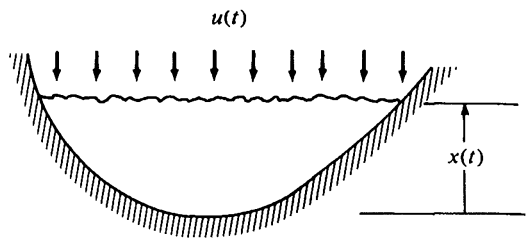


Figure 5-P13

(b) Repeat part (a) with the additional constraint that

$$\int_0^{100} u(t)\,dt = K \text{ (a known constant).}$$

(c) Determine the optimal control law if $J = -x(100)$, and

$$\int_0^{100} u(t)\,dt = K.$$

5-14. If the conditions for a time-invariant optimal control law are satisfied by a linear regulator problem, the constant **K** matrix must be a solution of the nonlinear algebraic equations

$$\mathbf{0} = -\mathbf{KA} - \mathbf{A}^T\mathbf{K} - \mathbf{Q} + \mathbf{KBR}^{-1}\mathbf{B}^T\mathbf{K}.$$

Using this result, determine the optimal control laws for:
(a) The first-order system $\dot{x}(t) = ax(t) + u(t)$ with performance measure

$$J = \int_0^{\infty} [qx^2(t) + ru^2(t)]\,dt, \qquad q, r > 0.$$

Show the variation of the pole of the closed-loop system for $0 < q/r < \infty$.
(b) The system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -4x_1(t) - 4x_2(t) + u(t)$$

and the performance measure

$$J = \int_0^{\infty} [20x_1^2(t) + 5x_2^2(t) + u^2(t)]\,dt.$$

Find the location of the poles of the controlled (closed-loop) system and compare with the pole locations for the open-loop system.

5-15. A set of state equations for the dc motor with constant armature current shown in Fig. 5-P15 is

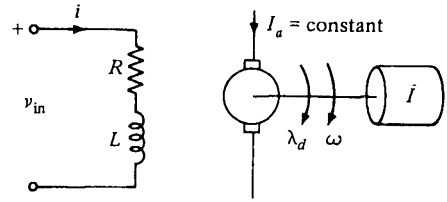


Figure 5-P15

$$\frac{di(t)}{dt} = -\frac{R}{L}i(t) + \frac{1}{L}v_{in}(t)$$

$$\frac{d\omega(t)}{dt} = \frac{K}{I}i(t).$$

$\lambda_d(t) = Ki(t)$ is the instantaneous torque developed; $K = k_t I_a$ is the product of the torque constant and the armature current, and I is the angular moment of inertia. The performance measure to be minimized is

$$J = \int_0^\infty [i^2(t) + \omega^2(t) + v_{in}^2(t)]\, dt,$$

and the admissible controls are not bounded. Assume that R, L, K, and I are equal to 1.0. Determine the optimal control law.

5-16. Consider a linear, completely observable system

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\,\mathbf{u}(t)$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t).$$

The performance measure to be minimized is

$$J = \tfrac{1}{2}\,\|\mathbf{y}(t_f)\|_{\mathbf{H}}^2 + \tfrac{1}{2}\int_{t_0}^{t_f} [\|\mathbf{y}(t)\|_{\mathbf{Q}(t)}^2 + \|\mathbf{u}(t)\|_{\mathbf{R}(t)}^2]\, dt.$$

$\mathbf{Q}(t)$ and $\mathbf{H}$ are real symmetric positive semi-definite matrices and $\mathbf{R}(t)$ is a real symmetric positive definite matrix. The admissible controls are not bounded, and t_f is specified.

(a) Show that this problem can be reduced to the form of linear regulator problems discussed in Section 5.2.

(b) Determine the optimal control law.

5-17. (a) The system

$$\dot{x}(t) = -x(t) + u(t)$$

is to be controlled to minimize the performance measure

$$J = \tfrac{1}{2}\int_0^{t_f} \{q[x(t) - r(t)]^2 + u^2(t)\}\, dt, \qquad q > 0.$$

Show that if $r(t) = \alpha\epsilon^{-t}$ (α is any real constant), this tracking problem can be put in the form of a linear regulator problem by defining a new state variable

$$\tilde{x}(t) \triangleq x(t) - r(t).$$

The admissible controls are not bounded, and t_f is specified.

(b) To generalize the result obtained in part (a), assume that the system is described by the nth-order differential equation

$$\frac{d^n}{dt^n}y(t) = -a_{n-1}\frac{d^{n-1}}{dt^{n-1}}y(t) - a_{n-2}\frac{d^{n-2}}{dt^{n-2}}y(t) - \cdots - a_0 y(t) + u(t).$$

The performance measure to be minimized is

$$J = \tfrac{1}{2}\int_0^{t_f}\{q[y(t) - r(t)]^2 + u^2(t)\}\,dt, \qquad q > 0.$$

Show that if $r(t)$ satisfies the differential equation

$$\left\{\frac{d^n}{dt^n} + a_{n-1}\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_0\right\}r(t) = 0,$$

then by defining the state variables as

$$x_i(t) = \frac{d^{i-1}}{dt^{i-1}}[y(t) - r(t)] \qquad i = 1, 2, \cdots, n$$

the problem can be reduced to a linear regulator problem. Find the optimal control law.

5-18. This problem is most easily done by using a digital computer.

(a) Determine the optimal control law for the attitude control of the spacecraft described in Example 2.2-1. Assume that $q_{11} = 4.0$, $q_{22} = 0.0$, $R = 1$, $x_1(0) = 10.0$, and $x_2(0) = 0.0$.

(b) Assume that $t_f = 10.0$ instead of ∞ and determine the optimal control law. Compare the optimal response obtained using this control law with the response obtained using the control law found in part (a).

(c) Repeat part (b) with $t_f = 4.0$.

5-19. Suppose that the missile in Example 5.4-1 is heading toward a target located at the fixed point $x = b$. If the pilot of the pursuing aircraft senses that he cannot protect point b, his orders are to look for other incoming missiles to pursue. You are to determine an algorithm for the pilot to use in deciding whether or not he can prevent the missile from reaching the point $x = b$.

(a) Find the expression in terms of a and b for the time required for the missile to reach point b.

(b) For interception to occur before the missile reaches b, the value of b must be greater than b_1. Find the relationship between a and b_1.

(c) Using a and b as axes, show the values for which:

(i) The missile is intercepted before reaching b.
(ii) Interception can be accomplished, but not before the missile reaches b.
(iii) Interception is impossible.

5-20. Using Pontryagin's minimum principle, verify the solution given for Example 1.1-4.

5-21. Using the results of Example 5.4-4, determine the optimal control law for transferring the system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = u(t)$$

from an arbitrary initial state to the point $[2, 2]^T$ in minimum time with $|u(t)| \leq 1.0$.

5-22. (a) Find the optimal control law for transferring the system

$$\dot{x}_1(t) = -x_1(t) - u(t)$$
$$\dot{x}_2(t) = -2x_2(t) - 2u(t)$$

from an arbitrary initial state to the origin in minimum time. The admissible controls are constrained by $|u(t)| \leq 1.0$.

(b) Generalize the results of part (a) to determine the time-optimal-control law for the system

$$\dot{x}_1(t) = a_1x_1(t) + a_1u(t)$$
$$\dot{x}_2(t) = a_2x_2(t) + a_2u(t).$$

Assume that $a_2 < a_1 < 0$.

5-23. Assume that the space vehicle shown in Fig. 5-P23 can be approximated by a particle of mass M, and

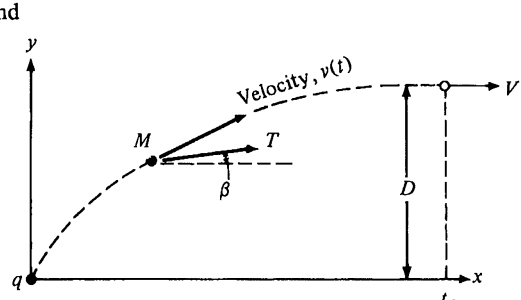


Figure 5-P23

(i) Aerodynamic and gravitational forces are negligible.
(ii) The motion is planar.

(iii) The mass of the vehicle is constant.
(iv) The thrust T is constant.
The thrust angle $\beta(t)$ is the control variable.
(a) Write state equations for the system.
(b) If the vehicle is to be transferred in minimum time from point q starting with zero velocity to an altitude D with vertical component of velocity equal to zero and horizontal velocity V, determine the costate equations and the required boundary condition relationships.
(c) Proceed with the solution of the problem in part (b) as far as possible using analytical methods.
(d) Repeat parts (b) and (c) if the vehicle is to maximize its horizontal range at the fixed final time $t_f = t_1$. The final altitude is again D.

5-24. (a) Consider the first-order system

$$\dot{x}(t) = 2x(t) + u(t)$$
$$|u(t)| \leq 1.$$

It is desired to transfer the system from an arbitrary initial state to the origin in minimum time; however, there are some states that cannot be transferred to the origin with any admissible control history, no matter how long is allowed. For these initial states an optimal solution does not exist.

Find the initial states for which there is no time-optimal control to the origin.

(b) Consider the system

$$\dot{x}_i(t) = a_i x_i(t) + b_i u(t)$$

with $|u(t)| \leq 1.0$, $b_i \neq 0$ for $i = 1, 2, \ldots, n$.
Find the initial states for which there is no time-optimal control to reach the origin.

5-25. Theorem 5.4-3 states that for a stationary, linear system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t)$$
$$|u(t)| \leq 1$$

to be transferred from an arbitrary initial state to the origin in minimum time, the optimal control, if it exists, is bang-bang and has at most $(n - 1)$ switchings if all of the eigenvalues of $\mathbf{A}$ are real. This problem points out why the eigenvalues of $\mathbf{A}$ must be real for the number of switchings to be at most $(n - 1)$. Consider the system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -2x_1(t) - 2x_2(t) + u(t)$$
$$|u(t)| \leq 1.$$

(a) Verify that the eigenvalues of A (the roots of the characteristic equation) are complex.
(b) Show that if an optimal control exists, it is always maximum effort; i.e., $|u(t)| = 1$.
(c) Show that the number of switchings may be more than $(n - 1)$ and indicate the conditions for which this will occur.

5-26. Attempt to determine the optimal control law for transferring the system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = x_3(t)$$
$$\dot{x}_3(t) = u(t)$$

from an arbitrary initial state to the origin in minimum time with $|u(t)| \leq 1.0$.

5-27. Suppose that in Example 5.1-2 the assumption was not made that the mass of the vehicle and the thrust are constant. In this case, the thrust is $T(t) = -k\dot{M}(t)$, where k is a known constant, and $\dot{M}(t)$ is the rate of change of mass. $\dot{M}(t)$ satisfies the constraint $-\eta \leq \dot{M}(t) \leq 0$ $(\eta > 0)$.
(a) Determine the modified state equations.
(b) Find the costate equations for minimum-time control of the vehicle.
(c) Determine a set of necessary conditions for the controls $T(t)$ and $\beta(t)$ to be optimal.
(d) Using Table 5-1, determine the required boundary condition relationships for missions *a*, *b*, *c*, *d* of Example 5.1-2.

5-28. Consider the system

$$\dot{x}(t) = ax(t) + u(t), \qquad a > 0,$$

which is to be transferred from an initial state x_0 to the origin. The admissible controls are constrained by $|u(t)| \leq 1.0$. Assume that x_0 is such that the origin can be reached by applying admissible controls.
(a) Determine the time-optimal control law.
(b) Determine the fuel-optimal control law.
(c) Compare the two optimal control laws and explain the difference between this problem and the case where $a < 0$.

5-29. The system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = x_1(t) + x_2(t)u(t)$$

is to be transferred from an initial state $\mathbf{x}_0$ to a target set $S(t)$ while

$$J = \int_0^{t_f} |u(t)|\, dt$$

is minimized.

The admissible controls are constrained by the relationship $|u(t)| \leq 1$. Assuming that an optimal control exists,

(a) Find the costate equations.

(b) Find an expression for the optimal control in terms of the extremal state and costate trajectories.

5-30. The system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), t) + \mathbf{B}(\mathbf{x}(t), t)\mathbf{u}(t)$$

is to be controlled to minimize the performance measure

$$J = \int_{t_0}^{t_f} [\lambda + \mathbf{u}^T(t)\mathbf{R}\mathbf{u}(t)]\, dt.$$

λ is ≥ 0, t_f is free, and $\mathbf{R}$ is a diagonal matrix with positive elements; i.e., $r_{ij} > 0$ for $i = j$, $r_{ij} = 0$ for $i \neq j$. Determine the form of the optimal control in terms of the extremal costates.

5-31. (a) Determine the costate equations and boundary conditions for the soft-lunar-landing vehicle described in Example 5.4-3.

(b) Determine necessary conditions for the landing to be accomplished with minimum fuel expenditure.

(c) Using the differential equation

$$\ddot{x}(t) = -g - \frac{k\dot{m}(t)}{m(t)} = -g - k\frac{d}{dt}[\ln m(t)]$$

show that if $\dot{x}(t_f) = 0$, the consumed fuel is a monotone increasing function of the final time t_f. What are the implications of this result?

5-32. Theorem 5.4-3 states that for a system of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$

to be transferred to the origin in minimum time there are at most $(n - 1)$ switchings of each control component if the eigenvalues of $\mathbf{A}$ are real. This can be proved by showing that $\mathbf{p}^{*T}(t)\mathbf{b}_i$ has at most $(n - 1)$ zeros ($\mathbf{b}_i$ is the ith column of $\mathbf{B}$). Use the results of Theorem 5.4-3 to determine the maximum number of switchings to transfer a system of this form to the origin using minimum fuel. Assume that the eigenvalues of $\mathbf{A}$ are real, and that an optimal control exists.

5-33. In Section 5.5 the performance measure

$$J_1 = \int_{t_0}^{t_f} \left[\lambda + \sum_{i=1}^{m} |u_i(t)|\right] dt$$

was used. There are other alternatives, however. For example, we could minimize elapsed time

$$J_2 = \int_{t_0}^{t_f} dt$$

subject to the constraint that the consumed fuel must satisfy

$$\int_{t_0}^{t_f} \left[\sum_{i=1}^{m} |u_i(t)| \right] dt \leq F \text{ (a constant)},$$

or we could minimize the consumed fuel

$$J_3 = \int_{t_0}^{t_f} \left[\sum_{i=1}^{m} |u_i(t)| \right] dt$$

with the constraint that $(t_f - t_o) \leq T$ (a constant).

(a) Determine the Hamiltonians for J_1, J_2, and J_3, assuming that the form of the state equations is

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), t) + \mathbf{B}(\mathbf{x}(t), t)\mathbf{u}(t),$$

and that $|u_i(t)| \leq 1$, $i = 1, 2, \ldots, m$.

(b) Compare the form of the costate equations for each of these performance measures.

5-34. Consider the system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -ax_2(t) + u(t),$$

where $a > 0$ and $|u(t)| \leq 1.0$. The system is to be transferred to the origin while minimizing the performance measure

$$J = \int_0^{t_f} [\lambda + |u(t)|] \, dt.$$

The final time is free and $\lambda > 0$.

(a) Determine the costate equations and the control that minimizes $\mathscr{H}$.

(b) What are the possible optimal control sequences?

(c) Show that a singular interval cannot exist.

(d) Determine the optimal control law.

5-35. A body M moving in a viscous fluid is shown in Fig. 5-P35. The controls u_1 and u_2 are thrusts in the x and y directions. Assume that the magnitude

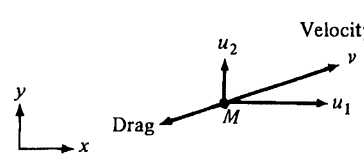


Figure 5-P35

of the drag force is $\alpha \cdot$[velocity of M]^2; the direction is opposite to the instantaneous velocity vector. If we assume planar motion and constant mass, and define $x_1 \triangleq x$, $x_2 \triangleq \dot{x}$, $x_3 \triangleq y$, $x_4 \triangleq \dot{y}$, the state equations are

$$\begin{aligned}
\dot{x}_1(t) &= x_2(t) \\
\dot{x}_2(t) &= -\alpha x_2(t)[x_2^2(t) + x_4^2(t)]^{1/2} + u_1(t) \\
\dot{x}_3(t) &= x_4(t) \\
\dot{x}_4(t) &= -\alpha x_4(t)[x_2^2(t) + x_4^2(t)]^{1/2} + u_2(t).
\end{aligned}$$

The system is to be transferred from the initial state $\mathbf{x}(0) = \mathbf{0}$ to $x_1(t_f) = e_1$ and $x_3(t_f) = e_3$ in minimum time. $x_2(t_f)$ and $x_4(t_f)$ are unspecified.

(a) Determine the costate equations.

(b) What are the required boundary conditions at $t = t_f$?

Hint. In parts (c) and (d) you will find it helpful to:

(i) Show the admissible control region in a two-dimensional picture.

(ii) Try to interpret minimization of $\mathscr{H}$ with respect to $\mathbf{u}$ geometrically.

(c) Determine an expression for the optimal control in terms of $\mathbf{x}^*(t)$, $\mathbf{p}^*(t)$ if the admissible controls are constrained by

$$u_1^2(t) + u_2^2(t) \leq 1.$$

(d) Repeat part (c) for the control constraints

$$|u_1(t)| + |u_2(t)| \leq 1.$$

(e) Which set of admissible controls, c or d, would you expect to yield a smaller value of the performance measure and why?

5-36. Consider the system discussed in Examples 5.5-5 and 5.6-2 with the performance measure

$$J = \int_0^{t_f} |u(t)|\, dt$$

and t_f free.

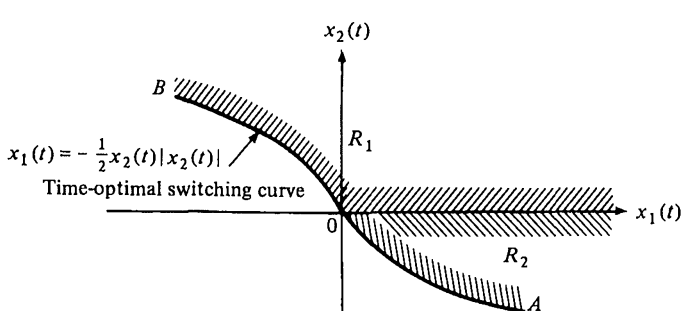


Figure 5-P36

(a) Determine the costate equations.
(b) If the initial state is in the region R_2 (not including the curve A-0) shown in Fig. 5-P36, show that the fuel-optimal control is not unique.
(c) If the initial state is in the region R_1 (not including the curve 0-B), show that a fuel-optimal control to the origin does not exist.
(d) Investigate the possibility of singular control intervals.

5-37. Consider a rocket in horizontal flight as shown in Fig. 5-P37. Assume that gravitational acceleration is constant, that the weight of the rocket is exactly balanced by the lift, and that the aerodynamic drag force is given by

$$D \triangleq \alpha x_1^2(t) + \frac{\beta x_2^2(t)}{x_1^2(t)} > 0,$$

where

$x_1 \triangleq$ the horizontal velocity.
$x_2 \triangleq m$, the mass of the rocket.
α and β are positive constants.

If we let $u(t) = -\dot{m}(t)$, the state equations are

$$\dot{x}_1(t) = \frac{cu(t)}{x_2(t)} - \frac{D}{x_2(t)}$$
$$\dot{x}_2(t) = -u(t).$$

c is the effective exhaust gas speed, a positive constant, and $0 \leq u(t) \leq u_{\max}$. It is desired to *maximize* the range of the rocket. The initial and final values of mass and velocity are specified, and the terminal time is free.
(a) Determine the costate equations and the boundary condition relationships.
(b) Investigate the possibility of singular control intervals.

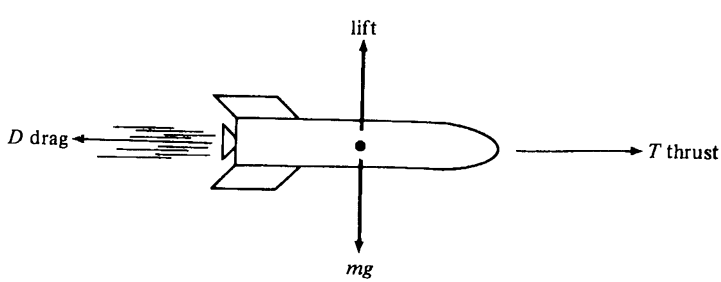


Figure 5-P37

5-38. The state equations for a linear system are

$$\dot{x}_1(t) = x_2(t) + u(t)$$
$$\dot{x}_2(t) = -u(t),$$

where $|u(t)| \leq 1$. The system is to minimize the performance measure

$$J = \int_0^{t_f} \tfrac{1}{2}[x_1^2(t)]\, dt.$$

The final time t_f is free and $\mathbf{x}(t_f) = \mathbf{0}$.

(a) Determine the costate equations and the required boundary conditions.

(b) Investigate the possibility of singular control intervals.

5-39. Consider the minimum-time control of a system of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t);$$

$\mathbf{A}$ and $\mathbf{B}$ are constant matrices, and $|u_i(t)| \leq 1$, $i = 1, 2, \ldots, m$.

Show that for a singular interval to exist it is necessary that the matrix

$$\mathbf{E}_j \triangleq \left[\mathbf{b}_j \;\middle|\; \mathbf{A}\mathbf{b}_j \;\middle|\; \cdots \;\middle|\; \mathbf{A}^{n-1}\mathbf{b}_j\right]$$

be singular for at least one value of $j (j = 1, 2, \ldots, m)$. $\mathbf{b}_j$ denotes the jth column of $\mathbf{B}$.

5-40. Investigate Example 5.4-3 to determine whether or not there can be any singular intervals.

5-41. (a) Show that if the performance measure

$$J = \int_0^{t_f} [\lambda + |u(t)|]\, dt, \qquad \lambda > 0$$

is used in Example 5.6-2, no singular intervals can exist.

(b) Attempt to generalize the results of part (a) to an arbitrary, stationary, single-input, linear system.

5-42. The system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = u(t)$$

is to be transferred to the origin in minimum time with admissible controls satisfying $|u(t)| \leq 1$. In addition, it is required that $|x_2(t)| \leq 2$ for $t \in [0, t^*]$.

(a) Determine a set of necessary conditions for optimal control.

(b) Determine the optimal control law.

5-43. The mass M described in Problem 5-35 is to be transferred from an arbitrary initial state to the target set

$$2x_1^2(t) + t^2x_1(t)x_2(t) + 5x_2^2(t) - 9 = 0$$

in minimum time.

(a) If the admissible controls are constrained by

$$\begin{aligned} |u_1(t)| &\leq 1 \\ |u_2(t)| &\leq 1 \\ |u_1(t)| + |u_2(t)| &\leq 1.5, \end{aligned}$$

show the admissible control region on a sketch of the u_1, u_2 plane

(b) Determine $\mathbf{u}^*(t)$ in terms of the extremal state and costate variables.

(c) Determine the boundary condition equations at $t = t_f$.

IV

Iterative Numerical Techniques for Finding Optimal Controls and Trajectories

6

Numerical Determination of Optimal Trajectories

In Chapter 5 variational techniques were used to derive necessary conditions for optimal control. In problems with linear plant dynamics and quadratic performance criteria (linear regulator and linear tracking systems), it was found that it is possible to obtain the optimal control *law* by numerically integrating a matrix differential equation of the Riccati type. Optimal control laws were also determined for several other simple examples by applying Pontryagin's minimum principle. In general, however, the variational approach leads to a nonlinear two-point boundary-value problem that cannot be solved analytically to obtain the optimal control law, or even an optimal open-loop control. Indeed, the difficulty of solving nonlinear two-point boundary-value problems analytically accounts for the fact that although many of the important variational concepts have been known for some time, only since the advent of digital computers have variational techniques been successfully applied to complex physical problems.

In this chapter we shall discuss four iterative numerical techniques for determining optimal controls and trajectories. Three of these techniques, steepest descent, variation of extremals, and quasilinearization, are procedures for solving nonlinear two-point boundary-value problems. The fourth technique, gradient projection, does not make use of the necessary conditions for optimality provided by the variational approach. Instead, the optimization problem is solved by minimizing a function of several variables subject to various constraining relationships. Each of these tech-

niques determines an open-loop optimal control, that is, the optimal control history associated with a specified set of initial conditions.

6.1 TWO-POINT BOUNDARY-VALUE PROBLEMS

Assuming that the state and control variables are not constrained by any boundaries, that the final time t_f is fixed, and that $\mathbf{x}(t_f)$ is free, we can summarize the two-point boundary-value problem that results from the variational approach by the equations

$$\dot{\mathbf{x}}^*(t) = \frac{\partial \mathscr{H}}{\partial \mathbf{p}} = \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \tag{6.1-1}$$

$$\dot{\mathbf{p}}^*(t) = -\frac{\partial \mathscr{H}}{\partial \mathbf{x}} = -\left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T \mathbf{p}^*(t) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \tag{6.1-2}$$

$$\mathbf{0} = \frac{\partial \mathscr{H}}{\partial \mathbf{u}} = \left[\frac{\partial \mathbf{a}}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T \mathbf{p}^*(t) + \frac{\partial g}{\partial \mathbf{u}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \tag{6.1-3}$$

$$\mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{6.1-4a}$$

$$\mathbf{p}^*(t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f)).\dagger \tag{6.1-4b}$$

From these five sets of conditions it is desired to obtain an explicit relationship for $\mathbf{x}^*(t)$ and $\mathbf{u}^*(t)$, $t \in [t_0, t_f]$. Notice that the expressions for $\mathbf{x}^*(t)$ and $\mathbf{u}^*(t)$ will be implicitly dependent on the initial state $\mathbf{x}_0$.

Let us assume that Eq. (6.1-3) can be solved to obtain an expression for $\mathbf{u}^*(t)$ in terms of $\mathbf{x}^*(t)$, $\mathbf{p}^*(t)$, and t; that is,

$$\mathbf{u}^*(t) = \mathbf{f}(\mathbf{x}^*(t), \mathbf{p}^*(t), t). \tag{6.1-5}$$

If this expression is substituted into Eqs. (6.1-1) and (6.1-2), we have a set of $2n$ first-order ordinary differential equations (called the *reduced differential equations*) involving only $\mathbf{x}^*(t)$, $\mathbf{p}^*(t)$, and t. The boundary conditions for these differential equations (which are generally nonlinear) are given by Eq. (6.1-4).

If the boundary conditions were all known at either t_0 or t_f, we could

† In the following discussion the * will be used only if all of Eqs. (6.1-1) through (6.1-4) are satisfied by a trajectory.

numerically integrate the reduced differential equations to obtain $\mathbf{x}^*(t)$, $\mathbf{p}^*(t)$, $t \in [t_0, t_f]$. The optimal control history could then be found by substituting $\mathbf{x}^*(t)$, $\mathbf{p}^*(t)$ into (6.1-5). Unfortunately, the boundary values are split, so this method cannot be applied. We hasten to point out that if the reduced differential equations are *linear*, the principle of superposition can be used to circumvent the complications caused by the split boundary values.† Thus, the difficulty of solving optimal control problems by using variational principles is caused by the combination of split boundary values and nonlinear differential equations.

Let us now discuss three iterative numerical techniques that have been used to solve nonlinear two-point boundary-value problems. The reader will notice that each of these techniques is based on the following general procedure:

An initial guess is used to obtain the solution to a problem in which one or more of the five necessary conditions (6.1-1) through (6.1-4) is not satisfied. This solution is then used to adjust the initial guess in an attempt to make the next solution come "closer" to satisfying all of the necessary conditions. If these steps are repeated and the iterative procedure converges, the necessary conditions (6.1-1) through (6.1-4) will eventually be satisfied.

6.2 THE METHOD OF STEEPEST DESCENT

Minimization of Functions by Steepest Descent

Let us begin our discussion of the method of steepest descent (or gradients) by considering an analogous calculus problem. Let f be a function of two independent variables y_1 and y_2; the value of the function at the point y_1, y_2 is denoted by $f(y_1, y_2)$. It is desired to find the point y_1^*, y_2^*, where f assumes its minimum value, $f(y_1^*, y_2^*)$.

If it is assumed that the variables y_1 and y_2 are not constrained by any boundaries, a necessary condition for y_1^*, y_2^* to be a point where f has a (relative) minimum is that the differential of f vanish at y_1^*, y_2^*, that is,

$$df(y_1^*, y_2^*) = \left[\frac{\partial f}{\partial y_1}(y_1^*, y_2^*)\right] \Delta y_1 + \left[\frac{\partial f}{\partial y_2}(y_1^*, y_2^*)\right] \Delta y_2$$
$$\triangleq \left[\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^*)\right]^T \Delta \mathbf{y} = 0. \tag{6.2-1}$$

$\partial f/\partial \mathbf{y}$ is called the gradient of f with respect to $\mathbf{y}$. Since y_1 and y_2 are independent, the components of $\Delta \mathbf{y}$ are independently arbitrary and (6.2-1)

† This matter is discussed in more detail in Section 6.4.

implies

$$\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^*) = \mathbf{0}. \tag{6.2-2}$$

In other words, for $f(\mathbf{y}^*)$ to be a relative minimum it is necessary that the gradient of f be zero at the point $\mathbf{y}^*$. Equation (6.2-2) represents two algebraic equations that are generally nonlinear. Suppose that these algebraic equations cannot be solved analytically for $\mathbf{y}^*$; how else might $\mathbf{y}^*$ be determined?

One possible approach is to visualize the minimization as a problem in hill climbing. Let us think of the function f as defining hills and valleys in the three-dimensional $y_1, y_2, f(y_1, y_2)$ space. One way to find the bottom of a valley is to pick a trial point $\mathbf{y}^{(0)}$ and climb in a downward direction until a point $\mathbf{y}^*$ is reached where moving in any direction increases the function value.† To make the climbing procedure efficient, we elect to climb in the direction of steepest descent, thus ensuring that the shortest distance is traveled in reaching the bottom of the hill. The direction of steepest descent at $\mathbf{y}^{(0)}$ is determined by evaluating the slope, or gradient, of the hill at the point $\mathbf{y}^{(0)}$. As shown in Fig. 6-1, the gradient vector is normal to the elevation contour. $\mathbf{z}(\mathbf{y}^{(0)})$ is the unit vector in the gradient direction at the point $\mathbf{y}^{(0)}$; that is,

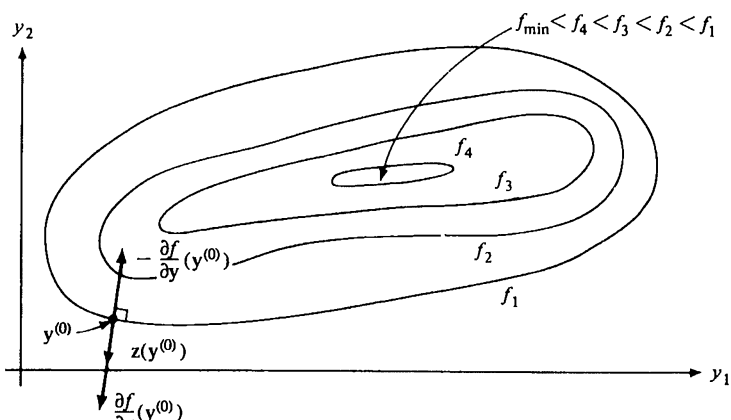


Figure 6-1 The gradient and several equal value contours of f

$$\mathbf{z}(\mathbf{y}^{(0)}) \triangleq \frac{\dfrac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^{(0)})}{\left\| \dfrac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^{(0)}) \right\|} = \frac{\dfrac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^{(0)})}{\sqrt{\left[\dfrac{\partial f}{\partial y_1}(\mathbf{y}^{(0)})\right]^2 + \left[\dfrac{\partial f}{\partial y_2}(\mathbf{y}^{(0)})\right]^2}}. \tag{6.2-3}$$

† If there are many hills and valleys, the point $\mathbf{y}^*$ determined by this procedure will depend on the starting point $\mathbf{y}^{(0)}$; thus $\mathbf{y}^*$ may be only a relative, or local, minimum.

Climbing in the direction of the vector $-\mathbf{z}(\mathbf{y}^{(0)})$, the change in $\mathbf{y}$ is given by

$$\Delta\mathbf{y} \triangleq \mathbf{y}^{(1)} - \mathbf{y}^{(0)} = -\tau\mathbf{z}(\mathbf{y}^{(0)}), \tag{6.2-4}$$

where $\tau > 0$ is the step size. With this selection for $\Delta\mathbf{y}$, the differential, which is a linear approximation to the change in f, becomes

$$df(\mathbf{y}^{(0)}) = -\tau\left[\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^{(0)})\right]^T \mathbf{z}(\mathbf{y}^{(0)}), \tag{6.2-5}$$

or, by using (6.2-3),

$$df(\mathbf{y}^{(0)}) = \frac{-\tau\left\{\left[\frac{\partial f}{\partial y_1}(\mathbf{y}^{(0)})\right]^2 + \left[\frac{\partial f}{\partial y_2}(\mathbf{y}^{(0)})\right]^2\right\}}{\left\|\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^{(0)})\right\|}$$

$$= -\tau\sqrt{\left[\frac{\partial f}{\partial y_1}(\mathbf{y}^{(0)})\right]^2 + \left[\frac{\partial f}{\partial y_2}(\mathbf{y}^{(0)})\right]^2}. \tag{6.2-5a}$$

Notice that this implies that

$$df(\mathbf{y}^{(0)}) \leq 0, \tag{6.2-6}$$

with the equality holding if and only if $\partial f/\partial \mathbf{y}$ is zero at $\mathbf{y}^{(0)}$.

We continue the iterative procedure by calculating $\mathbf{z}(\mathbf{y}^{(1)})$, the unit vector in the gradient direction at $\mathbf{y}^{(1)}$, and use the generalization of (6.2-4) to determine the next point, $\mathbf{y}^{(2)}$.

$$\Delta\mathbf{y} \triangleq \mathbf{y}^{(i+1)} - \mathbf{y}^{(i)} = -\tau\mathbf{z}(\mathbf{y}^{(i)}) \tag{6.2-4a}$$

A suitable value for the step size τ must also be selected. By inspection of Fig. 6-1 it is apparent that if τ is too large, then we overshoot the mark. On the other hand, if τ is too small, we are being overly timid; too much time is being spent measuring slopes and not enough time is spent climbing. In either case, the computation time may be excessive. Ideally, τ should be selected to minimize the total computation time; however, since this is a difficult problem in itself, various ad hoc strategies for choosing τ have been devised. One such strategy is to perform a single variable search to determine the value of τ that causes the largest decrease in f when moving in the direction of the vector $-\mathbf{z}(\mathbf{y}^{(0)})$. To use this technique, we would first determine the direction of steepest descent and then climb down in this direction until the function values no longer decreased. A new gradient direction would then be determined and the climbing process repeated. This procedure would be continued until a point were reached at which the gradient was zero. A few steps in the climbing process when τ is determined in this manner are shown in Fig. 6-2.

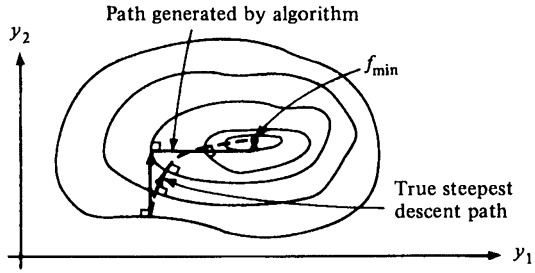


Figure 6-2 Steepest descent to find a minimum of a function

Minimization of Functionals by Steepest Descent

Let us now discuss an extension of the steepest descent concept which has been applied to optimal control problems by H. J. Kelley [K-8 and K-9] and A. E. Bryson, Jr. and W. F. Denham [B-5]. Suppose that a nominal control history $\mathbf{u}^{(i)}(t)$, $t \in [t_0, t_f]$, is known and used to solve the differential equations

$$\dot{\mathbf{x}}^{(i)}(t) = \mathbf{a}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), t) \tag{6.2-7}$$

$$\dot{\mathbf{p}}^{(i)}(t) = -\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), \mathbf{p}^{(i)}(t), t) \tag{6.2-8}$$

so that the nominal state-costate trajectory $\mathbf{x}^{(i)}$, $\mathbf{p}^{(i)}$ satisfies the boundary conditions

$$\mathbf{x}^{(i)}(t_0) = \mathbf{x}_0 \tag{6.2-9a}$$

$$\mathbf{p}^{(i)}(t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t_f)). \tag{6.2-9b}$$

If this nominal control history also satisfies

$$\frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), \mathbf{p}^{(i)}(t), t) = \mathbf{0}, \qquad t \in [t_0, t_f], \tag{6.2-10}$$

then $\mathbf{u}^{(i)}(t)$, $\mathbf{x}^{(i)}(t)$, and $\mathbf{p}^{(i)}(t)$ are extremal. Suppose that Eq. (6.2-10) is not satisfied; the variation of the augmented functional J_a on the nominal state-costate-control history is

$$\begin{aligned}\delta J_a = &\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t_f)) - \mathbf{p}^{(i)}(t_f)\right]^T \delta\mathbf{x}(t_f) \\ &+ \int_{t_0}^{t_f} \left\{\left[\dot{\mathbf{p}}^{(i)}(t) + \frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), \mathbf{p}^{(i)}(t), t)\right]^T \delta\mathbf{x}(t)\right. \\ &+ \left[\frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), \mathbf{p}^{(i)}(t), t)\right]^T \delta\mathbf{u}(t) \\ &\left. + [\mathbf{a}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), t) - \dot{\mathbf{x}}^{(i)}(t)]^T \delta\mathbf{p}(t)\right\} dt,\end{aligned} \tag{6.2-11}$$

where $\delta\mathbf{x}(t) \triangleq \mathbf{x}^{(i+1)}(t) - \mathbf{x}^{(i)}(t)$, $\delta\mathbf{u}(t) \triangleq \mathbf{u}^{(i+1)}(t) - \mathbf{u}^{(i)}(t)$, and

$$\delta\mathbf{p}(t) \triangleq \mathbf{p}^{(i+1)}(t) - \mathbf{p}^{(i)}(t).$$

If (6.2-7) through (6.2-9) are satisfied, then

$$\delta J_a = \int_{t_0}^{t_f} \left[\frac{\partial \mathscr{H}}{\partial \mathbf{u}}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), \mathbf{p}^{(i)}(t), t)\right]^T \delta\mathbf{u}(t)\, dt.\dagger \tag{6.2-12}$$

Recall that δJ_a is the linear part of the increment $\Delta J_a \triangleq J_a(\mathbf{u}^{(i+1)}) - J_a(\mathbf{u}^{(i)})$, and that if the norm of $\delta\mathbf{u}$, $\| \mathbf{u}^{(i+1)} - \mathbf{u}^{(i)} \|$, is small, the sign of ΔJ_a will be determined by the sign of δJ_a. Since our goal is to minimize J_a, we wish to make ΔJ_a negative. If we select the change in $\mathbf{u}$ as

$$\delta\mathbf{u}(t) = \mathbf{u}^{(i+1)}(t) - \mathbf{u}^{(i)}(t) = -\tau \frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t), \qquad t \in [t_0, t_f],\ddagger \tag{6.2-13}$$

with $\tau > 0$, then

$$\delta J_a = -\tau \int_{t_0}^{t_f} \left[\frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t)\right]^T \left[\frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t)\right] dt \leq 0, \tag{6.2-14}$$

because the integrand is nonnegative for all $t \in [t_0, t_f]$. The equality holds if and only if

$$\frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t) = \mathbf{0} \qquad \text{for all } t \in [t_0, t_f]. \tag{6.2-15}$$

Selecting $\delta\mathbf{u}$ in this manner, with $\| \delta\mathbf{u} \|$ sufficiently small, ensures that each value of the performance measure will be at least as small as the preceding value. Eventually, when J_a reaches a (relative) minimum the vector $\partial\mathscr{H}/\partial\mathbf{u}$ will be zero throughout the time interval $[t_0, t_f]$.

We have assumed that Eqs. (6.2-7) through (6.2-9) are satisfied. To see how this is accomplished, let us outline the algorithm as it would be executed if a digital computer were used.

The Steepest Descent Algorithm

The procedure we use to solve optimal control problems by the method of steepest descent is

1. Select a discrete approximation§ to the nominal control history $\mathbf{u}^{(0)}(t)$, $t \in [t_0, t_f]$, and store this in the memory of the digital computer. This can be done, for example, by subdividing the interval $[t_0, t_f]$ into N subintervals (generally of equal duration) and considering the control

† Henceforth we shall denote $\partial\mathscr{H}(\mathbf{x}^{(i)}(t), \mathbf{u}^{(i)}(t), \mathbf{p}^{(i)}(t), t)/\partial\mathbf{u}$ by $\partial\mathscr{H}^{(i)}(t)/\partial\mathbf{u}$.

‡ We shall assume that τ is a constant, although this is not a requirement.

§ A discrete approximation is required because the calculations are to be performed by a digital computer.

$\mathbf{u}^{(0)}$ as being piecewise-constant during each of these subintervals; that is,

$$\mathbf{u}^{(0)}(t) = \mathbf{u}^{(0)}(t_k), \qquad t \in [t_k, t_{k+1}), \qquad k = 0, 1, \ldots, N-1. \quad (6.2\text{-}16)$$

Let the iteration index i be zero.

2. Using the nominal control history $\mathbf{u}^{(i)}$, integrate the state equations from t_0 to t_f with initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$ and store the resulting state trajectory $\mathbf{x}^{(i)}$ as a piecewise-constant vector function.
3. Calculate $\mathbf{p}^{(i)}(t_f)$ by substituting $\mathbf{x}^{(i)}(t_f)$ from step 2 into Eq. (6.2-9b). Using this value of $\mathbf{p}^{(i)}(t_f)$ as the "initial condition" and the piecewise-constant values of $\mathbf{x}^{(i)}$ stored in step 2, integrate the costate equations from t_f to t_0, evaluate $\partial \mathscr{H}^{(i)}(t)/\partial \mathbf{u}$, $t \in [t_0, t_f]$, and store this function in piecewise-constant fashion. The costate trajectory does not need to be stored.
4. If

$$\left\| \frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}} \right\| \leq \gamma, \quad (6.2\text{-}17)$$

where γ is a preselected positive constant and

$$\left\| \frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}} \right\|^2 \triangleq \int_{t_0}^{t_f} \left[\frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t) \right]^T \left[\frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t) \right] dt, \quad (6.2\text{-}18)$$

terminate the iterative procedure, and output the extremal state and control. If the stopping criterion (6.2-17) is not satisfied, generate a new piecewise-constant control function given by

$$\mathbf{u}^{(i+1)}(t_k) = \mathbf{u}^{(i)}(t_k) - \tau \frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t_k), \qquad k = 0, \ldots, N-1, \quad (6.2\text{-}19)$$

where

$$\mathbf{u}^{(i)}(t) = \mathbf{u}^{(i)}(t_k) \qquad \text{for } t \in [t_k, t_{k+1}), \qquad k = 0, \ldots, N-1. \quad (6.2\text{-}20)$$

Replace $\mathbf{u}^{(i)}(t_k)$ by $\mathbf{u}^{(i+1)}(t_k)$, $k = 0, \ldots, N-1$, and return to step 2.

The value used for the termination constant γ will depend on the problem being solved and the accuracy desired of the solution. It may be desirable to perform several trial runs on a problem before γ is selected.

As mentioned previously, the step size τ is generally determined by some ad hoc strategy. One possible strategy is to select a value of τ which attempts to effect a certain value of ΔJ_a (perhaps some specified percentage of the preceding value of J_a). From Eqs. (6.2-14) and (6.2-18) observe that

$$\delta J_a = -\tau \left\| \frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}} \right\|^2 \leq 0. \tag{6.2-21}$$

To effect an approximate change of q percent in J_a, select τ as

$$\tau = \frac{\frac{q}{100} \left| J_a \right|}{\left\| \frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}} \right\|^2}. \tag{6.2-22}$$

This method of selecting τ generally requires the capability of intervening in the execution of the program to alter the value of q, because as J_a approaches a minimum, $\| \partial \mathscr{H}^{(i)}/\partial \mathbf{u} \| \to 0$; hence, if q is not decreased, the step size becomes large, and severe oscillations may result.

An alternative strategy for selecting τ is to use a single variable search. We choose an arbitrary starting value of τ, compute $\partial \mathscr{H}^{(i)}/\partial \mathbf{u}$, and find $\mathbf{u}^{(i+1)}$ using (6.2-19). Then a search among values of $\tau > 0$ is carried out until the smallest value of J_a is obtained. In other words, we move in the steepest descent direction until there is no further decrease in J_a.

An Illustrative Example

To illustrate the mechanics of the steepest descent procedure we have discussed, let us partially solve a simple example. Since all calculations will be done analytically, the piecewise-constant approximations mentioned previously will not be required.

Example 6.2-1. A first-order system is described by the state equation

$$\dot{x}(t) = -x(t) + u(t) \tag{6.2-23}$$

with initial condition $x(0) = 4.0$. It is desired to find $u(t)$, $t \in [0, 1]$, that minimizes the performance measure

$$J = x^2(1) + \int_0^1 \tfrac{1}{2} u^2(t)\, dt. \tag{6.2-24}$$

Notice that this problem is of the linear regulator type discussed in Section 5.2, and, therefore, can be solved without using iterative numerical techniques. The costate equation is

$$\dot{p}(t) = p(t) \tag{6.2-25}$$

with the boundary condition $p(1) = 2x(1)$. In addition, the optimal control and its costate must satisfy the relation

$$\frac{\partial \mathscr{H}}{\partial u} = u(t) + p(t) = 0. \tag{6.2-26}$$

As an initial guess for the optimal control, let us select $u^{(0)}(t) = 1.0$ throughout the interval [0, 1]. Integrating the state equation, using this control and the initial condition $x(0) = 4.0$, we obtain

$$x^{(0)}(t) = 3\epsilon^{-t} + 1; \tag{6.2-27}$$

hence, $p^{(0)}(1) = 2x^{(0)}(1) = 2[3\epsilon^{-1} + 1]$. Using this value for $p^{(0)}(1)$ and integrating the costate equation backward in time, we obtain

$$p^{(0)}(t) = 2\epsilon^{-1}[3\epsilon^{-1} + 1]\epsilon^{t}, \tag{6.2-28}$$

which makes

$$\frac{\partial \mathscr{H}^{(0)}}{\partial u}(t) = 1 + 2\epsilon^{-1}[3\epsilon^{-1} + 1]\epsilon^{t}. \tag{6.2-29}$$

If $u^{(0)}(t)$ had been the optimal control, then $\partial \mathscr{H}^{(0)}(t)/\partial u$ would have been identically zero. Assuming that our stopping criterion is not satisfied, we find that the next trial control is

$$u^{(1)}(t) = u^{(0)}(t) - \tau \frac{\partial \mathscr{H}^{(0)}}{\partial u}(t), \tag{6.2-30}$$

which if $\tau = 0.1$ gives

$$u^{(1)}(t) = 1.0 - 0.1[1 + 2\epsilon^{-1}[3\epsilon^{-1} + 1]\epsilon^{t}]. \tag{6.2-31}$$

To continue the iterative algorithm, we would repeat the preceding steps, using this revised control history. Eventually the iterative procedure should converge to the optimal control history, $u^*(t)$.

The preceding example indicates the steps involved in carrying out one iteration of the steepest descent algorithm. Let us now use this algorithm to determine the optimal trajectory and control for a continuous stirred-tank chemical reactor. This chemical engineering problem will also provide a basis for comparing the steepest descent method with other numerical techniques to be discussed in the following sections of this chapter.

A Continuous Stirred-Tank Chemical Reactor

Example 6.2-2. The state equations for a continuous stirred-tank chemical reactor are given below [L-5]. The flow of a coolant through a coil inserted in the reactor is to control the first-order, irreversible exothermic reaction taking place in the reactor. The states of the plant are $x_1(t) = T(t)$ (the

deviation from the steady-state temperature) and $x_2(t) = C(t)$ (the deviation from the steady-state concentration). $u(t)$, the normalized control variable, represents the effect of coolant flow on the chemical reaction. The state equations are

$$\begin{aligned}\dot{x}_1(t) &= -2[x_1(t) + 0.25] + [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right] \\ &\quad - [x_1(t) + 0.25]u(t) \\ \dot{x}_2(t) &= 0.5 - x_2(t) - [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right]\end{aligned} \tag{6.2-32}$$

with initial conditions $\mathbf{x}(0) = [0.05 \quad 0]^T$. The performance measure to be minimized is

$$J = \int_0^{0.78} [x_1^2(t) + x_2^2(t) + Ru^2(t)]\,dt, \tag{6.2-33}$$

indicating that the desired objective is to maintain the temperature and concentration close to their steady-state values without expending large amounts of control effort. R is a weighting factor that we shall select (arbitrarily) as 0.1. The costate equations are determined from the Hamiltonian,

$$\begin{aligned}\mathscr{H}(\mathbf{x}(t), u(t), \mathbf{p}(t)) &= x_1^2(t) + x_2^2(t) + Ru^2(t) \\ &\quad + p_1(t)\left[-2[x_1(t) + 0.25] + [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right]\right. \\ &\quad \left. - [x_1(t) + 0.25]u(t)\right] + p_2(t)\left[0.5 - x_2(t)\right. \\ &\quad \left. - [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right]\right]\end{aligned} \tag{6.2-34}$$

as

$$\begin{aligned}\dot{p}_1(t) &= -\frac{\partial \mathscr{H}}{\partial x_1} = -2x_1(t) + 2p_1(t) \\ &\quad - p_1(t)[x_2(t) + 0.5]\left[\frac{50}{[x_1(t) + 2]^2}\right]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right] \\ &\quad + p_1(t)u(t) + p_2(t)[x_2(t) + 0.5]\left[\frac{50}{[x_1(t) + 2]^2}\right]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right] \\ \dot{p}_2(t) &= -\frac{\partial \mathscr{H}}{\partial x_2} = -2x_2(t) - p_1(t)\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right] \\ &\quad + p_2(t)\left[1 + \exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right]\right].\end{aligned} \tag{6.2-35}$$

The algebraic relation that must be satisfied is

$$\frac{\partial \mathscr{H}}{\partial u} = 2Ru(t) - p_1(t)[x_1(t) + 0.25] = 0. \tag{6.2-36}$$

Since the final states are free and not explicitly present in the performance measure, the boundary conditions at $t = t_f$ are $\mathbf{p}(t_f) = \mathbf{0}$.

A program was written in FORTRAN IV for the IBM 360/67 digital computer. Numerical integration was carried out using a fourth-order Runge-Kutta-Gill method with double-precision arithmetic and an integration interval of 0.1 unit.

The norm used was

$$\left\| \frac{\partial \mathscr{H}}{\partial u} \right\|^2 = \int_0^{0.78} \left[\frac{\partial \mathscr{H}}{\partial u}(t) \right]^2 dt, \tag{6.2-37}$$

and the iterative procedure was terminated when either $\| \partial \mathscr{H}/\partial u \| \leq 10^{-2}$ or $|J^{(i)} - J^{(i+1)}| \leq 10^{-6}$. To ensure that a monotonically decreasing sequence of performance indices was generated, each trial control was required to provide a smaller performance measure than the preceding control. This was accomplished by halving τ and re-generating any trial control that increased the performance measure.

With $u^{(0)}(t) = 1.0$, $t \in [0, 0.78]$, and an initial τ equal to 1.0, the value of the performance measure as a function of the number of iterations is as shown in Fig. 6-3. Notice that the first four iterations reduce the performance measure significantly; however, the final 13 iterations yield only very slight improvement. This type of progress is typical of the steepest descent method. The optimal control and the optimal state

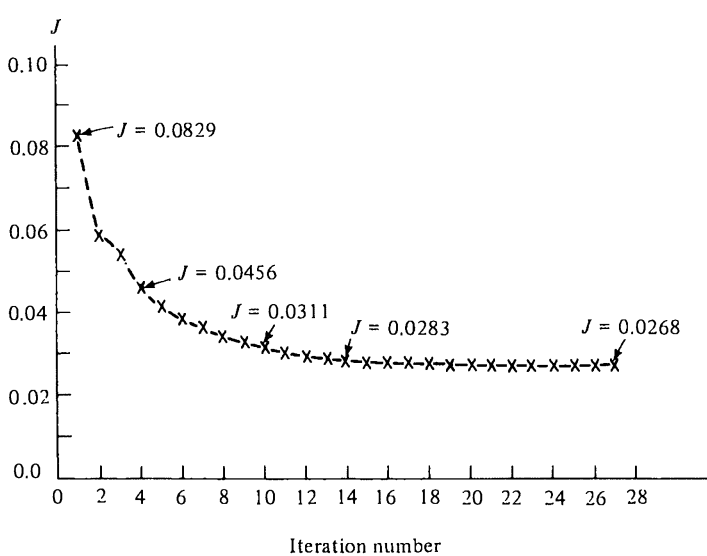


Figure 6-3 Performance measure reduction by the steepest descent method—stirred-tank reactor problem

trajectory obtained by the iterative solution are shown in Figs. 6-4 and 6-5.

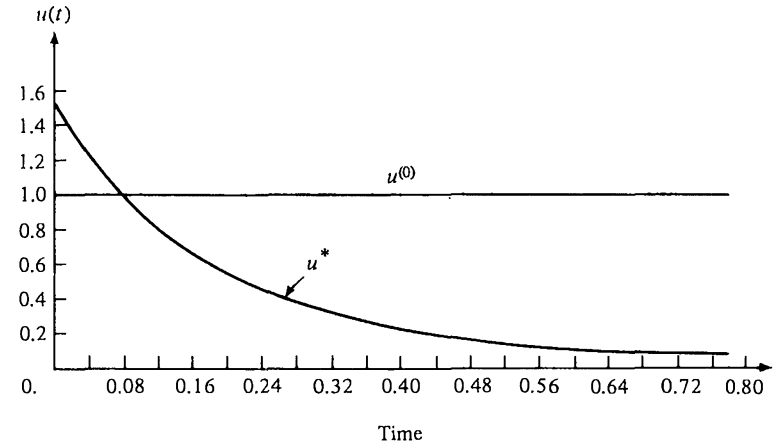


Figure 6-4 The optimal control for the stirred-tank reactor (steepest descent solution)

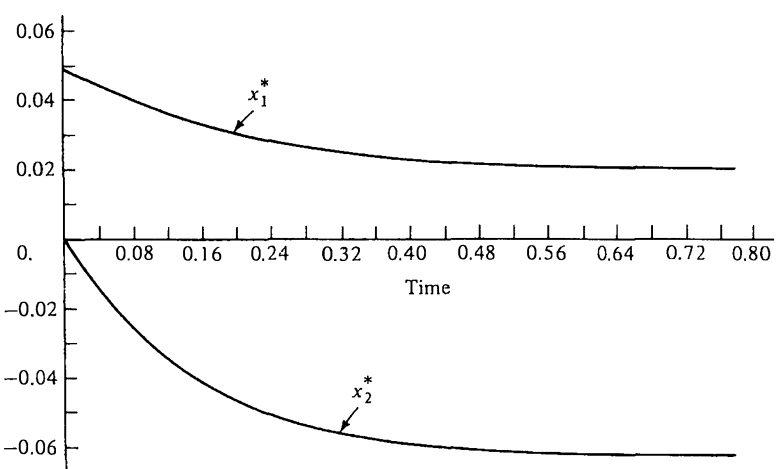


Figure 6-5 The optimal trajectory for the stirred-tank reactor (steepest descent solution)

To illustrate the effects of various initial step sizes and different initial guesses for the control history, three additional solutions were obtained. The results of these computer runs are summarized in Table 6-1.

Table 6-1 SUMMARY OF STEEPEST DESCENT SOLUTION OF THE STIRRED-TANK CHEMICAL REACTOR PROBLEM

Initial control $u^{(0)}(t)$, $t \in [0., 0.78]$	*Initial* τ	*Number of iterations required*	*Minimum value of* J, J^*	*Final* τ	*Stopping criterion satisfied*
1.0	1.00	27	0.02681	0.25	NORM
1.0	0.25	48	0.02682	0.25	NORM
0.0	1.00	7	0.02678	0.25	NORM
0.0	0.25	11	0.02680	0.25	NORM

Features of the Steepest Descent Algorithm

To conclude our discussion of the steepest descent method, let us summarize the important features of the algorithm.

Initial Guess. A nominal control history, $\mathbf{u}^{(0)}(t)$, $t \in [t_0, t_f]$, must be selected to begin the numerical procedure. In selecting the nominal control we utilize whatever physical insight we have about the problem.

Storage Requirements. The current trial control $\mathbf{u}^{(i)}$, the corresponding state trajectory $\mathbf{x}^{(i)}$, and the gradient history $\partial \mathscr{H}^{(i)}/\partial \mathbf{u}$, are stored. If storage must be conserved, the state values needed to determine $\partial \mathscr{H}^{(i)}/\partial \mathbf{u}$ can be obtained by reintegrating the state equations with the costate equations. If this is done $\mathbf{x}^{(i)}$ does not need to be stored; however, the computation time will increase. Generating the required state values in this manner may make the results of the backward integration more accurate, because the piecewise-constant approximation for $\mathbf{x}^{(i)}$ need not be used.

Convergence. The method of steepest descent is generally characterized by ease of starting—the initial guess for the control is not usually crucial. On the other hand, as a minimum is approached, the gradient becomes small and the method has a tendency to converge slowly.

Computations Required. In each iteration numerical integration of $2n$ first-order ordinary differential equations is required. In addition, the time history of $\partial \mathscr{H}^{(i)}/\partial \mathbf{u}$ at the times t_k, $k = 0, 1, \ldots, N-1$, must be evaluated. To speed up the iterative procedure, a single variable search may be used to determine the step size for the change in the trial control.

Stopping Criterion. The iterative procedure is terminated when a criterion such as $\|\partial \mathscr{H}^{(i)}/\partial \mathbf{u}\| < \gamma_1$ or $|J^{(i)} - J^{(i+1)}| < \gamma_2$ is satisfied; γ_1 and γ_2 are preselected positive numbers.

Modifications Required for Fixed End Point Problems. One way to modify the procedure we have discussed for problems in which some or all of the final states are fixed is to use the penalty function approach. For example, if the desired final state is denoted by $\mathbf{x}_d(t_f)$, we can add a term to the performance measure of the form

$$\tfrac{1}{2}[\mathbf{x}_d(t_f) - \mathbf{x}(t_f)]^T\mathbf{H}[\mathbf{x}_d(t_f) - \mathbf{x}(t_f)],$$

where $\mathbf{H}$ is a diagonal matrix with large positive elements, and treat $\mathbf{x}(t_f)$ as if it were free. Doing this, we find that the boundary conditions become $\mathbf{p}(t_f) = \mathbf{H}[\mathbf{x}(t_f) - \mathbf{x}_d(t_f)]$. By using this technique, fixed and free end point problems can be solved with the same computer program. Adding the penalty term to the performance index penalizes deviations of the final states from their desired values. For an alternative approach to fixed end point problems, see the discussion in reference [B-5].

6.3 VARIATION OF EXTREMALS

The iterative numerical technique that we shall discuss in this section is called *variation of extremals*, because every trajectory generated by the algorithm satisfies Eqs. (6.1-1) through (6.1-3) and hence is an extremal. To illustrate the basic concept of the algorithm, let us consider a simple example.

A First-Order Optimal Control Problem

Suppose that a first-order system

$$\dot{x}(t) = a(x(t), u(t), t) \tag{6.3-1}$$

is to be controlled to minimize a performance measure of the form

$$J = \int_{t_0}^{t_f} g(x(t), u(t), t)\, dt, \tag{6.3-2}$$

where $x(t_0) = x_0$ is given, t_0 and t_f are specified, and the admissible state and control values are not constrained by any boundaries. If the equation [corresponding to (6.1-3)]

$$\frac{\partial \mathscr{H}}{\partial u} = 0 \tag{6.3-3}$$

is solved for the control in terms of the state and costate and substituted in the state and costate equations, the reduced differential equations

$$\begin{aligned} \dot{x}(t) &= a(x(t), p(t), t) \\ \dot{p}(t) &= d(x(t), p(t), t) \end{aligned} \tag{6.3-4}$$

are obtained. In general, d is a nonlinear function of $x(t)$, $p(t)$, and t. Since $h = 0$ in the performance measure, Eq. (6.1-4b) gives $p(t_f) = 0$. To determine an optimal trajectory, we must find a solution of Eq. (6.3-4) that satisfies the boundary conditions $x(t_0) = x_0$, $p(t_f) = 0$.

If $p(t_0)$ were known, Eq. (6.3-4) could be solved by using numerical integration. Since this is not the case, suppose we guess a value $p^{(0)}(t_0)$ for the initial costate, and use this initial value to integrate numerically (6.3-4) from t_0 to t_f; we denote the costate solution obtained by this integration as $p^{(0)}$. If we should guess an initial costate value that causes $p^{(0)}(t_f)$ to be zero, the two-point boundary-value problem is solved. In general, however, it will turn out that the final costate will not equal zero. Notice that the value obtained for the terminal costate, $p^{(0)}(t_f)$, will depend on the number chosen for $p^{(0)}(t_0)$; in other words, $p(t_f)$ *is a function of* $p(t_0)$. Unfortunately, an analytical expression for this function is not known, nor is it readily determined; however, *values* of the function [such as $p^{(0)}(t_f)$] can be found by using selected values of $p(t_0)$ to integrate numerically the reduced state-costate equations. The method of variation of extremals is an algorithm that uses the observed values of $p(t_f)$ to adjust systematically the guessed values of $p(t_0)$. One technique for making systematic adjustments of the initial costate values is based on Newton's method for finding roots of nonlinear equations [F-1].

A geometric interpretation of Newton's method is provided by Fig. 6-6, where a possible curve of $p(t_f)$ as a function of $p(t_0)$ is shown. (Unfortunately, we do not really know what this curve looks like—if we did, our problem would be solved.) Newton's method consists of finding the tangent to the $p(t_f)$ versus $p(t_0)$ curve at an arbitrary starting point q and extrapolating this

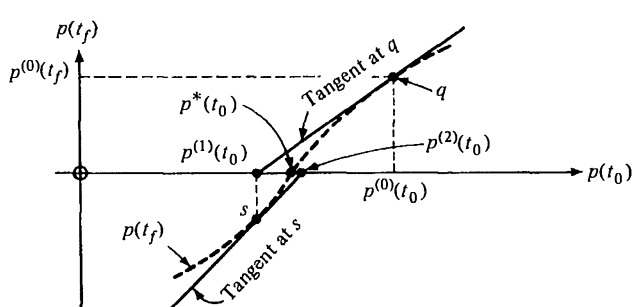


Figure 6-6 A typical relationship between $p(t_f)$ and $p(t_0)$

tangent to determine the point where it intersects the desired value of $p(t_f)$ (zero in this problem). This completes one iteration of the algorithm; the next iteration consists of extrapolating the tangent at point s to determine $p^{(2)}(t_0)$. The iterative procedure continues until the value of $p^{(i)}(t_f)$ obtained is sufficiently close to zero to satisfy a specified termination criterion. Notice that for small changes in $p(t_0)$ the approximation of the nonlinear curve by its tangent is quite good; hence if the initial costate values generated by the iterative procedure approach $p^*(t_0)$, convergence of the algorithm should be quite rapid.

The slope of the curve at point q, which is needed to determine the equation of the tangent, can be found approximately by perturbing the value of $p^{(0)}(t_0)$ and evaluating (by integration of the reduced state and costate equations) the perturbed value of the final costate $p^{(0)}(t_f) + \delta p^{(0)}(t_f)$; that is,

$$\begin{aligned}(\text{slope at } q) &= \left.\frac{dp(t_f)}{dp(t_0)}\right|_{p^{(0)}(t_0)} \\ &\doteq \frac{\delta p^{(0)}(t_f)}{\delta p^{(0)}(t_0)}.\end{aligned} \tag{6.3-5}$$

The tangent curve is described by the equation

$$p(t_f) = m \cdot p(t_0) + b, \tag{6.3-6}$$

where m is the slope of the tangent given by (6.3-5) and b is the $p(t_f)$ axis intercept. Since $p(t_f) = p^{(0)}(t_f)$ for $p(t_0) = p^{(0)}(t_0)$, the intercept is

$$b = p^{(0)}(t_f) - mp^{(0)}(t_0); \tag{6.3-7}$$

hence,

$$p(t_f) = mp(t_0) + [p^{(0)}(t_f) - mp^{(0)}(t_0)]. \tag{6.3-8}$$

To find the point on the tangent curve $p^{(1)}(t_0)$ where $p(t_f) = 0$, we substitute $p(t_f) = 0$ and $p(t_0) = p^{(1)}(t_0)$ into (6.3-8) with the result

$$0 = mp^{(1)}(t_0) + [p^{(0)}(t_f) - mp^{(0)}(t_0)]. \tag{6.3-9}$$

Solving for $p^{(1)}(t_0)$, the next trial value of $p(t_0)$, we obtain

$$\begin{aligned}p^{(1)}(t_0) &= p^{(0)}(t_0) - [m]^{-1} p^{(0)}(t_f) \\ &\triangleq p^{(0)}(t_0) - \left[\left.\frac{dp(t_f)}{dp(t_0)}\right|_{p^{(0)}(t_0)}\right]^{-1} p^{(0)}(t_f).\end{aligned} \tag{6.3-10}$$

Another way of deriving Eq. (6.3-10) is to solve the relationship $m = [p^{(0)}(t_f) - 0]/[p^{(0)}(t_0) - p^{(1)}(t_0)]$ (obtained by inspection of Fig. 6-6) for $p^{(1)}(t_0)$.

In general, using this algorithm gives

$$p^{(i+1)}(t_0) = p^{(i)}(t_0) - \left[\frac{dp(t_f)}{dp(t_0)}\bigg|_{p^{(i)}(t_0)}\right]^{-1} p^{(i)}(t_f) \tag{6.3-10a}$$

as the expression for the $(i + 1)$st trial value of $p(t_0)$.

To illustrate how the iterative procedure is carried out, let us solve a simple numerical example.

Example 6.3-1. Assume that the reduced state-costate equations are given by

$$\begin{aligned} \dot{x}(t) &= -2x(t) - p(t) + 6 \\ \dot{p}(t) &= 4x(t) + 3p(t) \end{aligned} \tag{6.3-11}$$

and that the boundary conditions are $x(0) = 3$, $p(1) = 0$.

The iterative procedure begins by guessing a value for $p(0)$; suppose we guess $p^{(0)}(0) = 0$. Since Eqs. (6.3-11) are linear, time-invariant differential equations, the solution

$$\begin{aligned} x^{(0)}(t) &= -4\epsilon^{-t} - 2\epsilon^{2t} + 9 \\ p^{(0)}(t) &= \quad 4\epsilon^{-t} + 8\epsilon^{2t} - 12 \end{aligned} \tag{6.3-12}$$

can be obtained easily by using analytical methods.

Next, we perturb $p^{(0)}(0)$ by a small amount, say $\delta p^{(0)}(0) = +0.001$. If we use $p(0) = 0.001$, the solution obtained for $p(t)$ is

$$p^{(0)}(t) + \delta p^{(0)}(t) = 4\epsilon^{-t} + 8\epsilon^{2t} - 12 - \frac{0.001}{3}\epsilon^{-t} + \frac{0.004}{3}\epsilon^{2t}. \tag{6.3-13}$$

We need to store only the trajectory values at $t = 1$ to use Newton's method. From (6.3-5) we obtain

$$\frac{dp(t_f)}{dp(t_0)}\bigg|_{p(t_0)=0} = \frac{-\frac{0.001}{3}\epsilon^{-1} + \frac{0.004}{3}\epsilon^{2}}{0.001} = -\frac{1}{3}\epsilon^{-1} + \frac{4}{3}\epsilon^{2}, \tag{6.3-14}$$

and substituting this value in (6.3-10) gives as the new value of $p(t_0)$

$$\begin{aligned} p^{(1)}(t_0) &= 0 - [-\tfrac{1}{3}\epsilon^{-1} + \tfrac{4}{3}\epsilon^{2}]^{-1}[4\epsilon^{-1} + 8\epsilon^{2} - 12] \\ &= -4.993. \end{aligned} \tag{6.3-15}$$

Then, using this value for $p(t_0)$ yields

$$\begin{aligned} x^{(1)}(t) &= -0.336\epsilon^{2t} - 5.664\epsilon^{-t} + 9 \\ p^{(1)}(t) &= 1.342\epsilon^{2t} + 5.664\epsilon^{-t} - 12. \end{aligned} \tag{6.3-16}$$

It is easily verified that $x^{(1)}(t)$, $p^{(1)}(t)$ satisfy the specified boundary conditions $x(0) = 3$, $p(1) = 0$; hence, the iterative procedure has converged in only one iteration. Although the reader may suspect that this has occurred because of an especially fortuitous initial guess, the procedure would have converged in one iteration regardless of the initial guess, because the reduced differential equations are linear. It is left as an exercise for the reader (Problem 6-2) to verify that if a two-point boundary-value problem is linear, the method of variation of extremals converges in one iteration regardless of the initial guess selected for the missing boundary conditions.

Before considering the generalization of the technique we have discussed to higher-order systems, we note that if the desired value of the final costate $p(t_f)$ is some nonzero *constant* p_f, then Eq. (6.3-10a) must be modified to read

$$p^{(i+1)}(t_0) = p^{(i)}(t_0) - \left[\frac{dp(t_f)}{dp(t_0)}\bigg|_{p^{(i)}(t_0)}\right]^{-1}[p^{(i)}(t_f) - p_f]. \tag{6.3-17}$$

Extensions Required for Systems of 2n Differential Equations

We have shown how the method of variation of extremals can be used to solve a two-point boundary-value problem involving two first-order differential equations. If we have $2n$ first-order differential equations (n state equations and n costate equations), the matrix generalization of Eq. (6.3-10a) is

$$\mathbf{p}^{(i+1)}(t_0) = \mathbf{p}^{(i)}(t_0) - [\mathbf{P}_p(\mathbf{p}^{(i)}(t_0), t_f)]^{-1}\mathbf{p}^{(i)}(t_f), \tag{6.3-18}$$

where $\mathbf{P}_p(\mathbf{p}^{(i)}(t_0), t)$ is the $n \times n$ matrix of partial derivatives of the components of $\mathbf{p}(t)$ with respect to each of the components of $\mathbf{p}(t_0)$, evaluated at $\mathbf{p}^{(i)}(t_0)$; that is,

$$\mathbf{P}_p(\mathbf{p}^{(i)}(t_0), t) \triangleq \begin{bmatrix} \dfrac{\partial p_1(t)}{\partial p_1(t_0)} & \dfrac{\partial p_1(t)}{\partial p_2(t_0)} & \cdots & \dfrac{\partial p_1(t)}{\partial p_n(t_0)} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial p_n(t)}{\partial p_1(t_0)} & \dfrac{\partial p_n(t)}{\partial p_2(t_0)} & \cdots & \dfrac{\partial p_n(t)}{\partial p_n(t_0)} \end{bmatrix}_{p^{(i)}(t_0)} \tag{6.3-19}$$

The $\mathbf{P}_p$ matrix indicates the influence of changes in the initial costate on the costate trajectory at time t; hence, we shall call $\mathbf{P}_p$ the *costate influence function matrix*. Notice that (6.3-18) requires that $\mathbf{P}_p$ be known only at the terminal time t_f.

Equation (6.3-18) is appropriate only if the desired value of the final costate is zero, which occurs if the term $h(\mathbf{x}(t_f))$ is missing from the performance measure. If, however, h is not absent from the performance measure, it can be shown (see Problem 6-1) that the appropriate equation for the iterative procedure is

$$\mathbf{p}^{(i+1)}(t_0) = \mathbf{p}^{(i)}(t_0) + \left\{\left[\left[\frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}(t_f))\right]\mathbf{P}_x(\mathbf{p}(t_0), t_f) - \mathbf{P}_p(\mathbf{p}(t_0), t_f)\right]_i\right\}^{-1} \cdot \left[\mathbf{p}(t_f) - \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}(t_f))\right]_i, \tag{6.3-20}$$

where $\mathbf{P}_x(\mathbf{p}^{(i)}(t_0), t_f)$ is the $n \times n$ *state influence function matrix*

$$\mathbf{P}_x(\mathbf{p}^{(i)}(t_0), t) \triangleq \begin{bmatrix} \frac{\partial x_1(t)}{\partial p_1(t_0)} & \frac{\partial x_1(t)}{\partial p_2(t_0)} & \cdots & \frac{\partial x_1(t)}{\partial p_n(t_0)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial x_n(t)}{\partial p_1(t_0)} & \frac{\partial x_n(t)}{\partial p_2(t_0)} & \cdots & \frac{\partial x_n(t)}{\partial p_n(t_0)} \end{bmatrix}_{\mathbf{p}^{(i)}(t_0)} \tag{6.3-21}$$

evaluated at $t = t_f$. The notation $[\ ]_i$ means that the enclosed terms are evaluated on the ith trajectory, and

$$\frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}(t_f))$$

is the matrix whose jkth element is

$$\left[\frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}(t_f))\right]_{jk} \triangleq \frac{\partial^2 h}{\partial x_j \partial x_k}(\mathbf{x}(t_f)).$$

Notice that if $h = 0$, (6.3-20) reduces to Eq. (6.3-18).

In order to use Eq. (6.3-20) in an iterative manner, we must first determine the influence function matrices. Let us now discuss how these matrices can be computed.

Determination of the Influence Function Matrices

Conceptually we may think of finding the $\mathbf{P}_p$ and $\mathbf{P}_x$ matrices at $t = t_f$ by the following finite difference procedure:

1. Using $\mathbf{p}(t_0) = \mathbf{p}^{(i)}(t_0)$ and $\mathbf{x}(t_0) = \mathbf{x}_0$, integrate the reduced state and

costate equations from t_0 to t_f, and store the resulting values of $\mathbf{p}^{(i)}(t_f)$ and $\mathbf{x}^{(i)}(t_f)$.

2. Perturb the first component of the vector $\mathbf{p}^{(i)}(t_0)$ by an amount $\delta p_1(t_0)$; again integrate the reduced state and costate equations from t_0 to t_f. The first column of $\mathbf{P}_p(\mathbf{p}^{(i)}(t_0), t_f)$ is found from the relationship

$$\left.\frac{\partial \mathbf{p}(t_f)}{\partial p_1(t_0)}\right|_{\mathbf{p}^{(i)}(t_0)} \doteq \frac{\delta \mathbf{p}(t_f)}{\delta p_1(t_0)}, \tag{6.3-22}$$

where $\delta \mathbf{p}(t_f)$ is found by subtracting the value of $\mathbf{p}^{(i)}(t_f)$ generated in step 1 from the value of $\mathbf{p}(t_f)$ generated by using the perturbed value of $\mathbf{p}(t_0)$.

Similarly, the first column of $\mathbf{P}_x(\mathbf{p}^{(i)}(t_0), t_f)$ is found from

$$\left.\frac{\partial \mathbf{x}(t_f)}{\partial p_1(t_0)}\right|_{\mathbf{p}^{(i)}(t_0)} \doteq \frac{\delta \mathbf{x}(t_f)}{\delta p_1(t_0)}. \tag{6.3-23}$$

3. The remaining columns of the influence function matrices at $t = t_f$ are generated by perturbing each of the components of $\mathbf{p}(t_0)$, *with all other components at their initial settings*, and evaluating the ratios of the changes in $\mathbf{p}(t_f)$ and $\mathbf{x}(t_f)$ to the change in the appropriate component of $\mathbf{p}(t_0)$.

If the influence function matrices are computed by using this procedure, the integrations of steps 1, 2, and 3 would probably be performed simultaneously.

One rather apparent difficulty with this finite difference procedure is the selection of the perturbations. Relatively large perturbations may cause the difference approximations of the partial derivatives to be inaccurate, whereas, on the other hand, very small perturbations may significantly increase the effects of inaccuracies caused by numerical integration, and truncation and round-off errors. These difficulties can be avoided, however, by using a different method to evaluate the influence function matrices; let us now discuss this alternative procedure.

Assume that $\partial \mathscr{H}/\partial \mathbf{u}$ has been solved for $\mathbf{u}(t)$ and used to obtain the reduced state and costate differential equations

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \frac{\partial \mathscr{H}}{\partial \mathbf{p}}(\mathbf{x}(t), \mathbf{p}(t), t) \\ \dot{\mathbf{p}}(t) &= -\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}(t), \mathbf{p}(t), t). \end{aligned} \tag{6.3-24}$$

Taking the partial derivatives of these equations with respect to the initial value of the costate vector gives

$$\begin{aligned}\frac{\partial}{\partial \mathbf{p}(t_0)}[\dot{\mathbf{x}}(t)] &= \frac{\partial}{\partial \mathbf{p}(t_0)}\left[\frac{\partial \mathscr{H}}{\partial \mathbf{p}}(\mathbf{x}(t), \mathbf{p}(t), t)\right] \\ \frac{\partial}{\partial \mathbf{p}(t_0)}[\dot{\mathbf{p}}(t)] &= \frac{\partial}{\partial \mathbf{p}(t_0)}\left[-\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}(t), \mathbf{p}(t), t)\right].\end{aligned} \tag{6.3-25}$$

If it is assumed that $\partial[d\mathbf{x}/dt]/\partial \mathbf{p}(t_0)$ and $\partial[d\mathbf{p}/dt]/\partial \mathbf{p}(t_0)$ are continuous with respect to $\mathbf{p}(t_0)$ and t, the order of differentiation can be interchanged on the left side of (6.3-25); doing this and using the chain rule on the right-hand side, we obtain

$$\begin{aligned}\frac{d}{dt}\left[\frac{\partial \mathbf{x}(t)}{\partial \mathbf{p}(t_0)}\right] &= \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{p}\partial \mathbf{x}}(\mathbf{x}(t), \mathbf{p}(t), t)\right]\frac{\partial \mathbf{x}(t)}{\partial \mathbf{p}(t_0)} \\ &\quad + \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{p}^2}(\mathbf{x}(t), \mathbf{p}(t), t)\right]\frac{\partial \mathbf{p}(t)}{\partial \mathbf{p}(t_0)} \\ \frac{d}{dt}\left[\frac{\partial \mathbf{p}(t)}{\partial \mathbf{p}(t_0)}\right] &= \left[-\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}^2}(\mathbf{x}(t), \mathbf{p}(t), t)\right]\frac{\partial \mathbf{x}(t)}{\partial \mathbf{p}(t_0)} \\ &\quad - \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}\partial \mathbf{p}}(\mathbf{x}(t), \mathbf{p}(t), t)\right]\frac{\partial \mathbf{p}(t)}{\partial \mathbf{p}(t_0)}.\end{aligned} \tag{6.3-26}$$

The indicated partial derivatives are $n \times n$ matrices having jkth elements

$$\left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{p}\partial \mathbf{x}}\right]_{jk} \triangleq \frac{\partial^2 \mathscr{H}}{\partial p_j \partial x_k}, \quad \text{and} \quad \left[\frac{\partial \mathbf{x}(t)}{\partial \mathbf{p}(t_0)}\right]_{jk} \triangleq \frac{\partial x_j(t)}{\partial p_k(t_0)}, \quad \text{etc.}$$

Notice that if the definitions of the influence function matrices given by (6.3-19) and (6.3-21) are used, Eq. (6.3-26) becomes

$$\begin{aligned}\frac{d}{dt}[\mathbf{P}_x(\mathbf{p}^{(t)}(t_0), t)] &= \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{p}\partial \mathbf{x}}(t)\right]_t \mathbf{P}_x(\mathbf{p}^{(t)}(t_0), t) \\ &\quad + \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{p}^2}(t)\right]_t \mathbf{P}_p(\mathbf{p}^{(t)}(t_0), t) \\ \frac{d}{dt}[\mathbf{P}_p(\mathbf{p}^{(t)}(t_0), t)] &= \left[-\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}^2}(t)\right]_t \mathbf{P}_x(\mathbf{p}^{(t)}(t_0), t) \\ &\quad + \left[-\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}\partial \mathbf{p}}(t)\right]_t \mathbf{P}_p(\mathbf{p}^{(t)}(t_0), t)\end{aligned} \tag{6.3-26a}$$

where $[\ ]_t$ indicates that the enclosed matrices are evaluated on the trajectory $\mathbf{x}^{(t)}$, $\mathbf{p}^{(t)}$ obtained by integrating the reduced state-costate equations with initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$, $\mathbf{p}(t_0) = \mathbf{p}^{(t)}(t_0)$. Notice that (6.3-26a) represents a set of $2n^2$ first-order differential equations involving the influence function matrices. The matrices $\mathbf{P}_x(\mathbf{p}^{(t)}(t_0), t_f)$ and $\mathbf{P}_p(\mathbf{p}^{(t)}(t_0), t_f)$ can be obtained by integrating these differential equations simultaneously with the reduced state-costate differential equations. The appropriate initial conditions for the influence function equations are

$$\mathbf{P}_x(\mathbf{p}^{(i)}(t_0), t_0) = \left.\frac{\partial \mathbf{x}(t_0)}{\partial \mathbf{p}(t_0)}\right|_{\mathbf{p}^{(i)}(t_0)} = \mathbf{0} \quad \text{(the } n \times n \text{ zero matrix)} \tag{6.3-27a}$$

$$\mathbf{P}_p(\mathbf{p}^{(i)}(t_0), t_0) = \left.\frac{\partial \mathbf{p}(t_0)}{\partial \mathbf{p}(t_0)}\right|_{\mathbf{p}^{(i)}(t_0)} = \mathbf{I} \quad \text{(the } n \times n \text{ identity matrix).} \tag{6.3-27b}$$

Equation (6.3-27a) follows because a change in any of the components of $\mathbf{p}(t_0)$ does not affect the value of $\mathbf{x}(t_0)$; the state values are specified at time t_0. A change in the jth component of $\mathbf{p}(t_0)$ changes only $p_j(t_0)$; hence the result shown in (6.3-27b) is obtained.

The Variation of Extremals Algorithm

Before solving a numerical example, let us first outline the steps required to carry out the variation of extremals method:

1. Form the reduced differential equations by solving $\partial \mathscr{H}/\partial \mathbf{u} = \mathbf{0}$ for $\mathbf{u}(t)$ in terms of $\mathbf{x}(t)$, $\mathbf{p}(t)$, and substituting in the state and costate equations [which then contain only $\mathbf{x}(t)$, $\mathbf{p}(t)$, and t].
2. Guess $\mathbf{p}^{(0)}(t_0)$, an initial value for the costate, and set the iteration index i to zero.
3. Using $\mathbf{p}(t_0) = \mathbf{p}^{(i)}(t_0)$ and $\mathbf{x}(t_0) = \mathbf{x}_0$ as initial conditions, integrate the reduced state-costate equations and the influence function equations (6.3-26a), with initial conditions (6.3-27), from t_0 to t_f. Store only the values $\mathbf{p}^{(i)}(t_f)$, $\mathbf{x}^{(i)}(t_f)$, and the $n \times n$ matrices $\mathbf{P}_p(\mathbf{p}^{(i)}(t_0), t_f)$ and $\mathbf{P}_x(\mathbf{p}^{(i)}(t_0), t_f)$.
4. Check to see if the termination criterion $\| \mathbf{p}^{(i)}(t_f) - \partial h(\mathbf{x}^{(i)}(t_f))/\partial \mathbf{x} \| < \gamma$ is satisfied. If it is, use the final iterate of $\mathbf{p}^{(i)}(t_0)$ to reintegrate the state and costate equations and print out (or graph) the optimal trajectory and the optimal control. If the stopping criterion is not satisfied, use the iteration equation (6.3-20) to determine the value for $\mathbf{p}^{(i+1)}(t_0)$, increase i by one, and return to step 3.

Notice that steps 1 and 2 are performed off-line by the user or programmer; the computer program consists of steps 3 and 4. Let us now illustrate the use of variation of extremals to re-solve the continuous stirred-tank chemical reactor problem discussed in Section 6.2.

Example 6.3-2. The appropriate equations are the same as in Example 6.2-2. To determine the reduced differential equations, we solve (6.2-36) for $u(t)$ to obtain

$$u(t) = \frac{1}{2R} p_1(t)[x_1(t) + 0.25]. \tag{6.3-28}$$

Substituting (6.3-28) with $R = 0.1$ in the state and costate equations gives

$$\dot{x}_1(t) = -[x_1(t) + 0.25][2 + 5p_1(t)[x_1(t) + 0.25]] + [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right] \tag{6.3-29}$$

$$\dot{x}_2(t) = 0.5 - x_2(t) - [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right]$$

$$\dot{p}_1(t) = -2[x_1(t) - p_1(t)] + 50\frac{[p_2(t) - p_1(t)][x_2(t) + 0.5]}{[x_1(t) + 2]^2} \times \exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right] + 5p_1^2(t)[x_1(t) + 0.25] \tag{6.3-30}$$

$$\dot{p}_2(t) = -2x_2(t) + p_2(t) + [p_2(t) - p_1(t)]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right]$$

as the reduced state-costate equations.

The influence function matrix differential equations are

$$\dot{\mathbf{P}}_x(t) = \begin{bmatrix} -2 + \dfrac{50[x_2 + 0.5]}{[x_1 + 2]^2}\alpha - \dfrac{[x_1 + 0.25]p_1}{R}, & \alpha \\ \dfrac{-50[x_2 + 0.5]}{[x_1 + 2]^2}\alpha, & -1 - \alpha \end{bmatrix}_t \mathbf{P}_x(t) + \begin{bmatrix} \dfrac{-[x_1 + 0.25]^2}{2R}, & 0 \\ 0 & 0 \end{bmatrix}_t \mathbf{P}_p(t)$$

$$\dot{\mathbf{P}}_p(t) = \begin{bmatrix} -2 + [p_2 - p_1]\left[\dfrac{100[23 - x_1][x_2 + 0.5]}{[x_1 + 2]^4}\right]\alpha + \dfrac{p_1^2}{2R}, & \dfrac{50[p_2 - p_1]}{[x_1 + 2]^2}\alpha \\ \dfrac{50[p_2 - p_1]}{[x_1 + 2]^2}\alpha, & -2 \end{bmatrix}_t \mathbf{P}_x(t) + \begin{bmatrix} 2 - \dfrac{50[x_2 + 0.5]}{[x_1 + 2]^2}\alpha + \dfrac{[x_1 + 0.25]p_1}{R}, & \dfrac{50[x_2 + 0.5]}{[x_1 + 2]^2}\alpha \\ -\alpha, & 1 + \alpha \end{bmatrix}_t \mathbf{P}_p(t), \dagger \tag{6.3-31}$$

where $\alpha \triangleq \exp\left[\dfrac{25x_1}{x_1 + 2}\right]$.

The boundary conditions for integrating these eight differential equations are $\mathbf{P}_x(0) = \mathbf{0}$, $\mathbf{P}_p(0) = \mathbf{I}$. The state and costate values appearing on the right side of (6.3-31) are obtained from the integration of the reduced state-costate equations with initial conditions $\mathbf{x}(0) = [0.05 \quad 0.00]^T$, $\mathbf{p}(0) = \mathbf{p}^{(i)}(0)$. After integrating (6.3-29) through (6.3-31) from $t = 0.0$ to $t = 0.78$, the matrix $\mathbf{P}_p(0.78)$ is used to determine $\mathbf{p}^{(i+1)}(0)$ from (6.3-20) with $h = 0$; that is,

† For simplicity the argument t has been omitted from the expressions involving $\mathbf{x}(t)$ and $\mathbf{p}(t)$, and the argument $\mathbf{p}^{(i)}(t_0)$ has been omitted from the influence functions.

$$\mathbf{p}^{(i+1)}(0) = \mathbf{p}^{(i)}(0) - [\mathbf{P}_p(0.78)]_i^{-1}\mathbf{p}^{(i)}(0.78). \tag{6.3-32}$$

The initial guess used to start the iterative procedure was

$$\mathbf{p}^{(0)}(0) = \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix}, \tag{6.3-33}$$

and

$$|p_1(0.78)| + |p_2(0.78)| \leq 10^{-5} \tag{6.3-34}$$

was the norm used as a stopping criterion.

The method converged after 5 iterations (with a norm of 1.71×10^{-6}) to the costate values

$$\mathbf{p}^*(0) = \begin{bmatrix} 1.0782 \\ 0.1918 \end{bmatrix},$$

which yield as the minimum value of the performance measure

$$J^* = 0.02660.$$

Figures 6-7 and 6-8 show the optimal control and trajectory. Comparing Figs. 6-7 and 6-8 with Figs. 6-4 and 6-5, we see that although the minimum value of J agrees to three decimal places with the steepest descent results, the trajectories and controls are discernibly different—indicating that the performance measure is rather insensitive to control variations in the vicinity of the optimum. Table 6-2 shows the costs and norm changes that occur during the iterative procedure.

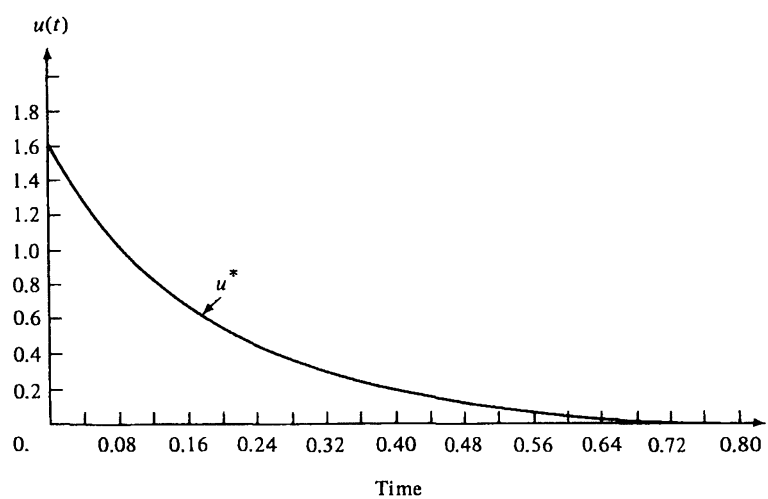


Figure 6-7 The optimal control for the stirred-tank reactor (variation of extremals solution)

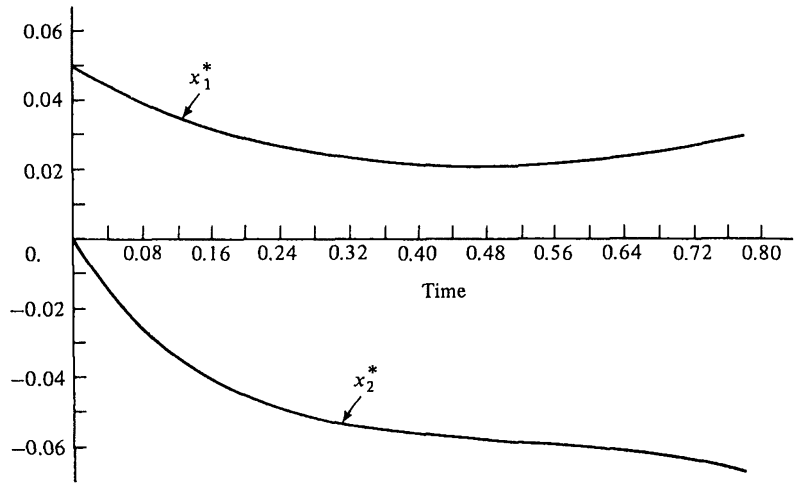


Figure 6-8 The optimal trajectory for the stirred-tank reactor (variation of extremals solution)

Table 6-2 VARIATION OF EXTREMALS SOLUTION OF THE STIRRED-TANK CHEMICAL REACTOR PROBLEM

Iteration	*Norm*	J†	p(0)	p(0.78)
0	2.97×10^{0}	0.09973	$\begin{bmatrix}1.0000\\0.5000\end{bmatrix}$	$\begin{bmatrix}2.3370\\0.6353\end{bmatrix}$
1	4.90×10^{-1}	0.04786	$\begin{bmatrix}1.2372\\0.2720\end{bmatrix}$	$\begin{bmatrix}0.4846\\0.0048\end{bmatrix}$
2	1.50×10^{-1}	0.02943	$\begin{bmatrix}1.0897\\0.2236\end{bmatrix}$	$\begin{bmatrix}0.1250\\0.0252\end{bmatrix}$
3	2.83×10^{-2}	0.02668	$\begin{bmatrix}1.0752\\0.1975\end{bmatrix}$	$\begin{bmatrix}0.0200\\0.0083\end{bmatrix}$
4	1.09×10^{-3}	0.02660	$\begin{bmatrix}1.0780\\0.1920\end{bmatrix}$	$\begin{bmatrix}0.0007\\0.0003\end{bmatrix}$
5	1.71×10^{-6}	0.02660	$\begin{bmatrix}1.0782\\0.1918\end{bmatrix}$	$\begin{bmatrix}0.0000012\\0.0000005\end{bmatrix}$

† Notice that this is the value of J associated with an extremal trajectory that satisfies the required boundary conditions only when the iterative procedure has converged.

An initial guess of

$$\mathbf{p}(0) = \begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix}$$

was also tried. The iterative procedure required 16 iterations to converge, and large excursions of the initial costate values (see Fig. 6-9) and the norm were observed. The norm for the initial guess was 11.77; after four iterations this value had grown to 933.69. These fluctuations and the number of iterations required for convergence point out the importance of making a good initial guess.

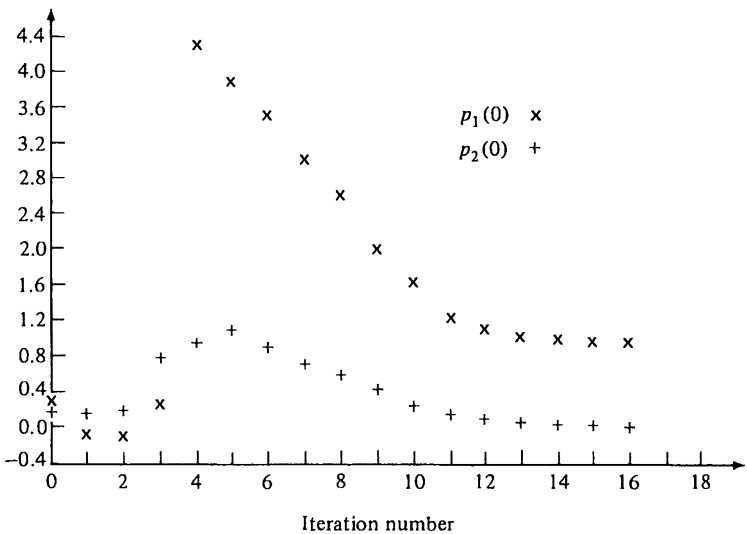


Figure 6-9 Variation of initial costate values for a starting guess $\mathbf{p}^{(0)}(0) = \begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix}$

Features of the Variation of Extremals Algorithm

As we did previously with the method of steepest descent, let us conclude our discussion of variation of extremals by reviewing the important characteristics of the algorithm.

Initial Guess. To begin the procedure, a guess for the initial costate $\mathbf{p}(t_0)$ must be made. [Actually, it may be better to guess $\mathbf{x}(t_f)$ instead, since we probably have more knowledge about the final values of the states from the physical nature of the problem. If we do elect to guess $\mathbf{x}(t_f)$ instead of $\mathbf{p}(t_0)$,

then we must integrate backward in time and modify the iteration equations appropriately to generate the next guess for $\mathbf{x}(t_f)$.]

Storage Requirements. No trajectories need to be stored; only the values of the influence function matrices at $t = t_f$, the value of $\mathbf{p}^{(i)}(t_0)$, the given initial state value, and the appropriate desired boundary conditions are retained in computer memory.

Convergence. Once $\mathbf{p}^{(i)}(t_0)$ is sufficiently close to $\mathbf{p}^*(t_0)$, the method of variation of extremals will generally converge quite rapidly; however, if the initial guess for $\mathbf{p}(t_0)$ is very poor, the method may not converge at all. Making a good initial guess is a difficult matter, because we have no physical insight to guide us in selecting $\mathbf{p}^{(0)}(t_0)$. The sensitivity of the solution of the differential equations is the culprit that makes the initial guess for $\mathbf{p}(t_0)$ so crucial; Fig. 6-10 illustrates how a small difference in the values of the initial costate may cause tremendous differences in the final values of the costate. Sometimes

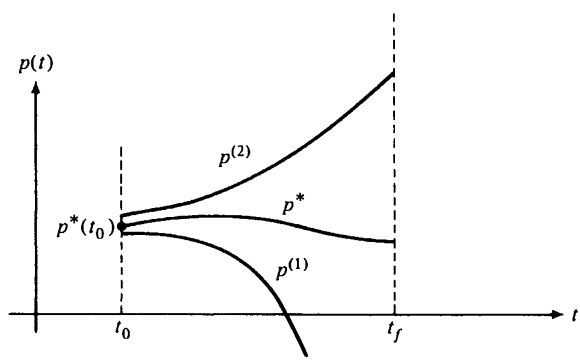


Figure 6-10 Sensitivity of the costate solution to changes in $p(t_0)$

this difficulty can be circumvented by using

$$\mathbf{p}^{(i+1)}(t_0) = \mathbf{p}^{(i)}(t_0) + \tau\left\{\left[\left[\frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}(t_f))\right]\mathbf{P}_x(\mathbf{p}(t_0), t_f) - \mathbf{P}_p(\mathbf{p}(t_0), t_f)\right]_i\right\}^{-1} \cdot \left[\mathbf{p}(t_f) - \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}(t_f))\right]_i, \tag{6.3-35}$$

where $0 < \tau \leq 1$ is a step size adjustment factor, instead of (6.3-20). To prevent the procedure from trying to correct the error in $\mathbf{p}(t_0)$ in one step, we can make τ small during the early iterations, and then gradually increase it to 1.0 as the procedure begins to converge.

Computations Required. $2n(n + 1)$ first-order differential equations must be numerically integrated and an $n \times n$ matrix inverted in each iteration.

Stopping Criterion. The iterative procedure is terminated when

$$\left\| \mathbf{p}^{(i)}(t_f) - \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t_f)) \right\| \leq \gamma,$$

where γ is a preselected positive termination constant.

Modifications Required for Fixed End Point Problems. If some or all of the final states are fixed, we must modify Eq. (6.3-20) to adjust $\mathbf{p}(t_0)$ iteratively on the basis of deviations of the final states from their desired values (see Problem 6-3).

6.4 QUASILINEARIZATION

In Chapter 4, we indicated that the combination of split boundary conditions and nonlinear differential equations is what makes nonlinear two-point boundary-value problems difficult to solve. Numerical integration can be used to solve *nonlinear* differential equations if a complete set of boundary conditions is specified at either the initial time or the final time—the method of variation of extremals consists of solving a sequence of such problems. In the method of quasilinearization, which we will introduce in this section, a sequence of *linear* two-point boundary-value problems is solved. Let us begin our discussion by illustrating a noniterative procedure for solving linear differential equations with split boundary conditions.

Solution of Linear Two-Point Boundary-Value Problems

Consider the two first-order, linear differential equations with time-varying coefficients

$$\begin{aligned} \dot{x}(t) &= a_{11}(t)x(t) + a_{12}(t)p(t) + e_1(t) \\ \dot{p}(t) &= a_{21}(t)x(t) + a_{22}(t)p(t) + e_2(t) \end{aligned} \tag{6.4-1}$$

with specified boundary conditions $x(t_0) = x_0$ and $p(t_f) = p_f$. a_{11}, a_{12}, a_{21}, a_{22}, e_1, and e_2 are known functions of time, and t_0, t_f, x_0, and p_f are known constants. It is desired to find a solution, $x^*(t)$, $p^*(t)$, $t \in [t_0, t_f]$, which satisfies the given boundary conditions. Notice that these differential equations are linear, but the boundary values are split.

First, suppose that we were to generate by numerical integration a solution, $x^H(t)$, $p^H(t)$, $t \in [t_0, t_f]$, of the homogeneous differential equations

$$\begin{aligned} \dot{x}(t) &= a_{11}(t)x(t) + a_{12}(t)p(t) \\ \dot{p}(t) &= a_{21}(t)x(t) + a_{22}(t)p(t) \end{aligned} \tag{6.4-2}$$

with arbitrary assumed values for the initial conditions; as a convenient choice, let $x^H(t_0) = 0$ and $p^H(t_0) = 1$.

Next, we could determine a particular solution, $x^P(t)$, $p^P(t)$, to the nonhomogeneous equations (6.4-1) by numerical integration, using $x^P(t_0) = x_0$ and $p^P(t_0) = 0$ as initial conditions. Since the differential equations are linear, the principle of superposition applies, and

$$\begin{aligned} x(t) &= c_1 x^H(t) + x^P(t) \\ p(t) &= c_1 p^H(t) + p^P(t) \end{aligned} \tag{6.4-3}$$

is a solution of (6.4-1) for any value of the constant c_1. We wish to find the solution that satisfies the specified boundary conditions. This can be accomplished by observing that

$$p(t_f) = p_f = c_1 p^H(t_f) + p^P(t_f). \tag{6.4-4}$$

Solving this for c_1, which is the only unknown quantity, gives

$$c_1 = \frac{p_f - p^P(t_f)}{p^H(t_f)}. \tag{6.4-5}$$

The required solution of (6.4-1) is then given by Eq. (6.4-3) with this value of c_1. Notice that the boundary condition $x(t_0) = x_0$ is satisfied for any choice of c_1 because of the judicious choices $x^H(t_0) = 0$ and $x^P(t_0) = x_0$.

The principle of superposition enabled us to obtain the solution of the linear two-point boundary-value problem in terms of the solution of a *linear algebraic equation*. The following example illustrates the required calculations.

Example 6.4-1. Find the solution of the differential equations

$$\begin{aligned} \dot{x}(t) &= -2x(t) - p(t) + 6 \\ \dot{p}(t) &= 4x(t) + 3p(t) \end{aligned} \tag{6.4-6}$$

which satisfies the boundary conditions $x(0) = 3$, $p(1) = 0$.

Proceeding as outlined previously, let $x^H(0) = 0$, $p^H(0) = 1$ be the initial conditions for the homogeneous equations

$$\begin{aligned} \dot{x}(t) &= -2x(t) - p(t) \\ \dot{p}(t) &= 4x(t) + 3p(t). \end{aligned} \tag{6.4-7}$$

Integrating Eq. (6.4-7) with these initial conditions gives

$$\begin{aligned} x^H(t) &= \tfrac{1}{3}\epsilon^{-t} - \tfrac{1}{3}\epsilon^{2t} \\ p^H(t) &= -\tfrac{1}{3}\epsilon^{-t} + \tfrac{4}{3}\epsilon^{2t}. \end{aligned} \tag{6.4-8}$$

To find a particular solution, we integrate (6.4-6) with the boundary conditions $x^p(0) = 3 = x(0)$, $p^p(0) = 0$, which yields the result

$$\begin{aligned} x^p(t) &= -4\epsilon^{-t} - 2\epsilon^{2t} + 9 \\ p^p(t) &= 4\epsilon^{-t} + 8\epsilon^{2t} - 12. \end{aligned} \tag{6.4-9}$$

The complete solution is then

$$\begin{aligned} x(t) &= c_1[\tfrac{1}{3}\epsilon^{-t} - \tfrac{1}{3}\epsilon^{2t}] - 4\epsilon^{-t} - 2\epsilon^{2t} + 9 \\ p(t) &= c_1[-\tfrac{1}{3}\epsilon^{-t} + \tfrac{4}{3}\epsilon^{2t}] + 4\epsilon^{-t} + 8\epsilon^{2t} - 12, \end{aligned} \tag{6.4-10}$$

where the value of c_1 that makes $p(1) = 0$ is to be determined. Notice that $x(0) = 3$ regardless of the value of c_1. To find c_1 we need only solve the linear algebraic equation

$$0 = c_1[-\tfrac{1}{3}\epsilon^{-1} + \tfrac{4}{3}\epsilon^{2}] + 4\epsilon^{-1} + 8\epsilon^{2} - 12, \tag{6.4-11}$$

which gives $c_1 = -4.993$. The solution is then

$$\begin{aligned} x(t) &= -0.336\epsilon^{2t} - 5.664\epsilon^{-t} + 9 \\ p(t) &= 1.342\epsilon^{2t} + 5.664\epsilon^{-t} - 12. \end{aligned} \tag{6.4-12}$$

It is easily verified by substitution that this is a solution of the original differential equations and comparing (6.4-12) with (6.3-16) we observe that this is the same solution as was obtained by using variation of extremals in Example 6.3-1.

In this particular problem the differential equations were simple enough so that numerical integration was not required; unfortunately, this is not generally the case. It should also be pointed out that the values assumed for $x^H(0)$, $p^H(0)$, $x^p(0)$, and $p^p(0)$ require that only one algebraic equation be solved; however, other initial values can also be used to obtain the same result for the complete solution.

The preceding discussion can be extended to include systems of *linear* differential equations of arbitrary order, but before doing this let us discuss how the linear differential equations arise.

Linearization of the Reduced State-Costate Equations

In general, the state and costate differential equations are nonlinear. Let us consider a simple situation in which there is one state equation and one costate equation. Assume that $\partial\mathscr{H}/\partial u = 0$ has been solved for $u(t)$ and substituted in the state-costate equations to obtain the reduced differential equations

$$\begin{aligned} \dot{x}(t) &= a(x(t), p(t), t) \\ \dot{p}(t) &= d(x(t), p(t), t) \end{aligned} \tag{6.4-13}$$

where a and d are nonlinear functions of $x(t)$, $p(t)$, and t. Let $x^{(0)}(t)$, $p^{(0)}(t)$, $t \in [t_0, t_f]$, be a known trajectory and $x^{(1)}(t)$, $p^{(1)}(t)$, $t \in [t_0, t_f]$, be any other trajectory. By performing a Taylor series expansion of the differential equations (6.4-13) about $x^{(0)}(t)$, $p^{(0)}(t)$ and retaining only terms of up to first order, we obtain

$$\begin{aligned}
\dot{x}^{(1)}(t) \doteq \dot{x}^{(0)}(t) &+ \left[\frac{\partial a}{\partial x}(x^{(0)}(t), p^{(0)}(t), t)\right][x^{(1)}(t) - x^{(0)}(t)] \\
&+ \left[\frac{\partial a}{\partial p}(x^{(0)}(t), p^{(0)}(t), t)\right][p^{(1)}(t) - p^{(0)}(t)] \\
\dot{p}^{(1)}(t) \doteq \dot{p}^{(0)}(t) &+ \left[\frac{\partial d}{\partial x}(x^{(0)}(t), p^{(0)}(t), t)\right][x^{(1)}(t) - x^{(0)}(t)] \\
&+ \left[\frac{\partial d}{\partial p}(x^{(0)}(t), p^{(0)}(t), t)\right][p^{(1)}(t) - p^{(0)}(t)]
\end{aligned} \tag{6.4-14}$$

or, substituting $a(x^{(0)}(t), p^{(0)}(t), t)$ for $\dot{x}^{(0)}(t)$ and $d(x^{(0)}(t), p^{(0)}(t), t)$ for $\dot{p}^{(0)}(t)$,

$$\begin{aligned}
\dot{x}^{(1)}(t) &= a(x^{(0)}(t), p^{(0)}(t), t) \\
&\quad + \left[\frac{\partial a}{\partial x}(x^{(0)}(t), p^{(0)}(t), t)\right][x^{(1)}(t) - x^{(0)}(t)] \\
&\quad + \left[\frac{\partial a}{\partial p}(x^{(0)}(t), p^{(0)}(t), t)\right][p^{(1)}(t) - p^{(0)}(t)] \\
\dot{p}^{(1)}(t) &= d(x^{(0)}(t), p^{(0)}(t), t) \\
&\quad + \left[\frac{\partial d}{\partial x}(x^{(0)}(t), p^{(0)}(t), t)\right][x^{(1)}(t) - x^{(0)}(t)] \\
&\quad + \left[\frac{\partial d}{\partial p}(x^{(0)}(t), p^{(0)}(t), t)\right][p^{(1)}(t) - p^{(0)}(t)].
\end{aligned} \tag{6.4-14a}$$

To demonstrate that these equations are linear, we can rewrite them, using the fact that $x^{(0)}$ and $p^{(0)}$ are *known* functions of time, as

$$\begin{aligned}
\dot{x}^{(1)}(t) &= a_{11}(t)x^{(1)}(t) + a_{12}(t)p^{(1)}(t) + e_1(t) \\
\dot{p}^{(1)}(t) &= a_{21}(t)x^{(1)}(t) + a_{22}(t)p^{(1)}(t) + e_2(t),
\end{aligned} \tag{6.4-15}$$

where

$$a_{11}(t) \triangleq \frac{\partial a}{\partial x}(x^{(0)}(t), p^{(0)}(t), t), \qquad a_{12}(t) \triangleq \frac{\partial a}{\partial p}(x^{(0)}(t), p^{(0)}(t), t),$$

$$a_{21}(t) \triangleq \frac{\partial d}{\partial x}(x^{(0)}(t), p^{(0)}(t), t), \qquad a_{22}(t) \triangleq \frac{\partial d}{\partial p}(x^{(0)}(t), p^{(0)}(t), t),$$

$$\begin{aligned}
e_1(t) \triangleq a(x^{(0)}(t), p^{(0)}(t), t) &- \left[\frac{\partial a}{\partial x}(x^{(0)}(t), p^{(0)}(t), t)\right]x^{(0)}(t) \\
&- \left[\frac{\partial a}{\partial p}(x^{(0)}(t), p^{(0)}(t), t)\right]p^{(0)}(t)
\end{aligned}$$

and

$$e_2(t) \triangleq d(x^{(0)}(t), p^{(0)}(t), t) - \left[\frac{\partial d}{\partial x}(x^{(0)}(t), p^{(0)}(t), t)\right] x^{(0)}(t) - \left[\frac{\partial d}{\partial p}(x^{(0)}(t), p^{(0)}(t), t)\right] p^{(0)}(t)$$

are all known functions of time.

By expanding the differential equations (6.4-13) about a trajectory $x^{(0)}$, $p^{(0)}$, we have obtained a set of *ordinary, linear, time-varying, nonhomogeneous differential equations;* these linear differential equations can be solved by using the procedure we have discussed previously in this section.

One Iteration of the Numerical Procedure

The method of quasilinearization consists of solving a *sequence* of linearized two-point boundary-value problems. We now know how to:

1. Linearize nonlinear differential equations.
2. Solve linear two-point boundary-value problems.

The following example illustrates how these two steps go together to constitute one iteration of the quasilinearization algorithm.

Example 6.4-2. A nonlinear first-order system is described by the differential equation

$$\dot{x}(t) = x^2(t) + u(t). \tag{6.4-16}$$

The initial condition is $x(0) = 3.0$, and the performance measure to be minimized is

$$J = \int_0^1 [2x^2(t) + u^2(t)]\, dt. \tag{6.4-17}$$

From the Hamiltonian

$$\mathcal{H}(x(t), u(t), p(t)) = 2x^2(t) + u^2(t) + p(t)x^2(t) + p(t)u(t), \tag{6.4-18}$$

the costate equation is

$$\dot{p}(t) = -\frac{\partial \mathcal{H}}{\partial x} = -4x(t) - 2p(t)x(t). \tag{6.4-19}$$

The algebraic relationship that must be satisfied is

$$\frac{\partial \mathcal{H}}{\partial u} = 0 = 2u(t) + p(t). \tag{6.4-20}$$

Observe that $\mathscr{H}$ is quadratic in $u(t)$, and

$$\frac{\partial^2 \mathscr{H}}{\partial u^2} = 2 > 0, \tag{6.4-21}$$

so that

$$u(t) = -\tfrac{1}{2}p(t) \tag{6.4-22}$$

is guaranteed to minimize the Hamiltonian. The nonlinear two-point boundary-value problem that is to be solved is then specified by the reduced state-costate equations

$$\begin{aligned} \dot{x}(t) &= x^2(t) - \tfrac{1}{2}p(t) \\ \dot{p}(t) &= -4x(t) - 2p(t)x(t), \end{aligned} \tag{6.4-23}$$

the boundary condition $x(0) = 3.0$, and, from equation (5.1-18), $p(1) = 0$.

Linearization of the reduced differential equations (6.4-23) about a nominal trajectory $x^{(0)}, p^{(0)}$ gives

$$\begin{aligned} \dot{x}^{(1)}(t) &= [x^{(0)}(t)]^2 - \tfrac{1}{2}p^{(0)}(t) + 2x^{(0)}(t)[x^{(1)}(t) - x^{(0)}(t)] \\ &\quad - \tfrac{1}{2}[p^{(1)}(t) - p^{(0)}(t)] \\ \dot{p}^{(1)}(t) &= -4x^{(0)}(t) - 2p^{(0)}(t)x^{(0)}(t) \\ &\quad - [4 + 2p^{(0)}(t)][x^{(1)}(t) - x^{(0)}(t)] \\ &\quad - 2x^{(0)}(t)[p^{(1)}(t) - p^{(0)}(t)], \end{aligned} \tag{6.4-24}$$

which, when rearranged, becomes

$$\begin{aligned} \dot{x}^{(1)}(t) &= [2x^{(0)}(t)]x^{(1)}(t) - \tfrac{1}{2}p^{(1)}(t) - [x^{(0)}(t)]^2 \\ \dot{p}^{(1)}(t) &= -[4 + 2p^{(0)}(t)]x^{(1)}(t) - [2x^{(0)}(t)]p^{(1)}(t) \\ &\quad + [2x^{(0)}(t)p^{(0)}(t)], \end{aligned} \tag{6.4-24a}$$

where each of the bracketed quantities is a known function of time. Notice that these differential equations are of the form given by (6.4-1), and hence can be solved for $x^{(1)}(t), p^{(1)}(t)$, $t \in [0, 1]$, by guessing $x^{(0)}(t), p^{(0)}(t)$, $t \in [0, 1]$, and using the procedure described previously. The new trajectory $x^{(1)}, p^{(1)}$ can then be used in place of $x^{(0)}, p^{(0)}$ to repeat the process.

Let us now discuss the generalization of these steps to systems of $2n$ differential equations. As expected, the generalization leads to similar equations, but with matrices replacing scalar quantities.

Assume that $\partial\mathscr{H}/\partial\mathbf{u} = \mathbf{0}$ has been solved for $\mathbf{u}(t)$ and substituted in the state and costate equations to obtain the reduced differential equations

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{p}(t), t) \tag{6.4-25}$$

$$\dot{\mathbf{p}}(t) = -\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}(t), \mathbf{p}(t), t). \tag{6.4-26}$$

The specified boundary conditions are $\mathbf{x}(t_0) = \mathbf{x}_0$ and $\mathbf{p}(t_f) = \mathbf{p}_f$, where $\mathbf{p}_f$ is an $n \times 1$ matrix of *constants;* t_f is assumed to be specified, and $\mathbf{x}(t_f)$ is free.

The first step is to linearize the differential equations (6.4-25) and (6.4-26). This is accomplished by expanding these differential equations in a Taylor series about a known trajectory $\mathbf{x}^{(i)}(t)$, $\mathbf{p}^{(i)}(t)$, $t \in [t_0, t_f]$, and retaining only terms of up to first order. The linearized reduced differential equations are

$$\begin{aligned}\dot{\mathbf{x}}^{(i+1)}(t) &= \mathbf{a}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t) \\ &\quad + \left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t)\right]\left[\mathbf{x}^{(i+1)}(t) - \mathbf{x}^{(i)}(t)\right] \\ &\quad + \left[\frac{\partial \mathbf{a}}{\partial \mathbf{p}}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t)\right]\left[\mathbf{p}^{(i+1)}(t) - \mathbf{p}^{(i)}(t)\right] \qquad (6.4\text{-}27)\end{aligned}$$

$$\begin{aligned}\dot{\mathbf{p}}^{(i+1)}(t) &= -\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t) \\ &\quad - \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}^2}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t)\right]\left[\mathbf{x}^{(i+1)}(t) - \mathbf{x}^{(i)}(t)\right] \\ &\quad - \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}\partial \mathbf{p}}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t)\right]\left[\mathbf{p}^{(i+1)}(t) - \mathbf{p}^{(i)}(t)\right], \qquad (6.4\text{-}28)\end{aligned}$$

where the jkth elements of the indicated matrices are

$$\left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}\right]_{jk} = \frac{\partial a_j}{\partial x_k}, \qquad \left[\frac{\partial \mathbf{a}}{\partial \mathbf{p}}\right]_{jk} = \frac{\partial a_j}{\partial p_k}, \qquad \left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}^2}\right]_{jk} = \frac{\partial^2 \mathscr{H}}{\partial x_j \partial x_k},$$

and

$$\left[\frac{\partial^2 \mathscr{H}}{\partial \mathbf{x}\partial \mathbf{p}}\right]_{jk} = \frac{\partial^2 \mathscr{H}}{\partial x_j \partial p_k}.$$

Notice that the differential equations (6.4-27) and (6.4-28) can be written

$$\dot{\mathbf{x}}^{(i+1)}(t) = \mathbf{A}_{11}(t)\mathbf{x}^{(i+1)}(t) + \mathbf{A}_{12}(t)\mathbf{p}^{(i+1)}(t) + \mathbf{e}_1(t) \qquad (6.4\text{-}27a)$$

$$\dot{\mathbf{p}}^{(i+1)}(t) = \mathbf{A}_{21}(t)\mathbf{x}^{(i+1)}(t) + \mathbf{A}_{22}(t)\mathbf{p}^{(i+1)}(t) + \mathbf{e}_2(t), \qquad (6.4\text{-}28a)$$

or, in partitioned matrix form,

$$\begin{aligned}\begin{bmatrix}\dot{\mathbf{x}}^{(i+1)}(t) \\ \hline \dot{\mathbf{p}}^{(i+1)}(t)\end{bmatrix} &= \left[\begin{array}{c|c}\mathbf{A}_{11}(t) & \mathbf{A}_{12}(t) \\ \hline \mathbf{A}_{21}(t) & \mathbf{A}_{22}(t)\end{array}\right]\begin{bmatrix}\mathbf{x}^{(i+1)}(t) \\ \hline \mathbf{p}^{(i+1)}(t)\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1(t) \\ \hline \mathbf{e}_2(t)\end{bmatrix} \\ &\triangleq \mathbf{A}(t)\begin{bmatrix}\mathbf{x}^{(i+1)}(t) \\ \hline \mathbf{p}^{(i+1)}(t)\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1(t) \\ \hline \mathbf{e}_2(t)\end{bmatrix}, \qquad (6.4\text{-}29)\end{aligned}$$

where the matrices $\mathbf{A}_{11}(t) \triangleq \partial \mathbf{a}/\partial \mathbf{x}$, $\mathbf{A}_{12}(t) \triangleq \partial \mathbf{a}/\partial \mathbf{p}$, $\mathbf{A}_{21}(t) \triangleq -\partial^2 \mathscr{H}/\partial \mathbf{x}^2$, $\mathbf{A}_{22}(t) \triangleq -\partial^2 \mathscr{H}/\partial \mathbf{x} \partial \mathbf{p}$, $\mathbf{e}_1(t) \triangleq -\mathbf{A}_{11}(t)\mathbf{x}(t) - \mathbf{A}_{12}(t)\mathbf{p}(t) + \mathbf{a}$, and $\mathbf{e}_2(t) \triangleq -\mathbf{A}_{21}(t)\mathbf{x}(t) - \mathbf{A}_{22}(t)\mathbf{p}(t) - \partial \mathscr{H}/\partial \mathbf{x}$ are evaluated at $\mathbf{x}^{(i)}(t)$, $\mathbf{p}^{(i)}(t)$ and hence are *known* functions of time.

An initial guess, $\mathbf{x}^{(0)}(t)$, $\mathbf{p}^{(0)}(t)$, $t \in [t_0, t_f]$, is used to evaluate these functions of time at the beginning of the first iteration. The next step is to generate *n solutions* to the $2n$ homogeneous differential equations

$$\begin{aligned} \dot{\mathbf{x}}^{(i+1)}(t) &= \mathbf{A}_{11}(t)\mathbf{x}^{(i+1)}(t) + \mathbf{A}_{12}(t)\mathbf{p}^{(i+1)}(t) \\ \dot{\mathbf{p}}^{(i+1)}(t) &= \mathbf{A}_{21}(t)\mathbf{x}^{(i+1)}(t) + \mathbf{A}_{22}(t)\mathbf{p}^{(i+1)}(t) \end{aligned} \tag{6.4-30}$$

by numerical integration. These solutions will be denoted by $\mathbf{x}^{H1}$, $\mathbf{p}^{H1}$; $\mathbf{x}^{H2}$, $\mathbf{p}^{H2}$; . . . ; $\mathbf{x}^{Hn}$, $\mathbf{p}^{Hn}$—the iteration superscript $(i+1)$ being understood. As boundary conditions for generating these solutions we shall use

$$\begin{aligned} \mathbf{x}^{H1}(t_0) &= \mathbf{0}, \qquad \mathbf{p}^{H1}(t_0) = [1 \quad 0 \quad 0 \quad \cdots \quad 0]^T \\ \mathbf{x}^{H2}(t_0) &= \mathbf{0}, \qquad \mathbf{p}^{H2}(t_0) = [0 \quad 1 \quad 0 \quad \cdots \quad 0]^T \\ &\vdots \qquad\qquad \vdots \\ \mathbf{x}^{Hn}(t_0) &= \mathbf{0}, \qquad \mathbf{p}^{Hn}(t_0) = [0 \quad 0 \quad \cdots \quad 0 \quad 1]^T. \end{aligned} \tag{6.4-31}$$

Next, we generate one particular solution, denoted by $\mathbf{x}^P$, $\mathbf{p}^P$, by numerically integrating Eq. (6.4-29) from t_0 to t_f, using the boundary conditions $\mathbf{x}^P(t_0) = \mathbf{x}_0$, $\mathbf{p}^P(t_0) = \mathbf{0}$. Using the principle of superposition, we find that the complete solution of (6.4-29) is of the form

$$\mathbf{x}^{(i+1)}(t) = c_1\mathbf{x}^{H1}(t) + c_2\mathbf{x}^{H2}(t) + \cdots + c_n\mathbf{x}^{Hn}(t) + \mathbf{x}^P(t) \tag{6.4-32}$$

$$\mathbf{p}^{(i+1)}(t) = c_1\mathbf{p}^{H1}(t) + c_2\mathbf{p}^{H2}(t) + \cdots + c_n\mathbf{p}^{Hn}(t) + \mathbf{p}^P(t), \tag{6.4-33}$$

where the values of $c_1, c_2, \ldots, c_n$ which make $\mathbf{p}^{(i+1)}(t_f) = \mathbf{p}_f$ are to be determined. To find the appropriate values of the c's, we let $t = t_f$ and write Eq. (6.4-33) as

$$\mathbf{p}_f = [\mathbf{p}^{H1}(t_f) \,|\, \mathbf{p}^{H2}(t_f) \,|\, \ldots \,|\, \mathbf{p}^{Hn}(t_f)]\mathbf{c} + \mathbf{p}^P(t_f), \tag{6.4-34}$$

where only $\mathbf{c} \triangleq [c_1 \quad c_2 \quad \ldots \quad c_n]^T$ is unknown. Solving for $\mathbf{c}$ yields

$$\mathbf{c} = [\mathbf{p}^{H1}(t_f) \,|\, \mathbf{p}^{H2}(t_f) \,|\, \ldots \,|\, \mathbf{p}^{Hn}(t_f)]^{-1}[\mathbf{p}_f - \mathbf{p}^P(t_f)]. \tag{6.4-35}$$

Substituting $\mathbf{c}$ of Eq. (6.4-35) into (6.4-32) and (6.4-33) gives the $(i+1)$st

trajectory—this completes one iteration of the quasilinearization algorithm. The $(i+1)$st trajectory can then be used to begin another iteration, if required.

Notice that if we let $t = t_0$ in Eqs. (6.4-32) and (6.4-33) and substitute the boundary conditions given by (6.4-31), then

$$\begin{aligned} \mathbf{x}^{(i+1)}(t_0) &= \mathbf{x}^p(t_0) = \mathbf{x}_0 \\ \mathbf{p}^{(i+1)}(t_0) &= \mathbf{c}. \end{aligned} \tag{6.4-36}$$

Thus, the solution $\mathbf{x}^{(i+1)}$, $\mathbf{p}^{(i+1)}$ satisfies the initial condition $\mathbf{x}^{(i+1)}(t_0) = \mathbf{x}_0$ regardless of the value of $\mathbf{c}$. In addition, the initial costate for the $(i+1)$st trajectory is the value of $\mathbf{c}$ obtained from Eq. (6.4-35); we shall subsequently make use of this information to reduce storage requirements.

In deriving Eq. (6.4-35) from (6.4-32) and (6.4-33) it was assumed that the final costate $\mathbf{p}(t_f)$ is a specified constant. If, however,

$$\mathbf{p}(t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}(t_f)), \tag{6.4-37}$$

then Eq. (6.4-35) must be modified to read (see Problem 6-10)

$$\begin{aligned} \mathbf{c} = [\mathbf{p}^{H1}(t_f) - \mathbf{M}\mathbf{x}^{H1}(t_f) \,|\, \mathbf{p}^{H2}(t_f) - \mathbf{M}\mathbf{x}^{H2}(t_f) \,|\, \cdots \,|\, \mathbf{p}^{Hn}(t_f) \\ - \mathbf{M}\mathbf{x}^{Hn}(t_f)]^{-1}\left[\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t_f)) - \mathbf{M}\mathbf{x}^{(i)}(t_f) + \mathbf{M}\mathbf{x}^p(t_f) - \mathbf{p}^p(t_f)\right], \end{aligned} \tag{6.4-38}$$

where

$$\mathbf{M} \triangleq \frac{\partial^2 h}{\partial \mathbf{x}^2}(\mathbf{x}^{(i)}(t_f)).$$

Notice that if h is a linear function of $\mathbf{x}(t_f)$, for example,

$$h(\mathbf{x}(t_f)) = \mathbf{v}^T\mathbf{x}(t_f), \tag{6.4-39}$$

where $\mathbf{v}^T$ is a $1 \times n$ matrix of constants, then $\partial^2 h/\partial \mathbf{x}^2 = \mathbf{0}$ and $\partial h/\partial \mathbf{x} = \mathbf{v}$. In this case, Eq. (6.4-38) reduces to (6.4-35) with $\mathbf{p}_f = \mathbf{v}$.

The Quasilinearization Algorithm

Let us now summarize the iterative procedure for solving nonlinear two-point boundary-value problems by using the method of quasilinearization:

1. Form the reduced differential equations by solving $\partial \mathscr{H}/\partial \mathbf{u} = \mathbf{0}$ for

$\mathbf{u}(t)$ in terms of $\mathbf{x}(t)$, $\mathbf{p}(t)$, t, and substituting in the state and costate equations (which then contain only $\mathbf{x}(t)$, $\mathbf{p}(t)$, and t).

2. Using (6.4-27) and (6.4-28), determine the linearized reduced differential equations in terms of $\mathbf{x}^{(i)}(t)$, $\mathbf{p}^{(i)}(t)$, t.
3. Guess an initial trajectory $\mathbf{x}^{(0)}(t)$, $\mathbf{p}^{(0)}(t)$, $t \in [t_0, t_f]$, and let the iteration index i be zero.
4. Evaluate the matrices $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, $\mathbf{A}_{22}$, $\mathbf{e}_1$, and $\mathbf{e}_2$ of Eq. (6.4-29) on the trajectory $\mathbf{x}^{(i)}$, $\mathbf{p}^{(i)}$.
5. Numerically integrate the linear homogeneous differential equations (6.4-30) from t_0 to t_f, using the n sets of initial conditions given in (6.4-31), to obtain n homogeneous solutions. Compute a particular solution to (6.4-29) by numerical integration from t_0 to t_f; using the initial conditions $\mathbf{x}^p(t_0) = \mathbf{x}_0$ and $\mathbf{p}^p(t_0) = \mathbf{0}$. Generally, the n homogeneous solutions and the one particular solution are calculated by performing a single integration of $n(2n) + 2n = 2n(n + 1)$ differential equations. Store the values of the appropriate variables at $t = t_f$.
6. Use the values found in step 5 to determine $\mathbf{c}$ from Eq. (6.4-38).
7. Use $\mathbf{c}$ found in step 6 and Eqs. (6.4-32) and (6.4-33) to determine the $(i + 1)$st trajectory.
8. Compare the ith and $(i + 1)$st trajectories by calculating the norm

$$\left\| \begin{bmatrix} \mathbf{x}^{(i+1)} \\ \cdots\cdots \\ \mathbf{p}^{(i+1)} \end{bmatrix} - \begin{bmatrix} \mathbf{x}^{(i)} \\ \cdots \\ \mathbf{p}^{(i)} \end{bmatrix} \right\| \triangleq \sum_{j=1}^{n} \{ \max_t | x_j^{(i+1)}(t) - x_j^{(i)}(t) | + \max_t | p_j^{(i+1)}(t) - p_j^{(i)}(t) | \}. \quad (6.4\text{-}40)\dagger$$

If

$$\left\| \begin{bmatrix} \mathbf{x}^{(i+1)} \\ \cdots\cdots \\ \mathbf{p}^{(i+1)} \end{bmatrix} - \begin{bmatrix} \mathbf{x}^{(i)} \\ \cdots \\ \mathbf{p}^{(i)} \end{bmatrix} \right\| \leq \gamma, \quad (6.4\text{-}41)$$

where γ is a preselected termination constant, the iterative procedure has converged; go to step 9. If the termination criterion is not satisfied, return to step 4, using the trajectory $\mathbf{x}^{(i+1)}$, $\mathbf{p}^{(i+1)}$ in place of $\mathbf{x}^{(i)}$, $\mathbf{p}^{(i)}$.

9. Integrate the original nonlinear reduced state and costate equations with initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$, $\mathbf{p}(t_0) = \mathbf{c}$. Compare the results of this integration with the final trajectory $\mathbf{x}^{(i+1)}$, $\mathbf{p}^{(i+1)}$, using a suitable norm, and also with the specified boundary values at $t = t_f$ to verify that the sequence of solutions to the linearized differential equations has converged to the solution of the nonlinear differential

† There are, of course, other acceptable choices for the norm.

equations (6.4-25) and (6.4-26). Evaluate the optimal control history from the state and costate values on the $(i+1)$st trajectory, and print out (or graph) the optimal trajectory and the optimal control.

Steps 1 through 3 are performed off-line by the user or programmer; steps 4 through 9 are executed on a digital computer.

The Continuous Stirred-Tank Chemical Reactor Problem

For comparison with the methods of steepest descent and variation of extremals, let us again solve the stirred-tank reactor problem discussed in Sections 6.2 and 6.3, this time using the quasilinearization algorithm.

Example 6.4-3. The problem statement is given in Example 6.2-2. The reduced differential equations are given by Eqs. (6.3-29) and (6.3-30). Linearizing these nonlinear differential equations, using (6.4-29), we obtain

$$\begin{bmatrix} \dot{\mathbf{x}}^{(i+1)}(t) \\ \hline \dot{\mathbf{p}}^{(i+1)}(t) \end{bmatrix} = \mathbf{A}(t) \begin{bmatrix} \mathbf{x}^{(i+1)}(t) \\ \hline \mathbf{p}^{(i+1)}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{a}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t) \\ \hline -\dfrac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(t), \mathbf{p}^{(i)}(t), t) \end{bmatrix} - \mathbf{A}(t) \begin{bmatrix} \mathbf{x}^{(i)}(t) \\ \hline \mathbf{p}^{(i)}(t) \end{bmatrix}. \tag{6.4-42}$$

$\mathbf{A}(t)$ denotes the $2n \times 2n$ matrix

$$\mathbf{A}(t) = \left[\begin{array}{c|c|c|c} -2-10p_1\alpha_5+\alpha_4 & \alpha_1 & -5\alpha_5^2 & 0 \\ -\alpha_4 & -1-\alpha_1 & 0 & 0 \\ \hline -2+\alpha_2\alpha_6+5p_1^2 & \alpha_3\alpha_6 & 2+10p_1\alpha_5-\alpha_4 & \alpha_4 \\ \alpha_3\alpha_6 & -2 & -\alpha_1 & 1+\alpha_1 \end{array}\right]_{\mathbf{x}^{(i)}(t),\, \mathbf{p}^{(i)}(t)}$$

where

$$\alpha_1 \triangleq \exp\left[\frac{25x_1}{x_1+2}\right]$$

$$\alpha_2 \triangleq \frac{100[x_2+0.5][23-x_1]\alpha_1}{[x_1+2]^4}$$

$$\alpha_3 \triangleq \frac{50\alpha_1}{[x_1+2]^2}$$

$$\alpha_4 \triangleq [x_2+0.5]\alpha_3.$$

$$\alpha_5 \triangleq x_1+0.25$$

$$\alpha_6 \triangleq p_2-p_1$$

Although these differential equations look formidable, their derivation is not difficult, only tedious. It should be emphasized that the $\mathbf{A}(t)$ matrix and the $2n \times 1$ matrix containing $\mathbf{a}$ and $\partial\mathscr{H}/\partial\mathbf{x}$ are evaluated on the ith trajectory; hence, in the $(i+1)$st iteration these are known functions of time.

To begin the iterative procedure, the nominal state-costate history

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ p_1(t) \\ p_2(t) \end{bmatrix} = \mathbf{0} \qquad \text{for } t \in [0, 0.78]$$

was selected. The norm used to measure the deviation of successive trajectories generated by the iterative process is given by Eq. (6.4-40) with $\gamma = 1 \times 10^{-3}$. For this initial guess, the procedure converged in four iterations to a minimum cost of

$$J^* = 0.02660$$

with a norm of 0.00044. As a check, the initial costates generated in the final iteration were used to integrate the original nonlinear differential equations (6.3-29) and (6.3-30), and the norm of the deviation of this trajectory from the trajectory generated in the last iteration was calculated to be 0.00073. The optimal trajectory and control history obtained by using quasilinearization are identical (to three decimal places) with the results obtained using the variation of extremals algorithm.

An initial state-costate history of

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ p_1(t) \\ p_2(t) \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.5 \\ -0.5 \\ -0.5 \end{bmatrix} \qquad \text{for } t \in [0, 0.78]$$

was also tried. With this initial guess, quasilinearization converged in nine iterations to a minimum cost of

$$J^* = 0.02660$$

with a final norm of the deviation between successive trajectories of 0.000002. Again, the optimal control history and its trajectory are essentially identical to those found by using variation of extremals. The results obtained by using each of these initial trajectories are summarized in Table 6-3.

Table 6-3 QUASILINEARIZATION SOLUTION OF THE STIRRED-TANK CHEMICAL REACTOR PROBLEM

	$\begin{bmatrix} \mathbf{x} \\ -- \\ \mathbf{p} \end{bmatrix} = \mathbf{0}$		$\begin{bmatrix} \mathbf{x} \\ --- \\ \mathbf{p} \end{bmatrix} = -\mathbf{0.5}$	
Iteration	*Norm*	*Initial costate*, c	*Norm*	*Initial costate*, c
1	1.3254	$\begin{bmatrix} 1.0235 \\ 0.1306 \end{bmatrix}$	2.4800	$\begin{bmatrix} -0.2912 \\ 0.1466 \end{bmatrix}$
2	0.2083	$\begin{bmatrix} 1.0714 \\ 0.1871 \end{bmatrix}$	1.7803	$\begin{bmatrix} -1.2543 \\ -0.2121 \end{bmatrix}$
3	0.0192	$\begin{bmatrix} 1.0778 \\ 0.1917 \end{bmatrix}$	2.2027	$\begin{bmatrix} 0.4651 \\ 0.1234 \end{bmatrix}$
4	0.0004	$\begin{bmatrix} 1.0780 \\ 0.1918 \end{bmatrix}$	1.1977	$\begin{bmatrix} 1.1099 \\ 0.1221 \end{bmatrix}$
5	—	—	0.4800	$\begin{bmatrix} 1.0587 \\ 0.1702 \end{bmatrix}$
6	—	—	0.1326	$\begin{bmatrix} 1.0741 \\ 0.1918 \end{bmatrix}$
7	—	—	0.0170	$\begin{bmatrix} 1.0780 \\ 0.1917 \end{bmatrix}$
8	—	—	0.0002	$\begin{bmatrix} 1.0780 \\ 0.1917 \end{bmatrix}$
9	—	—	0.0000	$\begin{bmatrix} 1.0780 \\ 0.1918 \end{bmatrix}$

Features of the Quasilinearization Method

To conclude our discussion of quasilinearization, let us summarize the important features of the algorithm.

Initial Guess. An initial state-costate trajectory $\mathbf{x}^{(0)}(t)$, $\mathbf{p}^{(0)}(t)$, $t \in [t_0, t_f]$, must be selected to begin the iterative procedure. This initial trajectory, which is used for linearizing the nonlinear reduced differential equations, does not necessarily have to satisfy the specified boundary conditions; all subsequent iterates will do so, however. The primary requirement of the initial guess is that it not be so poor that it causes the algorithm to diverge. As usual, the

initial guess is made primarily on the basis of whatever physical information is available about the particular problem being solved.

Storage Requirements. From Eq. (6.4-32) with $t = t_0$ it is apparent that $\mathbf{p}^{(i+1)}(t_0) = \mathbf{c}$ if the values of $\mathbf{p}^{H1}(t_0), \ldots, \mathbf{p}^{Hn}(t_0)$, and $\mathbf{p}^{P}(t_0)$ are selected as suggested; therefore, once $\mathbf{c}$ is known, the $(i+1)$st trajectory can be generated by reintegrating Eq. (6.4-29) with the initial conditions $\mathbf{x}^{(i+1)}(t_0) = \mathbf{x}_0$ and $\mathbf{p}^{(i+1)}(t_0) = \mathbf{c}$. By doing this, there is no necessity for storing (presumably in piecewise-constant fashion) the n homogeneous solutions and the particular solution; hence we store only the linearizing state-costate trajectory, the specified value of $\mathbf{x}(t_0)$, the value of $\mathbf{c}$, $\mathbf{x}^{P}(t_f)$, $\mathbf{p}^{P}(t_f)$, and $\mathbf{x}^{HJ}(t_f)$, $\mathbf{p}^{HJ}(t_f)$, $j = 1, 2, \ldots, n$.

Convergence. McGill and Kenneth [M-4] have proved that the sequence of solutions of the linearized equations (6.4-29) converges (with a rate that is at least quadratic) to the solution of the nonlinear differential equations (6.4-25) and (6.4-26) if

1. The functions $\mathbf{a}$ and $\partial\mathscr{H}/\partial\mathbf{x}$ of Eqs. (6.4-25) and (6.4-26) are continuous.
2. The partial derivatives $\partial\mathbf{a}/\partial\mathbf{x}$, $\partial\mathbf{a}/\partial\mathbf{p}$, $\partial^2\mathscr{H}/\partial\mathbf{x}^2$, and $\partial^2\mathscr{H}/\partial\mathbf{x}\partial\mathbf{p}$ of Eqs. (6.4-27) and (6.4-28) exist and are continuous.
3. The partial derivative functions in 2 satisfy a Lipschitz condition with respect to $[\mathbf{x}(t) \mid \mathbf{p}(t)]^T$.
4. The norm of the deviation of the initial guess from the solutions of (6.4-25) and (6.4-26) is sufficiently small.

Computations Required. The integration of $2n(n+1)$ first-order linear differential equations and the inversion of an $n \times n$ matrix are required in each iteration. If the $(i+1)$st trajectory is generated by integration as discussed previously, an additional $2n$ linear differential equations must be integrated.

Stopping Criterion. The method of quasilinearization involves successive approximations to the solution of a system of nonlinear differential equations by a sequence of solutions of a system of linear differential equations. To ascertain whether or not the procedure has converged, a measure of the deviation of adjacent members of the sequence is used. For example, McGill and Kenneth [M-5] use the norm

$$M = \sum_{j=1}^{n} \{\max_t |x_j^{(i+1)}(t) - x_j^{(i)}(t)| + \max_t |p_j^{(i+1)}(t) - p_j^{(i)}(t)|\}, \quad (6.4\text{-}43)$$

where $x_j^{(i+1)}$ is the jth component of the state vector generated in the $(i+1)$st iteration. When two successive trajectories yield a value of M that is smaller than some preselected number γ, the iterative procedure is terminated—the

sequence of solutions of the approximating linear differential equations has converged. It remains to verify that the final iterate in this sequence of solutions has converged to the solution of the original nonlinear differential equations. We can accomplish this by integrating the nonlinear differential equations (6.4-25) and (6.4-26), using the value for $\mathbf{p}(t_0)$ determined in the final iteration and the specified value of $\mathbf{x}(t_0)$ as initial conditions. The boundary values obtained at $t = t_f$ from this numerical integration are then compared with the specified values at $t = t_f$ to verify that the solution to the nonlinear two-point boundary-value problem has been obtained.

Modifications for Fixed End Point Problems. We have discussed problems in which the final states are free; however, the quasilinearization algorithm is easily modified to deal with problems in which some or all of the states are specified at $t = t_f$. For example, suppose $\mathbf{x}(t_f)$ is specified. To determine $\mathbf{c}$, we solve Eq. (6.4-32) with $t = t_f$ rather than solve Eq. (6.4-33). If some of the components of $\mathbf{x}(t_f)$ are fixed and others free, we select the appropriate equations among (6.4-32) and (6.4-33), let $t = t_f$, and solve for $\mathbf{c}$.

6.5 SUMMARY OF ITERATIVE TECHNIQUES FOR SOLVING TWO-POINT BOUNDARY-VALUE PROBLEMS

So far, in this chapter we have considered three iterative numerical methods for the solution of nonlinear two-point boundary-value problems. The assumption was made that the states and controls are not constrained by any boundaries; if this is not the case, the computational techniques we have discussed must be modified.†

In each of the methods we have considered, the philosophy is to solve a sequence of problems in which one or more of the five necessary conditions [Eqs. (6.1-1) through (6.1-4)] is initially violated, but eventually satisfied if the iterative procedure converges. In the steepest descent method the algorithm terminates when $\partial\mathscr{H}/\partial\mathbf{u} \approx \mathbf{0}$ for all $t \in [t_0, t_f]$, the other four conditions having been satisfied throughout the iterative procedure. Convergence of the method of variation of extremals is indicated when the boundary condition $\mathbf{p}(t_f) = \partial h(\mathbf{x}(t_f))/\partial\mathbf{x}$ is satisfied. In quasilinearization, trajectories are generated that satisfy the boundary conditions; when a trajectory is obtained that also is a solution of the reduced state-costate equations, the procedure has converged.

As bases for comparing the numerical techniques, we have used the initial guess requirement, storage requirements, convergence properties,

† See [S-3].

computational requirements, stopping criteria, and modifications for fixed end point problems. Table 6-4 summarizes these and other characteristics of the three iterative methods.

It should be emphasized that the numerical techniques we have discussed

Table 6-4 A COMPARISON OF THE FEATURES OF THREE ITERATIVE METHODS FOR SOLVING NONLINEAR TWO-POINT BOUNDARY-VALUE PROBLEMS

Feature	*Steepest descent*	*Variation of extremals*	*Quasilinearization*
Initial guess	$\mathbf{u}(t)$, $t \in [t_0, t_f]$	$\mathbf{p}(t_0)$ [or $\mathbf{x}(t_f)$]	$\mathbf{x}(t)$, $\mathbf{p}(t)$, $t \in [t_0, t_f]$
Iterate to satisfy	$\frac{\partial \mathscr{H}}{\partial \mathbf{u}} \equiv \mathbf{0}$	$\mathbf{p}(t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}(t_f))$	State and costate equations
Importance of initial guess	Not usually crucial to convergence	Divergence may result from poor guess	Divergence may result from poor guess
Storage requirements	$\mathbf{u}^{(i)}(t)$, $\mathbf{x}^{(i)}(t)$, and $\frac{\partial \mathscr{H}^{(i)}}{\partial \mathbf{u}}(t)$, $t \in [t_0, t_f]$	$2[n \times n]$ matrices, boundary conditions	$\mathbf{x}^{(i)}(t)$, $\mathbf{p}^{(i)}(t)$, $t \in [t_0, t_f]$, $n \times n$ matrix, boundary conditions, $\mathbf{c}$
Convergence	Approaches a minimum rapidly, then slows down drastically	Once convergence begins (if it does), it is rapid	Converges quadratically in the vicinity of the optimum
Computations required	Integration of $2n$ differential equations, calculation of $\partial \mathscr{H}/\partial \mathbf{u}$, step size.	Integration of $2n(n+1)$ first-order differential equations, inversion of an $n \times n$ matrix.	Integration of $2n(n+1)$ first-order differential equations, inversion of an $n \times n$ matrix.
Modifications for fixed end point problems	Penalty function or see [B-5]	Adjust $\mathbf{p}(t_0)$ based on calculated values of $\mathbf{x}(t_f)$.	Solve for $\mathbf{c}$ from equation for $\mathbf{x}(t_f)$

may not always converge, and even if convergence occurs it may be only to a local minimum. By trying several different initial guesses, we can be reasonably sure of locating any other local minima that may exist, or, if the numerical procedure converges to the same control and trajectory for a variety of initial guesses, we have some assurance that a global minimum has been determined.

The difficulty of solving nonlinear two-point boundary-value problems has made iterative numerical techniques the subject of continuing research. When one is confronted with a problem of this type, it is useful to be familiar with many different techniques, perhaps trying several methods on a given problem, or a hybrid scheme may be useful. For example, the steepest de-

scent method may be used as a starting procedure and quasilinearization to close in on the solution.

6.6 GRADIENT PROJECTION

In this section we shall discuss an alternative approach to optimization introduced by J. B. Rosen [R-4, 5, 6] which does not involve the solution of nonlinear two-point boundary-value problems. Rosen's method, called gradient projection, is an iterative numerical procedure for finding an extremum of a function of several variables that are required to satisfy various constraining relations. If the function to be extremized (called the *objective function*) and the constraints are linear functions of the variables, the optimization problem is referred to as a *linear programming problem*; when nonlinear terms are present in the constraining relations or in the objective function, the problem is referred to as a *nonlinear programming problem.*

We shall first discuss gradient projection as it applies to nonlinear programming problems that have linear constraints, but nonlinear objective functions. Then we shall show how the gradient projection algorithm can be used to solve optimal control problems.

Minimization of Functions by the Gradient Projection Method

Example 6.6-1. To begin, let us consider a simple example. Let f be a function of two variables y_1 and y_2 and $f(y_1, y_2)$ denote the value of f at the point (y_1, y_2). The problem is to find the point (y_1^*, y_2^*) where f has its minimum value. The variables y_1 and y_2 are required to satisfy the linear inequality constraints

$$y_1 \geq 0 \tag{6.6-1a}$$

$$y_2 \geq 0 \tag{6.6-1b}$$

$$2y_1 - 5y_2 + 10 \geq 0 \tag{6.6-1c}$$

$$-4y_1 - 7y_2 + 22.5 \geq 0 \tag{6.6-1d}$$

$$-9y_1 - 2y_2 + 26.5 \geq 0 \tag{6.6-1e}$$

The set of points that satisfy all of these constraints is denoted by R and called the *admissible region.*† For this example, R is the interior and the boundary of the region whose boundary is determined by the lines labelled

† In the nomenclature of nonlinear programming the term *feasible* is used rather than *admissible.*

H_1, H_2, H_3, H_4, and H_5 in Fig. 6-11. H_1 is the line determined by the equation $y_1 = 0$, H_2 by the equation $y_2 = 0$, H_3 by the equation $2y_1 - 5y_2 + 10 = 0$, and so forth. Also shown in Fig. 6-11 are several equal-value contours of the function f. $-\partial f^{(i)}/\partial \mathbf{y}$ denotes $-\partial f(\mathbf{y}^{(i)})/\partial \mathbf{y}$ [the negative gradient of f at the point $\mathbf{y}^{(i)}$] and $\mathbf{P}[-\partial f^{(i)}/\partial \mathbf{y}]$, which is a projection of the vector $-\partial f^{(i)}/\partial \mathbf{y}$, is called the *gradient projection*.

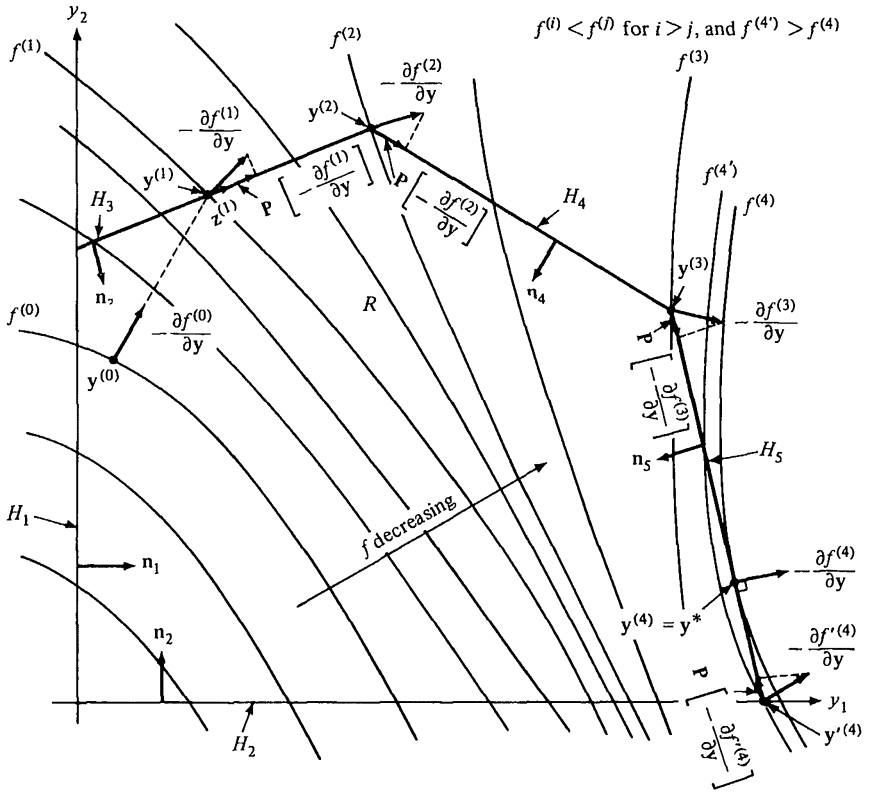


Figure 6-11 Gradient projection minimization of a function of two variables

Assume that the initial point $\mathbf{y}^{(0)}$ is in the admissible region as shown. The first step is to determine the gradient at the point $\mathbf{y}^{(0)}$. Since a minimum is sought, $\mathbf{y}$ is to be changed in the negative gradient direction as far as possible without violating any constraints, or until the function values begin to increase, whichever occurs first. In this example, $\mathbf{y}$ is changed in the direction of the vector $-\partial f^{(0)}/\partial \mathbf{y}$ until the line H_3, which is on the boundary of the admissible region, is encountered at the point $\mathbf{y}^{(1)}$. The negative gradient at $\mathbf{y}^{(1)}$ is $-\partial f^{(1)}/\partial \mathbf{y}$, as shown; however, if $\mathbf{y}$ were to be changed in the direction of $-\partial f^{(1)}/\partial \mathbf{y}$, the constraint H_3 would be violated, so we change $\mathbf{y}$ along the line H_3 in the direction of the *projection* onto H_3, $\mathbf{P}[-\partial f^{(1)}/\partial \mathbf{y}]$, of the vector $-\partial f^{(1)}/\partial \mathbf{y}$. $\mathbf{y}$ is changed in this direction until the point $\mathbf{y}^{(2)}$ at the intersection of H_3 and H_4 is

reached.† The next move is along the projection onto H_4 of $-\partial f^{(2)}/\partial \mathbf{y}$ to the point $\mathbf{y}^{(3)}$; notice that the function values encountered continue to decrease. From the point $\mathbf{y}^{(3)}$, we change $\mathbf{y}$ in the direction of the projected gradient $\mathbf{P}[-\partial f^{(3)}/\partial \mathbf{y}]$ until the point $\mathbf{y}'^{(4)}$ is reached. Upon evaluating the gradient at $\mathbf{y}'^{(4)}$, it is found that the projection $\mathbf{P}[-\partial f'^{(4)}/\partial \mathbf{y}]$ indicates a move back toward $\mathbf{y}^{(3)}$. By repeated interpolation along the line H_5, the point $\mathbf{y}^{(4)} = \mathbf{y}^*$, where f assumes its minimum value, is determined. Observe that $-\partial f^{(4)}/\partial \mathbf{y}$ is normal to H_5 and directed toward the inadmissible region, indicating that no further improvement can be obtained by moving along H_5, or by moving into the interior of the admissible region.

The preceding example illustrates the basic idea of the gradient projection algorithm: We change $\mathbf{y}$ in the direction of steepest descent until a minimum of the objective function is found. If moving in the direction of steepest descent would cause any of the constraints to be violated, $\mathbf{y}$ is changed along the projection of the negative gradient onto the boundary of the admissible region.

Let us now generalize this procedure to apply to a nonlinear function f of K variables, $y_1, y_2, \ldots, y_K$. Although the discussion applies to problems where the dimension K is arbitrary, we shall illustrate the concepts geometrically with two- and three-dimensional examples.

Fundamental Concepts and Definitions. The value of the objective function at the point $\mathbf{y}$ is denoted by $f(\mathbf{y})$, where $\mathbf{y}$ is a K vector. It is assumed that f is convex‡ and has continuous second partial derivatives in the admissible

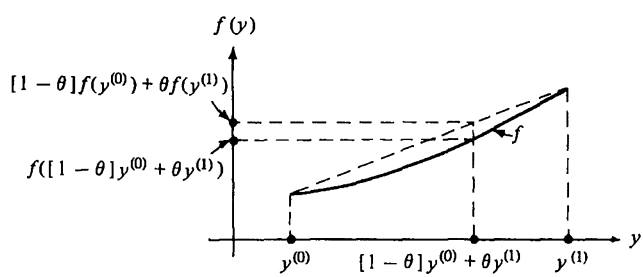


Figure 6-12 A convex function

† By intersection of H_3 and H_4 we mean the points that are on both H_3 and H_4; in this example the intersection is the single point $\mathbf{y}^{(2)}$.

‡ $f(\mathbf{y})$ is a *convex function* in the region R if

$$[1 - \theta]f(\mathbf{y}^{(0)}) + \theta f(\mathbf{y}^{(1)}) \geq f([1 - \theta]\mathbf{y}^{(0)} + \theta \mathbf{y}^{(1)}) \tag{6.6-2}$$

for $0 \leq \theta \leq 1$ and for all $\mathbf{y}^{(0)}$ and $\mathbf{y}^{(1)}$ in R. Equation (6.6-2) implies that linear interpolation between any two points $\mathbf{y}^{(0)}$ and $\mathbf{y}^{(1)}$ yields a value at least as large as the actual value of the function at the point of interpolation. Figure 6-12 shows a convex function of one variable.

region R. The variables $y_1, \ldots, y_K$ are constrained by L linear inequalities of the form

$$\sum_{j=1}^{K} n_{ji} y_j - v_i \geq 0, \qquad i = 1, 2, \ldots, L, \tag{6.6-3}$$

where the n_{ji} have been normalized so that

$$\sum_{j=1}^{K} (n_{ji})^2 = 1, \qquad i = 1, 2, \ldots, L,$$

and the v_i are specified constants. Any linear inequality can be put into the form of (6.6-3), and normalization is performed by simply dividing both sides of the inequality by a positive constant. These inequalities define a convex region R in a K-dimensional Euclidean space (E^K).† It is assumed that R is bounded; hence there must be at least $(K + 1)$ linear inequalities ($L \geq K + 1$). Under these assumptions, the problem is to find the minimum of a function f that is convex and has continuous second partial derivatives in a closed and bounded convex region R. If we define $\mathbf{n}_i \triangleq [n_{1i}, n_{2i}, \ldots, n_{Ki}]^T$, $i = 1, 2, \ldots, L$, the inequalities (6.6-3) can be written

$$\mathbf{n}_i^T \mathbf{y} - v_i \triangleq \lambda_i(\mathbf{y}) \geq 0, \qquad i = 1, 2, \ldots, L. \tag{6.6-3a}$$

The points that satisfy $\lambda_i(\mathbf{y}) = 0$ lie in a hyperplane (which we will denote by H_i) in the K-dimensional space. The boundary of R consists of all points that lie in at least one of the hyperplanes; that is, $\lambda_i(\mathbf{y}) = 0$ is satisfied for at least one i, and the interior of R consists of all points that satisfy $\lambda_i(\mathbf{y}) > 0$ for *all* $i = 1, 2, \ldots, L$. The unit vector $\mathbf{n}_i$ is orthogonal to H_i,‡ and if drawn so that it originates at a point in H_i, then $\mathbf{n}_i$ points toward the interior of R. For example, a three-dimensional space bounded by five planes is shown in Fig. 6-13. Notice that the intersection of two linearly independent planes is a straight line, and that the intersection of three linearly independent planes is a point.§ In a K-dimensional space, the intersection of two linearly independent hyperplanes defines a $(K - 2)$-dimensional subspace (or manifold) of E^K; the intersection of (K $-$ 1) linearly independent hyperplanes determines a line, and the intersection of K linearly independent hyperplanes determines a point.

† A region C is *convex* if the straight line joining any two points in C lies entirely within C. That is, if $\mathbf{y}^{(1)}, \mathbf{y}^{(2)} \in C$ then $\mathbf{y}^{(3)} = \theta \mathbf{y}^{(1)} + (1 - \theta)\mathbf{y}^{(2)} \in C$ for all $0 \leq \theta \leq 1$. Since R is defined by the linear inequalities (6.6-3), R is convex.

‡ Two vectors $\mathbf{n}$ and $\mathbf{w}$ are said to be orthogonal if the inner product $\mathbf{n}^T\mathbf{w} = 0$. (In two- or three-dimensional spaces the term *perpendicular* is often used.) A vector $\mathbf{n}$ is *orthogonal to a hyperplane* H_i if $\mathbf{n}^T\mathbf{w} = 0$ for all $\mathbf{w}$ in H_i.

§ q hyperplanes $H_1, H_2, \ldots, H_q$ are *linearly independent* if the corresponding unit normals $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_q$ are linearly independent; that is, if the linear combination $\alpha_1\mathbf{n}_1 + \alpha_2\mathbf{n}_2 + \cdots + \alpha_q\mathbf{n}_q$ is zero only if $\alpha_i = 0$, $i = 1, 2, \ldots, q$.

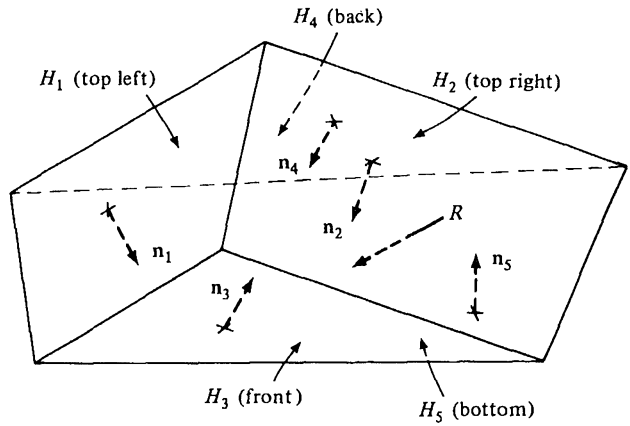


Figure 6-13 A closed convex region bounded by five planes

If we define

$$\mathbf{N}_L \triangleq [\mathbf{n}_1 \quad \mathbf{n}_2 \quad \cdots \quad \mathbf{n}_L],$$

a $K \times L$ matrix, and

$$\mathbf{v}_L \triangleq [v_1 \quad v_2 \quad \cdots \quad v_L]^T,$$

an L vector, then a set of linear constraints can be written collectively as

$$\mathbf{N}_L^T\mathbf{y} - \mathbf{v}_L = \boldsymbol{\lambda}(\mathbf{y}) \geq \mathbf{0},\dagger \tag{6.6-3b}$$

where $\boldsymbol{\lambda}(\mathbf{y}) \triangleq [\lambda_1(\mathbf{y}), \lambda_2(\mathbf{y}), \ldots, \lambda_L(\mathbf{y})]^T$. For example, the normalized version of the inequalities (6.6-1) is

$$y_1 \geq 0 \tag{6.6-4a}$$

$$y_2 \geq 0 \tag{6.6-4b}$$

$$\left[\frac{2}{\sqrt{29}}\right]y_1 - \left[\frac{5}{\sqrt{29}}\right]y_2 + \frac{10}{\sqrt{29}} \geq 0 \tag{6.6-4c}$$

$$-\left[\frac{4}{\sqrt{65}}\right]y_1 - \left[\frac{7}{\sqrt{65}}\right]y_2 + \frac{22.5}{\sqrt{65}} \geq 0 \tag{6.6-4d}$$

$$-\left[\frac{9}{\sqrt{85}}\right]y_1 - \left[\frac{2}{\sqrt{85}}\right]y_2 + \frac{26.5}{\sqrt{85}} \geq 0, \tag{6.6-4e}$$

and these can be written in the matrix form given by (6.6-3b) as

† This notation means that each component of the vector $\mathbf{N}_L^T\mathbf{y} - \mathbf{v}_L$ is greater than, or equal to, zero.

$$\begin{bmatrix} 1 & 0 & \frac{2}{\sqrt{29}} & -\frac{4}{\sqrt{65}} & -\frac{9}{\sqrt{85}} \\ 0 & 1 & -\frac{5}{\sqrt{29}} & -\frac{7}{\sqrt{65}} & -\frac{2}{\sqrt{85}} \end{bmatrix}^T \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ -\frac{10}{\sqrt{29}} \\ -\frac{22.5}{\sqrt{65}} \\ -\frac{26.5}{\sqrt{85}} \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6.6-5}$$

Suppose that $\mathbf{y}$ is a point that lies in the intersection of q linearly independent hyperplanes. These hyperplanes, which for convenience we shall assume are $H_1, H_2, \ldots, H_q$, are defined by the q linearly independent unit normals $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_q$, and the components $v_1, v_2, \ldots, v_q$ of the vector $\mathbf{v}_L$ in (6.6-3b). (Notice that $q \leq K$ because there can be at most K linearly independent vectors in the space E^K.) The q equations

$$[\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_q]^T\mathbf{y} - [v_1, v_2, \ldots, v_q]^T \triangleq \mathbf{N}_q^T\mathbf{y} - \mathbf{v}_q = \mathbf{0} \tag{6.6-6}$$

determine the points that lie in the intersection of $H_1, H_2, \ldots, H_q$; let us denote this intersection by Q'. Next, we consider the points $\mathbf{w}$ that satisfy

$$\mathbf{N}_q^T\mathbf{w} = \mathbf{0}; \tag{6.6-7}$$

$\mathbf{N}_q$ is the same matrix as defined in Eq. (6.6-6), so these points lie in the intersection of q linearly independent hyperplanes, each of which contains the origin. The intersection defined by (6.6-7), which is a $(K - q)$-dimensional subspace of E^K, will be denoted by Q. Notice that the intersections Q and Q' differ only by the vector $\mathbf{v}_q$. The linearly independent unit normals $\mathbf{n}_1$, $\mathbf{n}_2, \ldots, \mathbf{n}_q$, which make up $\mathbf{N}_q$, span a q-dimensional subspace of E^K, which we shall denote by $\bar{Q}$.† It can be shown that the subspaces Q and $\bar{Q}$ are orthogonal; that is, if $\mathbf{w}$ is any vector in Q and $\mathbf{s}$ is any vector in $\bar{Q}$, then $\mathbf{w}$ and $\mathbf{s}$ are orthogonal; $\mathbf{s}^T\mathbf{w} = 0$. In addition, the union of Q and $\bar{Q}$ is the entire K-dimensional space E^K.

Using $\mathbf{N}_q$ given in (6.6-6), let us define the $K \times K$ symmetric matrices

$$\tilde{\mathbf{P}}_q \triangleq \mathbf{N}_q[\mathbf{N}_q^T\mathbf{N}_q]^{-1}\mathbf{N}_q^T \tag{6.6-8}$$

$$\mathbf{P}_q \triangleq \mathbf{I} - \mathbf{N}_q[\mathbf{N}_q^T\mathbf{N}_q]^{-1}\mathbf{N}_q^T = \mathbf{I} - \tilde{\mathbf{P}}_q, \tag{6.6-9}$$

† A set of vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ *span* a q-dimensional subspace if every vector in the subspace can be expressed as a linear combination of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$; hence, every vector in $\bar{Q}$ can be written as a linear combination of $\mathbf{n}_1, \ldots, \mathbf{n}_q$.

where $\mathbf{I}$ is the $K \times K$ identity matrix. Since the unit vectors $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_q$ are linearly independent, the matrix $[\mathbf{N}_q^T\mathbf{N}_q]$ is nonsingular, and the inverse $[\mathbf{N}_q^T\mathbf{N}_q]^{-1}$ exists. It can be shown [R-4] that $\tilde{\mathbf{P}}_q$ is a projection matrix that takes any vector in E^K into $\tilde{Q}$, and $\mathbf{P}_q$ is a projection matrix that takes any vector in E^K into Q.

Let us now illustrate the utility of the projection matrix $\mathbf{P}_q$. Assume that $\mathbf{y}$ lies on the part of the boundary of R determined by the intersection Q' of the q linearly independent hyperplanes $H_1, H_2, \ldots, H_q$; that is, $\mathbf{y}$ satisfies (6.6-6), and let $-\partial f/\partial \mathbf{y}$ be the negative gradient of the function f at the point $\mathbf{y}$. We assert that the vector $\mathbf{s}$ defined by

$$\mathbf{s} \triangleq \mathbf{y} + \mathbf{P}_q\left[-\frac{\partial f}{\partial \mathbf{y}}\right] \tag{6.6-10}$$

satisfies

$$\mathbf{N}_q^T\mathbf{s} - \mathbf{v}_q = \mathbf{0}; \tag{6.6-11}$$

hence $\mathbf{s}$ also lies on the part of the boundary of R determined by the intersection Q' of $H_1, H_2, \ldots, H_q$. To show this, we substitute the expression for $\mathbf{s}$ in (6.6-10) and the definition of $\mathbf{P}_q$ given by (6.6-9) into (6.6-11) with the result

$$\mathbf{N}_q^T\left\{\mathbf{y} + [\mathbf{I} - \mathbf{N}_q[\mathbf{N}_q^T\mathbf{N}_q]^{-1}\mathbf{N}_q^T]\left[-\frac{\partial f}{\partial \mathbf{y}}\right]\right\} - \mathbf{v}_q = \mathbf{0} \tag{6.6-12}$$

or

$$\mathbf{N}_q^T\mathbf{y} - \mathbf{v}_q + [\mathbf{N}_q^T - \mathbf{N}_q^T]\left[-\frac{\partial f}{\partial \mathbf{y}}\right] = \mathbf{0}. \tag{6.6-13}$$

The coefficient of $-\partial f/\partial \mathbf{y}$ is the $q \times K$ zero matrix, and the first two terms on the left add to zero because $\mathbf{y}$ satisfies (6.6-6). Equation (6.6-11) is important because it indicates the procedure for changing $\mathbf{y}$ along the boundary of R in the direction of the projected gradient.

Calculation Requirements

Let us now discuss the calculations that are required by the gradient projection algorithm.

The Gradient. It is assumed that the expression for the function to be minimized is known. The components of the gradient vector are found by taking the partial derivatives of f with respect to $y_1, y_2, \ldots, y_K$. For example, if

$$f(\mathbf{y}) = y_1^2 - 80y_1 + 1600 + y_2^2 - 100y_2, \tag{6.6-14}$$

then

$$\frac{\partial f^{(i)}}{\partial \mathbf{y}} \triangleq \frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^{(i)}) = [2y_1^{(i)} - 80, 2y_2^{(i)} - 100]^T \tag{6.6-15}$$

is the gradient at the point $\mathbf{y}^{(i)}$. $-\partial f^{(i)}/\partial \mathbf{y}$ is obtained by changing the sign of each component of $\partial f^{(i)}/\partial \mathbf{y}$.

The Projection Matrix. From Eq. (6.6-9) it is seen that to determine the projection matrix $\mathbf{P}_q$ at some point $\mathbf{y}^{(i)}$, it is first necessary to find the matrix $\mathbf{N}_q$. This is done by forming the L vector $\boldsymbol{\lambda}(\mathbf{y}^{(i)}) = \mathbf{N}_L^T\mathbf{y}^{(i)} - \mathbf{v}_L$ of (6.6-3b) and checking the sign of each component of $\boldsymbol{\lambda}$. Since $\mathbf{y}^{(i)}$ is assumed to be an admissible point, each component of $\boldsymbol{\lambda}$ must be nonnegative. If λ_j (the jth component of $\boldsymbol{\lambda}$) is zero, the unit vector $\mathbf{n}_j$ is to be included in $\mathbf{N}_q$; if $\lambda_j > 0$, $\mathbf{n}_j$ is not included in $\mathbf{N}_q$. Once $\mathbf{N}_q$ is known, the matrices $\mathbf{N}_q^T\mathbf{N}_q$ and $[\mathbf{N}_q^T\mathbf{N}_q]^{-1}$ can be found and the projection matrix formed by using Eq. (6.6-9); that is,

$$\mathbf{P}_q = \mathbf{I} - \mathbf{N}_q[\mathbf{N}_q^T\mathbf{N}_q]^{-1}\mathbf{N}_q^T. \tag{6.6-9}$$

Subsequently, we shall see that only *one* vector $\mathbf{n}_q$ is added to, or dropped from, $\mathbf{N}_q$ at each stage of the iterative process; in addition to simplifying the determination of $\mathbf{N}_q$, this allows the matrix $[\mathbf{N}_q^T\mathbf{N}_q]^{-1}$ to be found by using recurrence relations† that do not require matrix inversion.

In Example 6.6-1, the projection matrix at the point $\mathbf{y}^{(0)}$ is the identity matrix, since $\mathbf{y}^{(0)}$ lies in the interior of the admissible region. At $\mathbf{y}^{(1)}$, on the other hand, if we form the vector $\boldsymbol{\lambda}$, we find that $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_4 > 0$, $\lambda_5 > 0$, and $\lambda_3 = 0$, indicating that

$$\mathbf{N}_q = \mathbf{n}_3 = \left[\frac{2}{\sqrt{29}}, \quad -\frac{5}{\sqrt{29}}\right]^T$$

Following the procedure outlined above, we obtain

$$\mathbf{N}_q^T\mathbf{N}_q = \begin{bmatrix} \frac{2}{\sqrt{29}} & -\frac{5}{\sqrt{29}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{29}} \\ -\frac{5}{\sqrt{29}} \end{bmatrix} = 1, \tag{6.6-16}$$

$$[\mathbf{N}_q^T\mathbf{N}_q]^{-1} = 1, \tag{6.6-17}$$

$$\mathbf{P}_q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \frac{2}{\sqrt{29}} \\ -\frac{5}{\sqrt{29}} \end{bmatrix} [1] \begin{bmatrix} \frac{2}{\sqrt{29}} & -\frac{5}{\sqrt{29}} \end{bmatrix} = \begin{bmatrix} 1 - \frac{4}{29} & \frac{10}{29} \\ \frac{10}{29} & 1 - \frac{25}{29} \end{bmatrix}. \tag{6.6-18}$$

† See [R-4], pp. 189ff.

The Maximum Allowable Step Size. In changing $\mathbf{y}$ in the direction of the projected gradient, it is necessary to know the maximum step size that can be used without causing any of the constraints to be violated. Assume that the projected gradient has been found by performing the matrix multiplication $\mathbf{P}_q[-\partial f^{(i)}/\partial \mathbf{y}]$. Let $\mathbf{z}^{(i)}$ represent the unit vector in the direction of the projected gradient; that is,

$$\mathbf{z}^{(i)} = \frac{\mathbf{P}_q\left[-\dfrac{\partial f^{(i)}}{\partial \mathbf{y}}\right]}{\left\|\mathbf{P}_q\left[-\dfrac{\partial f^{(i)}}{\partial \mathbf{y}}\right]\right\|}. \tag{6.6-19}$$

and define

$$\mathbf{y}' \triangleq \mathbf{y}^{(i)} + \tau \mathbf{z}^{(i)}, \tag{6.6-20}$$

where τ is a scalar that represents the step size. We wish to find the maximum value of τ for which all of the constraints are satisfied. $\mathbf{y}^{(i)}$ lies in the intersection Q', and (6.6-20) defines a line that also lies in Q' for all values of τ. Let

$$\mathbf{y}_j'^{(i+1)} \triangleq \mathbf{y}^{(i)} + \tau_j \mathbf{z}^{(i)} \tag{6.6-21}$$

be the point where this line intersects the hyperplane H_j; then

$$\mathbf{n}_j^T \mathbf{y}_j'^{(i+1)} - v_j = 0. \tag{6.6-22}$$

Substituting $\mathbf{y}_j'^{(i+1)}$ from (6.6-21) gives

$$\mathbf{n}_j^T \mathbf{y}^{(i)} + \tau_j \mathbf{n}_j^T \mathbf{z}^{(i)} - v_j = 0, \tag{6.6-23}$$

which when solved for τ_j yields

$$\tau_j = \frac{v_j - \mathbf{n}_j^T \mathbf{y}^{(i)}}{\mathbf{n}_j^T \mathbf{z}^{(i)}}. \tag{6.6-24}$$

τ_j is calculated for all hyperplanes not already in Q'; the *minimum positive value* of these τ_j's, denoted by τ_m, determines the maximum step that can be taken along the line (6.6-20) without violating any constraints. Thus,

$$\mathbf{y}'^{(i+1)} = \mathbf{y}^{(i)} + \tau_m \mathbf{z}^{(i)} \tag{6.6-25}$$

is the point most distant from $\mathbf{y}^{(i)}$ along the gradient projection for which no constraints are violated.

To illustrate this procedure, consider the point $\mathbf{y}^{(1)}$ shown in Fig. 6-11. The unit projected gradient vector $\mathbf{z}^{(1)}$ is in the same direction as the vector $\mathbf{P}[-\partial f^{(1)}/\partial \mathbf{y}]$, and is given by

$$\mathbf{z}^{(1)} = [0.930 \quad 0.372]^T.$$

Suppose that $\mathbf{y}^{(1)} = [1.0 \quad 2.4]^T$. Using (6.6-24) to solve for τ_j, $j = 1, 2, 4, 5$, we obtain

$$\tau_1 = -1.075, \qquad \tau_2 = -6.452, \qquad \tau_4 = 0.269, \quad \text{and} \quad \tau_5 = 1.393.$$

By inspection, τ_m is equal to $\tau_4 = 0.269$. In Fig. 6-11 notice that H_4 is the hyperplane closest to the point $\mathbf{y}^{(1)}$ when moving in the positive $\mathbf{z}^{(1)}$ direction. The negative values for τ_1 and τ_2 indicate that to reach H_1 and H_2, $\mathbf{y}$ would have to be changed in the negative $\mathbf{z}^{(1)}$ direction—this conclusion is also obtained by inspection of Fig. 6-11.

The point $\mathbf{y}'^{(2)}$ is found by substituting $\mathbf{z}^{(1)}$, $\mathbf{y}^{(1)}$, and the calculated value of τ_m into Eq. (6.6-25).

Interpolation. If the maximum step is taken to the point $\mathbf{y}'^{(i+1)}$, the next stage of the iterative procedure may indicate a step back toward the point $\mathbf{y}^{(i)}$; this would occur, for example, at $\mathbf{y}'^{(4)}$ in Fig. 6-11. To determine whether or not the maximum step size should be used, we form the inner product

$$\mathbf{z}^{(i)T}\left[-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}'^{(i+1)})\right], \tag{6.6-26}$$

where $\mathbf{z}^{(i)}$ is the unit projected gradient at the point $\mathbf{y}^{(i)}$. If this inner product is greater than or equal to zero, as would be the case, for example, in Fig. 6-14(a) and (b), then the maximum step is taken; that is,

$$\mathbf{y}^{(i+1)} = \mathbf{y}'^{(i+1)}. \tag{6.6-27}$$

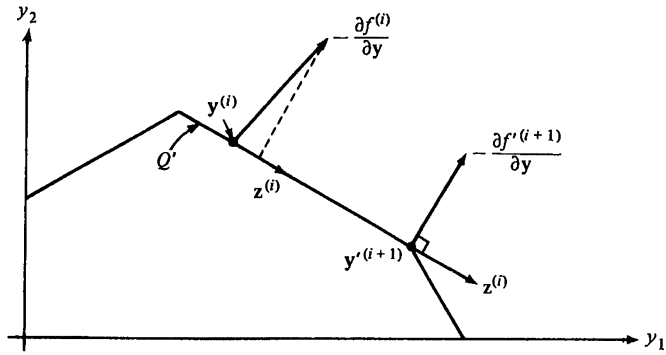
After the maximum step has been taken, the point $\mathbf{y}'^{(i+1)}$ lies in the intersection of Q' and the hyperplane H_m (which corresponds to τ_m); hence H_m is added to Q', and the new projection matrix $\mathbf{P}_{q+1}$ is calculated. On the other hand, if the inner product is negative, as, for example, in Fig. 6-14(c), the maximum step is not taken. Instead, interpolation is used to find the point

$$\mathbf{y}^{(i+1)} = \mathbf{y}^{(i)} + \theta\tau_m\mathbf{z}^{(i)} \qquad (0 < \theta < 1), \tag{6.6-28}$$

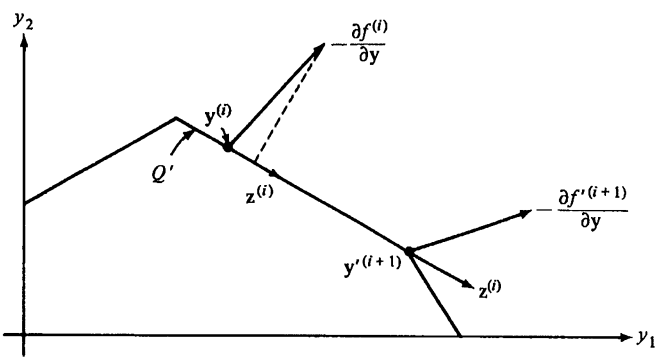
where

$$\mathbf{z}^{(i)T}\left[-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^{(i+1)})\right] = 0, \tag{6.6-29}$$

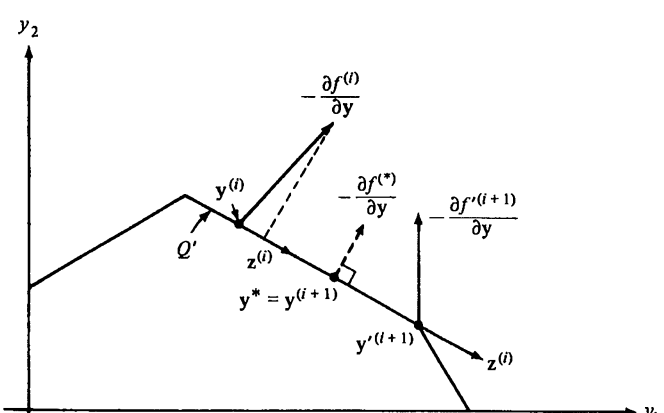
that is, the point where the gradient is orthogonal to Q'. A straightforward method for finding the appropriate value of θ is to use repeated linear interpolation as illustrated in Fig. 6-15. $\theta = 0$ corresponds to the point $\mathbf{y}^{(i)}$, and $\theta = 1$ corresponds to the point $\mathbf{y}'^{(i+1)}$. θ_1, the abscissa where the straight line from A to B has an ordinate of zero, is determined from the relationship



(a)



(b)



(c)

Figure 6-14 (a) $\mathbf{z}^{(i)T}[-\partial f(\mathbf{y}'^{(i+1)})/\partial \mathbf{y}] = 0$. No interpolation required. (b) $\mathbf{z}^{(i)T}[-\partial f(\mathbf{y}'^{(i+1)})/\partial \mathbf{y}] > 0$. No interpolation required. (c) $\mathbf{z}^{(i)T}[-\partial f(\mathbf{y}'^{(i+1)})/\partial \mathbf{y}] < 0$. Interpolation required

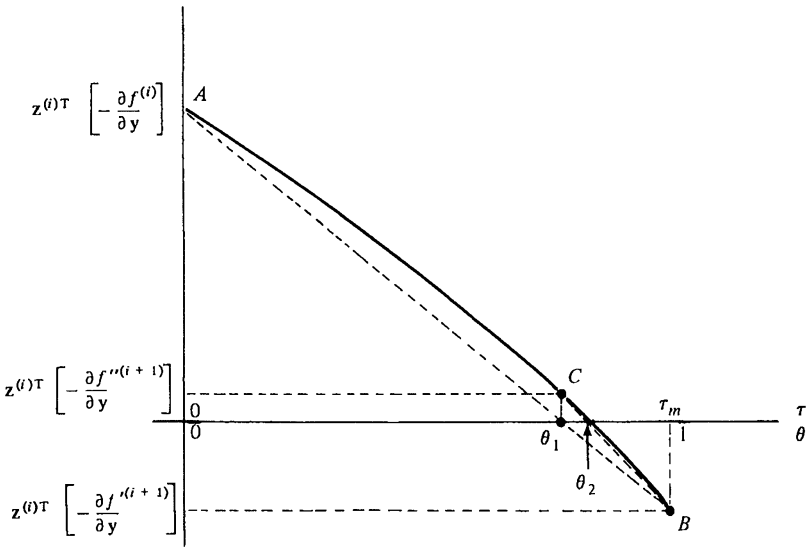


Figure 6-15 Repeated linear interpolation

$$\theta_1 = \frac{z^{(i)T}\left[-\dfrac{\partial f^{(i)}}{\partial y}\right]}{z^{(i)T}\left[-\dfrac{\partial f^{(i)}}{\partial y}\right] - z^{(i)T}\left[-\dfrac{\partial f''^{(i+1)}}{\partial y}\right]}. \tag{6.6-30}$$

Next, we evaluate the gradient at the point

$$y''^{(i+1)} = y^{(i)} + \theta_1 \tau_m z^{(i)} \tag{6.6-31}$$

and form the inner product

$$z^{(i)T}\left[-\frac{\partial f''^{(i+1)}}{\partial y}\right]. \tag{6.6-32}$$

If this inner product is positive, as in Fig. 6-15, we use points C and B to interpolate again (if the inner product had been negative, point C would have a negative ordinate and points A and C would be used for the next interpolation). This procedure is repeated until a point $y^{(i+1)}$ is found where the magnitude of the inner product is less than a preassigned small positive number ϵ_2; that is,

$$\left| z^{(i)T}\left[-\frac{\partial f^{(i+1)}}{\partial y}\right] \right| < \epsilon_2. \tag{6.6-33}$$

Necessary and Sufficient Conditions for a Constrained Global Minimum

Let us now state the theorem that provides the basis for the gradient projection algorithm.

THEOREM 6.6-1

Assume that f is a convex function with continuous second partial derivatives in a closed and bounded convex region R of E^K. Let $\mathbf{y}^*$ be a boundary point of R which lies on exactly q, $1 \leq q \leq K$, hyperplanes that are assumed to be linearly independent. Q' denotes the intersection of these hyperplanes. The point $\mathbf{y}^*$ is a constrained global minimum of f if, and only if,

$$\mathbf{P}_q\left[-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^*)\right] = \mathbf{0} \tag{6.6-34}$$

and

$$[\mathbf{N}_q^T\mathbf{N}_q]^{-1}\mathbf{N}_q^T\left[-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^*)\right] \leq \mathbf{0}. \tag{6.6-35}$$

A proof of this theorem is given in [R-4] and will not be repeated here. A few comments are in order, however:

1. The proof that (6.6-34) and (6.6-35) are necessary for $\mathbf{y}^*$ to be a constrained global minimum is a constructive procedure for obtaining a point with a smaller value of the objective function if *both* conditions are not satisfied at $\mathbf{y}^*$. Thus, the gradient projection algorithm follows directly from the proof.
2. If $\mathbf{y}^*$ is an *interior point* of R ($\mathbf{y}^*$ lies inside rather than on the boundary of R), then the projection matrix $\mathbf{P}_q$ is simply the $K \times K$ identity matrix, and Eq. (6.6-34) reduces to the familiar necessary and sufficient condition that

$$\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^*) = \mathbf{0}. \tag{6.6-36}$$

   The sufficiency follows from the assumption that f is a convex function.
3. It should be emphasized that this theorem gives necessary and sufficient conditions for $f(\mathbf{y}^*)$ to be a constrained *global* (or *absolute*) minimum. That is, if $\mathbf{y}^*$ satisfies (6.6-34) and (6.6-35), then $f(\mathbf{y}^*) \leq f(\mathbf{y})$ for *all* admissible $\mathbf{y}$.

Geometric Interpretation of the Necessary and Sufficient Conditions

Let us now discuss the geometric interpretation of the conditions given by Theorem 6.6-1. The requirement that $\mathbf{P}_q[-\partial f(\mathbf{y}^*)/\partial \mathbf{y}] = \mathbf{0}$ implies that either:

1. $-\partial f(\mathbf{y}^*)/\partial \mathbf{y} = \mathbf{0}$, which means that $\mathbf{y}^*$ is an interior point of R or that the *unconstrained* minimum coincides with the boundary as shown, for example, in Fig. 6-16; or
2. $-\partial f(\mathbf{y}^*)/\partial \mathbf{y} \neq \mathbf{0}$, in which case the gradient is orthogonal to the intersection of the q hyperplanes, as shown at the point $\mathbf{y}^{(4)}$ in Fig. 6-11.

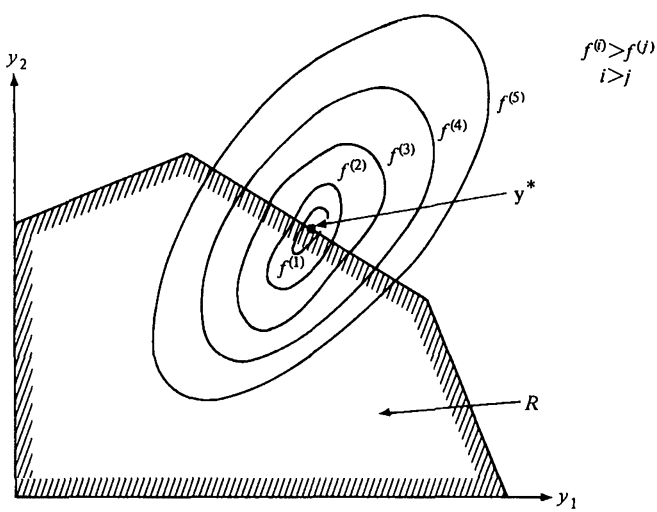


Figure 6-16 Coincidence of an unconstrained minimum with the boundary of R

If the second alternative applies, we must next ascertain whether the gradient is directed toward the interior of R, or outward; this is the role of Eq. (6.6-35).

Suppose that (6.6-34) is satisfied; then $-\partial f(\mathbf{y}^*)/\partial \mathbf{y}$ is in the space $\bar{Q}$ and can be written as a linear combination of $\mathbf{n}_1, \ldots, \mathbf{n}_q$, that is,

$$-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^*) = \sum_{i=1}^{q} r_i \mathbf{n}_i = \mathbf{N}_q \mathbf{r}. \tag{6.6-37}$$

Premultiplying both sides by $[\mathbf{N}_q^T \mathbf{N}_q]^{-1} \mathbf{N}_q^T$, we obtain

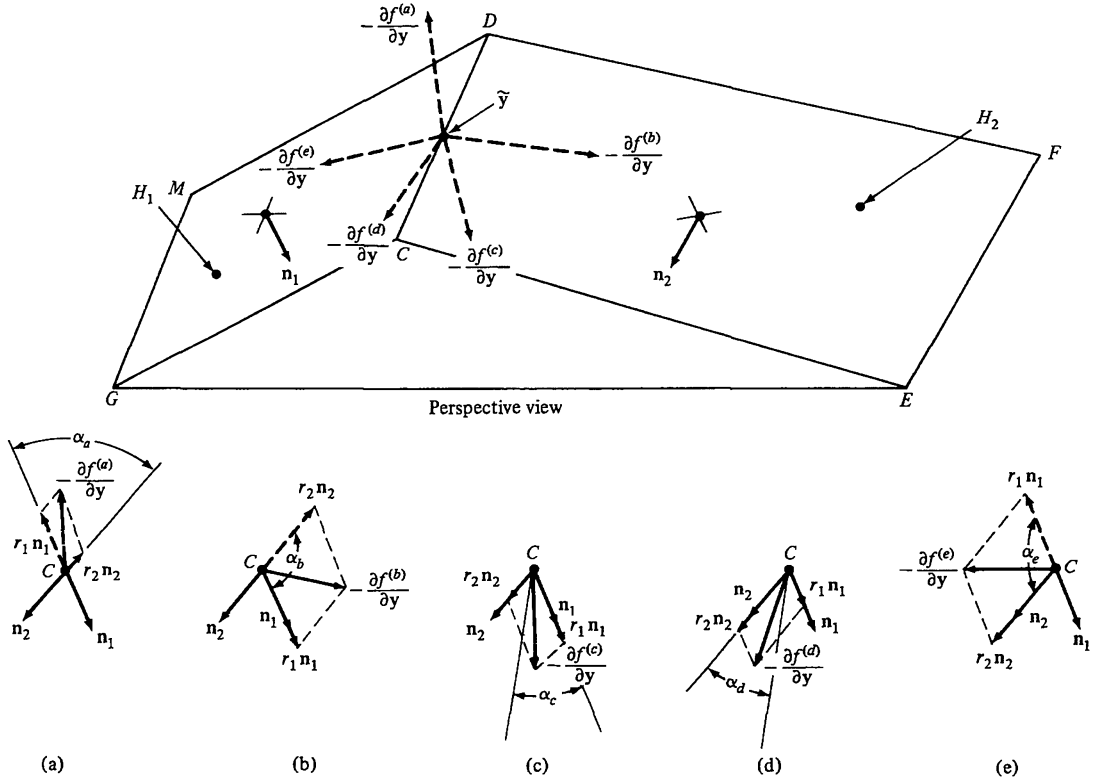


Figure 6-17 Several alternatives when $\mathbf{P}_q[-\partial f/\partial \mathbf{y}] = \mathbf{0}$. (a) Minimum attained, $r_1, r_2 < 0$. (b) Drop H_1, $r_1 > 0$, $r_2 < 0$. (c) Drop H_1, $r_1 > 0$, $r_2 > 0$, $r_1 > r_2$. (d) Drop H_2, $r_1 > 0$, $r_2 > 0$, $r_2 > r_1$. (e) Drop H_2, $r_1 < 0$, $r_2 > 0$

$$[\mathbf{N}_q^T\mathbf{N}_q]^{-1}\mathbf{N}_q^T\left[-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}^*)\right] = \mathbf{r}; \tag{6.6-38}$$

hence, (6.6-35) implies that

$$\mathbf{r} \leq \mathbf{0}. \tag{6.6-39}$$

Assume that the projected gradient is zero at some point $\tilde{\mathbf{y}}$. To determine if $\tilde{\mathbf{y}}$ is the minimum, we use Eq. (6.6-38) to calculate the vector $\mathbf{r}$. If each component of $\mathbf{r}$ is nonpositive, then $\tilde{\mathbf{y}} = \mathbf{y}^*$ is the constrained global minimum. If, however, one or more components of $\mathbf{r}$ is positive, this indicates that the objective function can be decreased by dropping the hyperplane corresponding to the largest positive component of $\mathbf{r}$ and proceeding with the iterative algorithm.

Figure 6-17 provides a geometric interpretation of the use of the condition (6.6-39). The subspace Q' consists of the line CD, which is the intersection of the planes H_1 and H_2, and the matrix $\mathbf{P}_q$ projects the gradient onto CD. In each of the cases shown, the gradient vector at $\tilde{\mathbf{y}}$ is orthogonal to CD; hence, the projected gradient is zero. If $-\partial f/\partial \mathbf{y}$ lies in the sector defined by α_a in Fig. 6-17(a), then in the expression

$$-\frac{\partial f}{\partial \mathbf{y}}(\tilde{\mathbf{y}}) = r_1\mathbf{n}_1 + r_2\mathbf{n}_2 = \mathbf{N}_q\mathbf{r}, \tag{6.6-40}$$

both r_1 and r_2 will be negative as shown, indicating that $\tilde{\mathbf{y}} = \mathbf{y}^*$ is the constrained global minimum. If the gradient is oriented as in Fig. 6-17(b), the perspective view indicates that the objective function can be decreased by moving toward EF in the plane H_2. For this case, r_2 will be negative and r_1 positive—which means that H_1 should be dropped from Q'. If $-\partial f/\partial \mathbf{y}$ is as shown in Fig. 6-17(c), then $r_1 > r_2 > 0$. Although both H_1 and H_2 could be dropped from Q', it is more convenient to drop at most one hyperplane in each iteration; therefore, H_1, which corresponds to the largest positive component of $\mathbf{r}$, would be dropped. The situation illustrated by Fig. 6-17(d) is similar to (c), except that $r_2 > r_1 > 0$, indicating that H_2 should be dropped. Finally, if $-\partial f/\partial \mathbf{y}$ is in the sector defined by α_e in Fig. 6-17(e), $r_1 < 0$ and $r_2 > 0$; hence, the plane H_2 would be removed from Q'.

Because of numerical inaccuracies, if $\|\mathbf{P}_q[-\partial f/\partial \mathbf{y}]\| \leq \epsilon_1$, where ϵ_1 is a small positive constant, we shall agree that $\mathbf{P}_q[-\partial f/\partial \mathbf{y}] \approx \mathbf{0}$, which indicates that the gradient vector is orthogonal to Q' and that the vector $\mathbf{r}$ should be computed to determine whether the global minimum has been found, or, if not, which hyperplane should be dropped. If the gradient projection is not orthogonal to Q', that is, $\|\mathbf{P}_q[-\partial f/\partial \mathbf{y}]\| > \epsilon_1$, it still may be desirable to drop a hyperplane from Q'. To see how this situation may occur, refer to Fig. 6-18. At the point $\tilde{\mathbf{y}}$, the matrix $\mathbf{P}_q$ projects $-\partial f/\partial \mathbf{y}$ onto Q' (the line

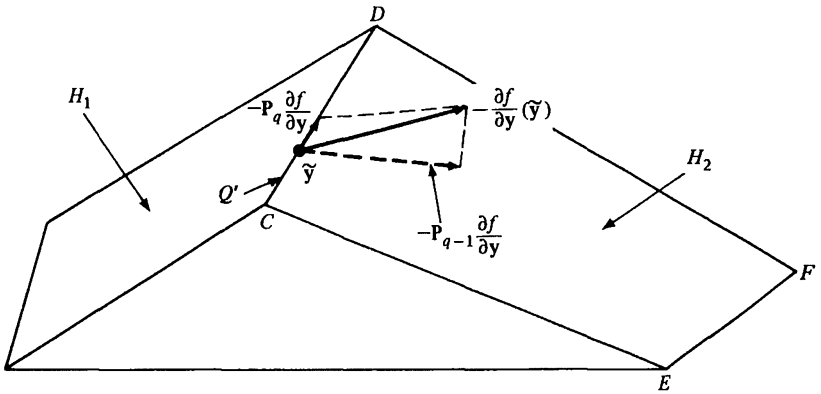


Figure 6-18 Dropping a hyperplane when $\|\mathbf{P}_q[-\partial f/\partial \mathbf{y}]\| > \epsilon_1$

CD) as shown; this projection is small, but nonzero. If H_1 were dropped from Q', however, the new projection matrix, $\mathbf{P}_{q-1}$, would project $-\partial f/\partial \mathbf{y}$ onto the plane H_2. The relative lengths of the two projections indicate that more is to be gained by moving toward EF on H_2 than by moving along CD; therefore, H_1 should be dropped from Q'. To detect whether a hyperplane should be dropped although $\|\mathbf{P}_q[-\partial f/\partial \mathbf{y}]\| > \epsilon_1$, Rosen suggests a test (which does not require the actual determination of each possible $\mathbf{P}_{q-1}$ formed by dropping exactly one hyperplane from Q') consisting of the following steps:

1. Let $\alpha_i \triangleq$ the sum of the absolute values of the elements of the ith row of the matrix $[\mathbf{N}_q^T\mathbf{N}_q]^{-1}$, calculate α_i, $i = 1, 2, \ldots, q$, and determine

$$\beta = \max_i \{\alpha_i\}. \tag{6.6-41}$$

2. Compute the vector $\mathbf{r}$ given by Eq. (6.6-38), and determine r_q, the maximum positive component of $\mathbf{r}$.
3. If $r_q > \beta$, drop the hyperplane H_q from Q'. If $r_q \leq \beta$, no hyperplane is dropped from Q'.

For a discussion of the theoretical basis for this procedure, the reader should refer to [R-4].

A Summary of the Gradient Projection Iterative Procedure

Let us now formalize the iterative procedure that was used in Example 6.6-1. It will be assumed that the initial point $\mathbf{y}^{(0)}$ is admissible and lies in the intersection Q' of q linearly independent hyperplanes. To determine the constrained global minimum:

1. Calculate the projection matrix $\mathbf{P}_q$, the gradient vector at the point $\mathbf{y}^{(i)}$, $-\partial f(\mathbf{y}^{(i)})/\partial \mathbf{y} \triangleq -\partial f^{(i)}/\partial \mathbf{y}$, the vector $\mathbf{r}$ given by Eq. (6.6-38), and the gradient projection $\mathbf{P}_q[-\partial f^{(i)}/\partial \mathbf{y}]$. If $\|\mathbf{P}_q[-\partial f^{(i)}/\partial \mathbf{y}]\| \leq \epsilon_1$, and $\mathbf{r} \leq \mathbf{0}$, then $\mathbf{y}^{(i)}$ is the constrained global minimum and the procedure is terminated; otherwise, go to step 2.
2. Determine whether or not a hyperplane should be dropped from Q'. If $\|\mathbf{P}_q[-\partial f^{(i)}/\partial \mathbf{y}]\| \leq \epsilon_1$, drop the hyperplane H_q, which corresponds to $r_q > 0$, form the projection matrix $\mathbf{P}_{q-1}$, and go to step 3.† The other alternative is that the norm of the gradient projection is greater than ϵ_1. In this case, calculate β given by (6.6-41). If $r_q > \beta$, drop the hyperplane H_q from Q'; if $r_q \leq \beta$, Q' remains unchanged.
3. Compute the normalized gradient projection $\mathbf{z}^{(i)}$ given by Eq. (6.6-19), and the maximum allowable step size τ_m, where τ_m is the *minimum positive value* of the τ_j's found by evaluating

$$\tau_j = \frac{v_j - \mathbf{n}_j^T \mathbf{y}^{(i)}}{\mathbf{n}_j^T \mathbf{z}^{(i)}} \tag{6.6-24}$$

for j corresponding to all hyperplanes not in the intersection Q'. The tentative next point $\mathbf{y}'^{(i+1)}$ is found from

$$\mathbf{y}'^{(i+1)} = \mathbf{y}^{(i)} + \tau_m \mathbf{z}^{(i)}. \tag{6.6-25}$$

4. Calculate the gradient at the point $\mathbf{y}'^{(i+1)}$, if

$$\mathbf{z}^{(i)T}\left[-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}'^{(i+1)})\right] \geq 0, \tag{6.6-42}$$

set $\mathbf{y}^{(i+1)} = \mathbf{y}'^{(i+1)}$; since $\mathbf{y}^{(i+1)}$ lies in the intersection of Q' and H_m (the hyperplane which corresponds to the step size τ_m determined in step 3), add H_m to Q', and return to step 1.

On the other hand, if

$$\mathbf{z}^{(i)T}\left[-\frac{\partial f}{\partial \mathbf{y}}(\mathbf{y}'^{(i+1)})\right] < 0, \tag{6.6-43}$$

find $\mathbf{y}^{(i+1)}$ by repeated linear interpolation as illustrated in Fig. 6-15; the appropriate equations are (6.6-30) and (6.6-31). The intersection Q' remains unchanged, and the computational algorithm begins another iteration by returning to step 1.

A flow chart of this procedure is shown in Fig. 6-19.

† Notice that at least one component of $\mathbf{r}$ must be positive; otherwise, the iterative procedure would have terminated in step 1.

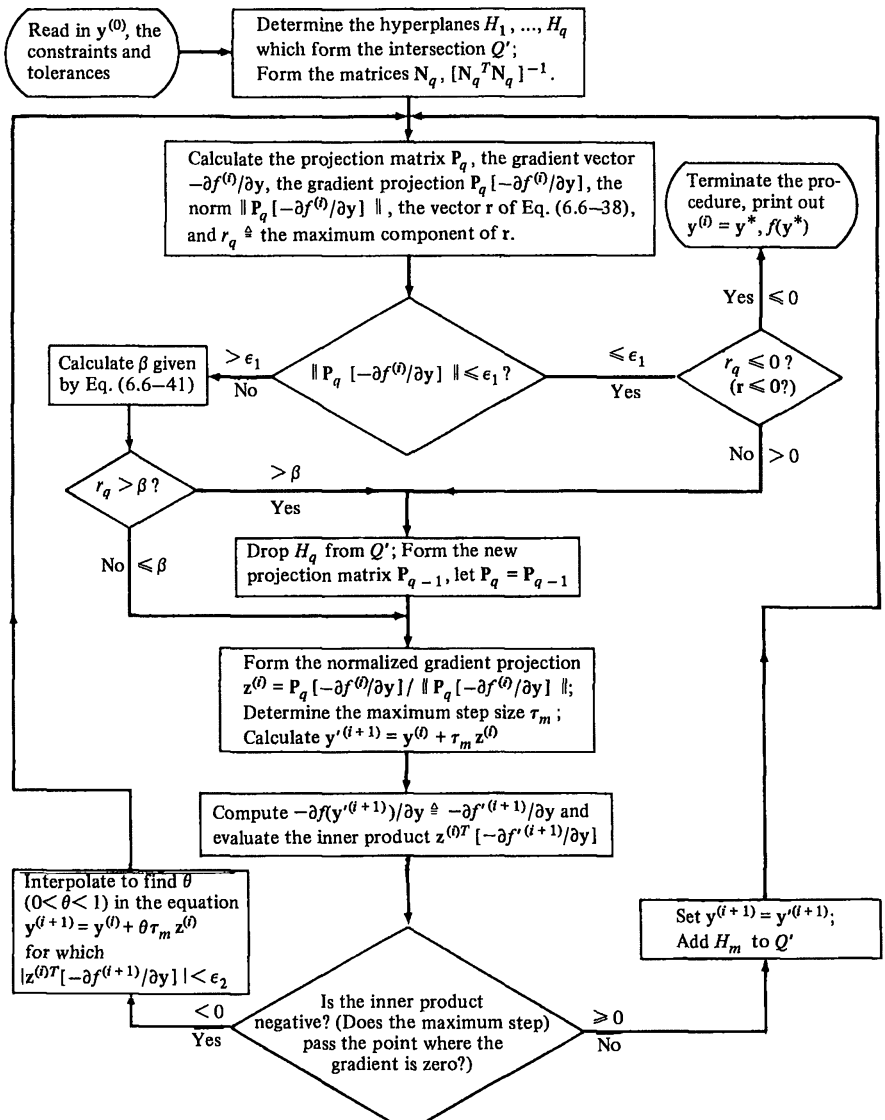


Figure 6-19 Flow chart of the gradient projection algorithm

To illustrate the procedure just summarized, let us return to Example 6.6-1. For simplicity, we shall assume that all calculations are exact so that consideration of tolerances is not required.

Referring to Fig. 6-11, we observe that the initial point $\mathbf{y}^{(0)}$ is the interior of the admissible region R; therefore, the projection matrix is the 2×2 identity matrix. Testing the norm of the gradient projection (which at $\mathbf{y}^{(0)}$ is simply the norm of the gradient), we find that this norm is nonzero. Since Q' is empty, calculation of the vector $\mathbf{r}$ is by-passed. The unit gradient projection $\mathbf{z}^{(0)} = [-\partial f^{(0)}/\partial \mathbf{y}]/\| -\partial f^{(0)}/\partial \mathbf{y} \|$ is calculated and used to find the maximum allowable step size by evaluating τ_j, $j = 1, 2, 3, 4, 5$ from Eq. (6.6-24). Using τ_3, which is the maximum allowable step size, yields $\mathbf{y}'^{(1)}$ on the line H_3 as the tentative next point. At $\mathbf{y}'^{(1)}$ it is found that the inner product $\mathbf{z}^{(0)T}[-\partial f'^{(1)}/\partial \mathbf{y}] > 0$; hence, no interpolation is required, $\mathbf{y}^{(1)} = \mathbf{y}'^{(1)}$, and H_3 is added to Q'; this completes the first iteration of the algorithm.

Returning to step 1 of the algorithm, we calculate the new projection matrix, the vector $\mathbf{r}$, and the gradient projection. The norm of the gradient projection is nonzero, and the test described previously indicates that no planes should be dropped. Computing the normalized gradient projection $\mathbf{z}^{(1)}$ and the maximum allowable step size τ_4, we find the tentative next point to be $\mathbf{y}'^{(2)}$. At $\mathbf{y}'^{(2)}$ it is found that $\mathbf{z}^{(1)T}[-\partial f'^{(2)}/\partial \mathbf{y}] > 0$; thus, no interpolation is required, $\mathbf{y}^{(2)}$ is set equal to $\mathbf{y}'^{(2)}$, and H_4 is added to Q'—which now consists of the point $\mathbf{y}^{(2)}$ (which is the intersection of H_3 and H_4). This completes the second iteration.

Calculating the new projection matrix, we find that the projection of $-\partial f^{(2)}/\partial \mathbf{y}$ onto the intersection of H_3 and H_4 is zero (the projection of any vector onto a point is zero); hence, Eq. (6.6-34) is satisfied. Using Eq. (6.6-38) to solve for $\mathbf{r}$ in

$$-\frac{\partial f^{(2)}}{\partial \mathbf{y}} = r_1 \mathbf{n}_3 + r_2 \mathbf{n}_4 = \mathbf{N}_q \mathbf{r} \tag{6.6-44}$$

gives $r_1 > 0$ and $r_2 < 0$. H_3 is dropped from Q', and the new projection matrix, which projects the gradient onto H_4, is computed. The normalized gradient projection $\mathbf{z}^{(2)}$ and the maximum step size τ_5 are determined and used to find the next tentative point $\mathbf{y}'^{(3)}$, located at the intersection of H_4 and H_5. The inner product $\mathbf{z}^{(2)T}[-\partial f'^{(3)}/\partial \mathbf{y}]$ is positive, so we set $\mathbf{y}^{(3)} = \mathbf{y}'^{(3)}$, and add H_5 to Q', thus completing the third iteration.

Returning to step 1 and calculating the new projection matrix, we find that the projection of $-\partial f^{(3)}/\partial \mathbf{y}$ onto the intersection of H_4 and H_5 is zero. Solving

$$-\frac{\partial f^{(3)}}{\partial \mathbf{y}} = r_1\mathbf{n}_4 + r_2\mathbf{n}_5$$
$$= \mathbf{N}_q\mathbf{r} \tag{6.6-45}$$

for $\mathbf{r}$ gives $r_1 > 0$, $r_2 < 0$, so H_4 is dropped from Q' and the projection matrix is recalculated. Projecting $-\partial f^{(3)}/\partial \mathbf{y}$ onto H_5 and moving the maximum allowable distance to the point $\mathbf{y}'^{(4)}$, we find that interpolation is required because $\mathbf{z}^{(3)T}[-\partial f'^{(4)}/\partial \mathbf{y}] < 0$. Performing linear interpolation repeatedly eventually obtains the point $\mathbf{y}^{(4)}$, where $\mathbf{z}^{(3)T}[-\partial f^{(4)}/\partial \mathbf{y}] = 0$. This completes the fourth iteration, the intersection Q' remaining unchanged.

At $\mathbf{y}^{(4)}$, $\mathbf{P}_q[-\partial f^{(4)}/\partial \mathbf{y}] = \mathbf{0}$; hence, Eq. (6.6-34) is satisfied. By inspection of Fig. 6-11 we observe that in the expression

$$-\frac{\partial f^{(4)}}{\partial \mathbf{y}} = r_1\mathbf{n}_5 \tag{6.6-46}$$

r_1 is negative; thus, *both* (6.6-34) and (6.6-35) are satisfied, and $\mathbf{y}^* = \mathbf{y}^{(4)}$ is the sought-after minimum.

Additional Features of the Gradient Projection Algorithm

Before establishing the connection between gradient projection and the solution of optimal control problems, let us first mention some additional features of Rosen's algorithm:†

1. Since at most one hyperplane is added or dropped at each stage in the iterative procedure, the matrix $[\mathbf{N}_q^T\mathbf{N}_q]^{-1}$, and hence $\mathbf{P}_q$ can be calculated from recurrence relations that do not require matrix inversion.
2. It may occur that a point calculated by the iterative procedure lies in the intersection of i hyperplanes, only $q < i$ of which are linearly independent; the gradient projection method contains provisions for dealing with such situations.
3. The algorithm provides a starting procedure for generating an admissible point (if one exists) from an arbitrary initial guess $\mathbf{y}^{(0)}$.
4. If f is a convex function in the admissible region of E^K and has continuous second partial derivatives with respect to each of the components of $\mathbf{y}$ in the admissible region R, then the gradient projection algorithm converges to a global minimum of f. If f is not convex in R, the algorithm will generally converge to a local minimum. To find the global minimum, one usually resorts to trying several

† For a complete discussion, refer to [R-4].

different starting points in order to determine as many local minima as possible; the point $\mathbf{y}^*$ which corresponds to the local minimum having the smallest value of f is then selected as the best possible point.

Determination of Optimal Trajectories by Using Gradient Projection

Let us now discuss a technique, also due to Rosen,† for solving optimal control problems by using the gradient projection algorithm. The problem is to find an admissible control history $\mathbf{u}^*$ that causes the system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t)) \tag{6.6-47}$$

with known initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ to follow an admissible trajectory $\mathbf{x}^*$ that minimizes the performance measure

$$J = h(\mathbf{x}(t_f)) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t))\, dt. \tag{6.6-48}$$

For simplicity of notation, we shall assume that time does not appear explicitly in either the state equations or the performance measure; the solution of time-varying problems requires only straightforward modifications of the procedure to be described. It is also assumed that the final time t_f is specified, and since the equations are time-invariant, we can let $t_0 = 0$. Although the technique to be presented applies to problems involving general *linear* constraints among the state and control variables, we shall restrict our discussion to problems with constraints of the form

$$M_{i-} \leq u_i(t) \leq M_{i+}, \qquad t \in [0, t_f], \qquad i = 1, 2, \ldots, m \tag{6.6-49a}$$

$$S_{i-} \leq x_i(t) \leq S_{i+}, \qquad t \in [0, t_f], \qquad i = 1, 2, \ldots, n \tag{6.6-49b}$$

$$x_i(t_j) = T_{ij}, \qquad t_j \text{ specified}, \qquad i = 1, 2, \ldots, n. \tag{6.6-49c}$$

M_{i-} and M_{i+} denote the lower and upper bounds on the ith control component, S_{i-} and S_{i+} are the lower and upper bounds on the ith state component, and T_{ij} is the required value of the state component x_i at the time t_j.

Since gradient projection is an algorithm for minimizing a function of several variables, we must first approximate the optimal control problem to be solved by a discrete problem. To accomplish this, let us approximate the state differential equations by difference equations, and the integral

† See [R-5] and [R-6].

term in the performance measure by a summation. We shall use the simplest approximating difference equation, that is,

$$\mathbf{x}(t + \Delta t) \approx \mathbf{x}(t) + \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t)) \cdot \Delta t. \tag{6.6-50}$$

Assuming that the state is observed and the control changed only at the instants $t = 0, \Delta t, 2\,\Delta t, \ldots, N\,\Delta t$, we let $t = k\Delta T$, and

$$\mathbf{x}([k + 1]\,\Delta t) = \mathbf{x}(k\,\Delta t) + \mathbf{a}(\mathbf{x}(k\,\Delta t), \mathbf{u}(k\,\Delta t)) \cdot \Delta t, \tag{6.6-51}$$

or, defining $\mathbf{x}(k) \triangleq \mathbf{x}(k\,\Delta t)$ and $\mathbf{u}(k) \triangleq \mathbf{u}(k\,\Delta t)$,

$$\begin{aligned} \mathbf{x}(k + 1) &= \mathbf{x}(k) + \mathbf{a}(\mathbf{x}(k), \mathbf{u}(k)) \cdot \Delta t \\ &\triangleq a_D(\mathbf{x}(k), \mathbf{u}(k)). \end{aligned} \tag{6.6-52}$$

The performance measure can be written

$$J = h(\mathbf{x}(N\,\Delta t)) + \int_0^{\Delta t} g(\mathbf{x}(t), \mathbf{u}(t))\,dt + \cdots + \int_{(N-1)\,\Delta t}^{N\Delta t} g(\mathbf{x}(t), \mathbf{u}(t))\,dt, \tag{6.6-53}$$

which is approximated by

$$J_D \approx h(\mathbf{x}(N\,\Delta t)) + \Delta t \sum_{k=0}^{N-1} g(\mathbf{x}(k\,\Delta t), \mathbf{u}(k\,\Delta t)), \tag{6.6-54}$$

or

$$J_D = h(\mathbf{x}(N)) + \Delta t \sum_{k=0}^{N-1} g(\mathbf{x}(k), \mathbf{u}(k)). \tag{6.6-54a}$$

This approximation to the performance measure (6.6-48) is a *function* of the variables $\mathbf{x}(0), \mathbf{x}(1), \ldots, \mathbf{x}(N)$, and $\mathbf{u}(0), \mathbf{u}(1), \ldots, \mathbf{u}(N-1)$. Recalling that the state vector is of dimension n, and the control vector of dimension m, we see that there are $n[N + 1]$ (state values) and mN (control values), or a total of $n[N + 1] + mN$ variables, contained in J_D. Notice that these variables are not independent, because the approximating state difference equations (6.6-52) must be satisfied. Since $\mathbf{x}(0)$ is specified, our problem is to find the $N \cdot n + N \cdot m$ variables that minimize J_D, and satisfy the approximating state difference equations (6.6-52) and the constraints (6.6-49).

In our discussion of gradient projection it was assumed that the constraining relations were linear; however, the state difference equations of (6.6-52) may be nonlinear. To circumvent this difficulty we shall use a tech-

nique employed in the method of quasilinearization: linearize the state equations about a nominal state-control history, and solve a sequence of linearized problems. In the limit, the sequence of solutions to the linearized problems will converge (if certain technical requirements† are satisfied) to the solution of the discrete nonlinear problem. Assume that the ith state-control history $(\mathbf{x}^{(i)}(0), \mathbf{x}^{(i)}(1), \ldots, \mathbf{x}^{(i)}(N); \mathbf{u}^{(i)}(0), \mathbf{u}^{(i)}(1), \ldots, \mathbf{u}^{(i)}(N-1))$ is known; the initial state-control history is guessed. By expanding the $(i+1)$st trajectory in a Taylor series about the ith trajectory and retaining only terms of up to first order, we have

$$\mathbf{x}^{(i+1)}(k+1) = \mathbf{x}^{(i)}(k+1) + \left[\frac{\partial \mathbf{a}_D}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k))\right][\mathbf{x}^{(i+1)}(k) - \mathbf{x}^{(i)}(k)] + \left[\frac{\partial \mathbf{a}_D}{\partial \mathbf{u}}(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k))\right][\mathbf{u}^{(i+1)}(k) - \mathbf{u}^{(i)}(k)]. \tag{6.6-55}$$

Substituting $\mathbf{a}_D(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k))$ for $\mathbf{x}^{(i)}(k+1)$ and rearranging, we obtain

$$\begin{aligned}\mathbf{x}^{(i+1)}(k+1) = {} & \left[\frac{\partial \mathbf{a}_D}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k))\right]\mathbf{x}^{(i+1)}(k) \\ & + \left[\frac{\partial \mathbf{a}_D}{\partial \mathbf{u}}(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k))\right]\mathbf{u}^{(i+1)}(k) + \mathbf{a}_D(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k)) \\ & - \left[\frac{\partial \mathbf{a}_D}{\partial \mathbf{x}}(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k))\right]\mathbf{x}^{(i)}(k) \\ & - \left[\frac{\partial \mathbf{a}_D}{\partial \mathbf{u}}(\mathbf{x}^{(i)}(k), \mathbf{u}^{(i)}(k))\right]\mathbf{u}^{(i)}(k).\end{aligned} \tag{6.6-56}$$

Since $\mathbf{x}^{(i)}$ and $\mathbf{u}^{(i)}$ are known, (6.6-56) can be written

$$\mathbf{x}^{(i+1)}(k+1) = \mathbf{A}(k)\mathbf{x}^{(i+1)}(k) + \mathbf{B}(k)\mathbf{u}^{(i+1)}(k) + \mathbf{c}(k), \tag{6.6-56a}$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{c}$ are *known* time-varying matrices of appropriate dimensions that depend on the ith state-control history.

We could, at this point, proceed to minimize the function J_D of Eq. (6.6-54a) subject to the linearized state equation constraints (6.6-56a), and any additional constraints (6.6-49), which for the discrete problem would be of the form

† Rosen [R-5] has proved that the procedure converges, provided that: (1) The admissible state and control values lie in compact, convex sets [the constraints (6.6-49) guarantee this]. (2) The functions g and h in the performance measure are convex. (3) Each component of $\mathbf{a}$ in the state equations is either convex or concave for the admissible state-control values. We note that the method may still converge even if these conditions are not all satisfied; however, convergence is not assured.

$$M_{i-} \leq u_i(k) \leq M_{i+}, \qquad k = 0, 1, \ldots, N-1, \qquad i = 1, 2, \ldots, m \tag{6.6-57a}$$

$$S_{i-} \leq x_i(k) \leq S_{i+}, \qquad k = 0, 1, \ldots, N, \qquad i = 1, 2, \ldots, n \tag{6.6-57b}$$

$$x_i(j) = T_{ij}, \qquad j \text{ specified}, \qquad i = 1, 2, \ldots, n; \tag{6.6-57c}$$

however, there is an additional benefit to be derived from the linearization we have just performed. Since $\mathbf{x}(0) = \mathbf{x}_0$ is specified, $\mathbf{x}^{(i)}(0) = \mathbf{x}_0$ for all i. Let us write out a few terms of the solution of Eq. (6.6-56a):

$$\begin{aligned} \mathbf{x}^{(i+1)}(1) &= \mathbf{A}(0)\mathbf{x}_0 + \mathbf{B}(0)\mathbf{u}^{(i+1)}(0) + \mathbf{c}(0) \\ &\triangleq \mathbf{x}_H(1) + \mathbf{D}_0^1\mathbf{u}^{(i+1)}(0), \end{aligned} \tag{6.6-58}$$

where

$$\begin{aligned} \mathbf{x}_H(1) &\triangleq \mathbf{A}(0)\mathbf{x}_0 + \mathbf{c}(0), \quad \text{and} \quad \mathbf{D}_0^1 \triangleq \mathbf{B}(0); \\ \mathbf{x}^{(i+1)}(2) &= \mathbf{A}(1)[\mathbf{x}_H(1) + \mathbf{D}_0^1\mathbf{u}^{(i+1)}(0)] + \mathbf{B}(1)\mathbf{u}^{(i+1)}(1) + \mathbf{c}(1) \\ &= \mathbf{A}(1)\mathbf{x}_H(1) + \mathbf{c}(1) + \mathbf{A}(1)\mathbf{B}(0)\mathbf{u}^{(i+1)}(0) + \mathbf{B}(1)\mathbf{u}^{(i+1)}(1) \\ &\triangleq \mathbf{x}_H(2) + \mathbf{D}_0^2\mathbf{u}^{(i+1)}(0) + \mathbf{D}_1^2\mathbf{u}^{(i+1)}(1), \end{aligned} \tag{6.6-59}$$

and, in general,

$$\begin{aligned} \mathbf{x}^{(i+1)}(k+1) = \mathbf{A}(k)\mathbf{x}_H(k) &+ \mathbf{c}(k) + \mathbf{A}(k)\ldots\mathbf{A}(1)\mathbf{B}(0)\mathbf{u}^{(i+1)}(0) \\ &+ \mathbf{A}(k)\ldots\mathbf{A}(2)\mathbf{B}(1)\mathbf{u}^{(i+1)}(1) + \cdots \\ &+ \mathbf{A}(k)\mathbf{B}(k-1)\mathbf{u}^{(i+1)}(k-1) \\ &+ \mathbf{B}(k)\mathbf{u}^{(i+1)}(k) \\ \triangleq \mathbf{x}_H(k+1) &+ \mathbf{D}_0^{k+1}\mathbf{u}^{(i+1)}(0) + \mathbf{D}_1^{k+1}\mathbf{u}^{(i+1)}(1) + \cdots \\ &+ \mathbf{D}_{k-1}^{k+1}\mathbf{u}^{(i+1)}(k-1) + \mathbf{D}_k^{k+1}\mathbf{u}^{(i+1)}(k) \\ = \mathbf{x}_H(k+1) &+ \sum_{l=0}^{N-1} [\mathbf{D}_l^{k+1}\mathbf{u}^{(i+1)}(l)]. \end{aligned} \tag{6.6-60}$$

$\mathbf{x}_H(k+1)$ is the part of the solution for $\mathbf{x}^{(i+1)}(k+1)$ that does not depend on the control values $\mathbf{u}^{(i+1)}(0), \ldots, \mathbf{u}^{(i+1)}(N-1)$, and $\mathbf{D}_l^{k+1}$ is an $n \times m$ matrix that determines the contribution of the control at the lth instant to the state value at the $(k+1)$st instant.† $\mathbf{x}_H(k+1)$ and the $\mathbf{D}$ matrices are found from the relationships

† Note that the superscript $k+1$ on the matrix $\mathbf{D}$ does not indicate the $(k+1)$st power of $\mathbf{D}$.

$$\mathbf{x}_H(k+1) = \mathbf{A}(k)\mathbf{x}_H(k) + \mathbf{c}(k), \qquad \mathbf{x}_H(0) = \mathbf{x}_0, \tag{6.6-61}$$

and

$$\mathbf{D}_l^{k+1} = \begin{cases} \mathbf{A}(k)\mathbf{A}(k-1)\ldots\mathbf{A}(l+1)\mathbf{B}(l), & \text{for } k > l \\ \mathbf{B}(l), & \text{for } k = l \\ \mathbf{0}, & \text{for } k < l. \end{cases} \tag{6.6-62}$$

If the entire discrete state history is written in terms of the discrete control history in partitioned matrix form, we have

$$\left[\begin{array}{c} \mathbf{x}^{(i+1)}(0) \\ \hline \mathbf{x}^{(i+1)}(1) \\ \hline \mathbf{x}^{(i+1)}(2) \\ \hline \vdots \\ \hline \mathbf{x}^{(i+1)}(N) \end{array}\right] = \left[\begin{array}{c|c|c|c} \mathbf{D}_0^0 & \mathbf{D}_1^0 & \cdots & \mathbf{D}_{N-1}^0 \\ \hline \mathbf{D}_0^1 & \mathbf{D}_1^1 & \cdots & \mathbf{D}_{N-1}^1 \\ \hline \mathbf{D}_0^2 & \mathbf{D}_1^2 & \cdots & \mathbf{D}_{N-1}^2 \\ \hline \vdots & \vdots & & \vdots \\ \hline \mathbf{D}_0^N & \mathbf{D}_1^N & \cdots & \mathbf{D}_{N-1}^N \end{array}\right] \left[\begin{array}{c} \mathbf{u}^{(i+1)}(0) \\ \hline \mathbf{u}^{(i+1)}(1) \\ \hline \mathbf{u}^{(i+1)}(2) \\ \hline \vdots \\ \hline \mathbf{u}^{(i+1)}(N-1) \end{array}\right] + \left[\begin{array}{c} \mathbf{x}_H(0) \\ \hline \mathbf{x}_H(1) \\ \hline \mathbf{x}_H(2) \\ \hline \vdots \\ \hline \mathbf{x}_H(N) \end{array}\right] \tag{6.6-63}$$

or

$$\mathscr{X}^{(i+1)} = \mathscr{D}\mathscr{U}^{(i+1)} + \mathscr{X}_H. \tag{6.6-63a}$$

Notice that because of (6.6-62) the $\mathscr{D}$ matrix will contain an upper triangular matrix of zeros, and that the matrices $\mathbf{D}_j^0, j = 0, \ldots, N-1$, in the top row are all defined to be zero.

Equation (6.6-63) is quite important because it allows us to reduce substantially the number of variables used in the gradient projection algorithm. This is accomplished by replacing $\mathbf{x}^{(i+1)}(k)$ in the expression (6.6-54a) for J_D by $\mathbf{D}_0^k\mathbf{u}^{(i+1)}(0) + \mathbf{D}_1^k\mathbf{u}^{(i+1)}(1) + \cdots + \mathbf{D}_{N-1}^k\mathbf{u}^{(i+1)}(N-1) + \mathbf{x}_H(k)$, for $k = 0, 1, \ldots, N$. In addition, if there are inequality constraints involving the states, for example,

$$\left[\begin{array}{c} S_{1-} \\ S_{2-} \\ \vdots \\ S_{n-} \end{array}\right] \leq \mathbf{x}(k) \leq \left[\begin{array}{c} S_{1+} \\ S_{2+} \\ \vdots \\ S_{n+} \end{array}\right], \qquad k = 0, 1, \ldots, N, \tag{6.6-57b}$$

these relationships can be expressed as *linear* constraints involving only the control values:

$$\begin{bmatrix} S_{1-} \\ S_{2-} \\ \cdot \\ \cdot \\ \cdot \\ S_{n-} \end{bmatrix} \leq \mathbf{D}_0^k \mathbf{u}^{(i+1)}(0) + \mathbf{D}_1^k \mathbf{u}^{(i+1)}(1) + \cdots + \mathbf{D}_{N-1}^k \mathbf{u}^{(i+1)}(N-1) + \mathbf{x}_H(k)$$

$$\leq \begin{bmatrix} S_{1+} \\ S_{2+} \\ \cdot \\ \cdot \\ \cdot \\ S_{n+} \end{bmatrix}, \qquad k = 0, 1, \ldots, N. \tag{6.6-64}$$

To recapitulate, the problem to be solved is now of the form: Find the control values that satisfy the constraints

$$\begin{bmatrix} M_{1-} \\ M_{2-} \\ \cdot \\ \cdot \\ \cdot \\ M_{m-} \end{bmatrix} \leq \mathbf{u}^{(i+1)}(k) \leq \begin{bmatrix} M_{1+} \\ M_{2+} \\ \cdot \\ \cdot \\ \cdot \\ M_{m+} \end{bmatrix}, \qquad k = 0, 1, \ldots, N-1, \tag{6.6-65a}$$

$$\begin{bmatrix} S_{1-} \\ S_{2-} \\ \cdot \\ \cdot \\ \cdot \\ S_{n-} \end{bmatrix} \leq \mathbf{D}_0^k \mathbf{u}^{(i+1)}(0) + \mathbf{D}_1^k \mathbf{u}^{(i+1)}(1) + \cdots + \mathbf{D}_{N-1}^k \mathbf{u}^{(i+1)}(N-1) + \mathbf{x}_H(k)$$

$$\leq \begin{bmatrix} S_{1+} \\ S_{2+} \\ \cdot \\ \cdot \\ \cdot \\ S_{n+} \end{bmatrix}, \qquad k = 0, 1, \ldots, N, \tag{6.6-65b}$$

$$\begin{bmatrix} T_{1j} \\ T_{2j} \\ \cdot \\ \cdot \\ \cdot \\ T_{nj} \end{bmatrix} = \mathbf{D}_0^j \mathbf{u}^{(i+1)}(0) + \mathbf{D}_1^j \mathbf{u}^{(i+1)}(1) + \cdots + \mathbf{D}_{N-1}^j \mathbf{u}^{(i+1)}(N-1)$$

$$+ \mathbf{x}_H(j), \qquad j \text{ specified} \tag{6.6-65c}$$

and minimize the *function of Nm variables*

$$J_D = h(\mathscr{U}^{(i+1)}) + \Delta t \sum_{k=0}^{N-1} g(\mathscr{U}^{(i+1)}). \tag{6.6-66}$$

This expression for the performance measure simply indicates that only the control values $\mathbf{u}^{(i+1)}(0), \ldots, \mathbf{u}^{(i+1)}(N-1)$ appear explicitly, since Eq. (6.6-63) has been used to eliminate the presence of the state values.

A Summary of the Procedure for Solving Optimal Control Problems by Using Gradient Projection. The procedure we use to solve for an optimal control and its trajectory is:

1. Approximate the state differential equations by difference equations and the integral term of the performance measure by a summation; linearize the state difference equations.
2. Determine the expressions, in literal form, for any state constraints and the performance measure J_D in terms of $\mathbf{x}_H(k)$ $(k = 0, \ldots, N)$, $\mathbf{u}^{(i+1)}(k)$ $(k = 0, \ldots, N-1)$, and the **D** matrices of Eq. (6.6-63).
3. Guess a nominal state trajectory and control history, $\mathbf{x}^{(0)}$, $\mathbf{u}^{(0)}$. Set the iteration index i to zero.
4. Using the state-control history $\mathbf{x}^{(i)}$, $\mathbf{u}^{(i)}$, calculate the **A**, **B**, and **c** matrices, and use these matrices to determine $\mathscr{X}_H$ and $\mathscr{D}$.
5. Substitute the numerical values of $\mathscr{X}_H$ and $\mathscr{D}$ into the expressions obtained in step 2 to determine the coefficients in the constraining equations and the performance measure.
6. Minimize the function J_D, using the gradient projection algorithm.
7. Determine $\mathbf{x}^{(i+1)}$ by evaluating Eq. (6.6-63a) with $\mathbf{u}^{(i+1)}$ found in step 6.
8. If the norm of the difference between successive control iterates is small, that is,

$$\| \mathscr{U}^{(i+1)} - \mathscr{U}^{(i)} \| \leq \gamma, \tag{6.6-67}$$

terminate the procedure and output $\mathbf{x}^{(i+1)}$, $\mathbf{u}^{(i+1)}$, and the minimum value of J_D; otherwise increase i by one and return to step 4.

The reader will notice that in the above procedure the role of the gradient projection algorithm, described earlier and shown in Fig. 6-19, is as a subroutine that is called in step 6. Also note that steps 1 through 3 in the procedure are done off-line by the user; a digital computer program is used to perform steps 4 through 8.

To illustrate the details of the procedure, let us return to the continuous stirred-tank chemical reactor problem that was solved previously by using variational techniques.

Example 6.6-2. The state differential equations are

$$\begin{aligned}\dot{x}_1(t) &= -2[x_1(t) + 0.25] + [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right] \\ &\quad - [x_1(t) + 0.25]u(t) \qquad (6.6\text{-}68)\\ \dot{x}_2(t) &= 0.5 - x_2(t) - [x_2(t) + 0.5]\exp\left[\frac{25x_1(t)}{x_1(t) + 2}\right],\end{aligned}$$

and

$$J = \int_0^{0.78} [x_1^2(t) + x_2^2(t) + 0.1u^2(t)]\,dt \qquad (6.6\text{-}69)$$

is the performance measure to be minimized. The initial state value is $\mathbf{x}(0) = [0.05 \quad 0]^T$.

The approximating difference equations are

$$\begin{aligned}x_1(k+1) &= x_1(k) + \Delta t\Big[-2[x_1(k) + 0.25] \\ &\quad + [x_2(k) + 0.5]\exp\left[\frac{25x_1(k)}{x_1(k) + 2}\right] - [x_1(k) + 0.25]u(k)\Big] \qquad (6.6\text{-}70)\\ x_2(k+1) &= x_2(k) + \Delta t\Big[0.5 - x_2(k) \\ &\quad - [x_2(k) + 0.5]\exp\left[\frac{25x_1(k)}{x_1(k) + 2}\right]\Big] \qquad (6.6\text{-}71)\end{aligned}$$

and

$$\begin{aligned}J_D &= \Delta t \sum_{k=0}^{N-1} [x_1^2(k) + x_2^2(k) + 0.1u^2(k)] \\ &= \Delta t \sum_{k=0}^{N-1} [\mathbf{x}^T(k)\mathbf{x}(k) + 0.1u^2(k)] \qquad (6.6\text{-}72)\end{aligned}$$

is the approximate performance measure. Linearizing the difference equations gives

$$\begin{aligned}x_1^{(i+1)}(k+1) &= \left\{1 + \Delta t\left[-2 + \frac{50[x_2(k) + 0.5]\alpha_1}{[x_1(k) + 2]^2} - u(k)\right]\right\}_i x_1^{(i+1)}(k) \\ &\quad + \{\Delta t\alpha_1\}_i x_2^{(i+1)}(k) \\ &\quad + \{-\Delta t[x_1(k) + 0.25]\}_i u^{(i+1)}(k) \\ &\quad + \left\{\Delta t\left[-0.5 + 0.5\alpha_1 - \frac{50[x_2(k) + 0.5]x_1(k)\alpha_1}{[x_1(k) + 2]^2}\right.\right. \\ &\quad \left.\left. + x_1(k)u(k)\right]\right\}_t \qquad (6.6\text{-}73)\\ &\triangleq a_{11}(k)x_1^{(i+1)}(k) + a_{12}(k)x_2^{(i+1)}(k) + b_1(k)u^{(i+1)}(k) + c_1(k). \qquad (6.6\text{-}73a)\end{aligned}$$

$$x_2^{(i+1)}(k+1) = \left\{\frac{-50\,\Delta t[x_2(k)+0.5]\alpha_1}{[x_1(k)+2]^2}\right\}_i x_1^{(i+1)}(k) + \{1+\Delta t[-1-\alpha_1]\}_i x_2^{(i+1)}(k) + \left\{\Delta t\left[0.5-0.5\alpha_1+\frac{50x_1(k)[x_2(k)+0.5]\alpha_1}{[x_1(k)+2]^2}\right]\right\}_i \tag{6.6-74}$$

$$\triangleq a_{21}(k)x_1^{(i+1)}(k) + a_{22}(k)x_2^{(i+1)}(k) + c_2(k), \tag{6.6-74a}$$

where $\alpha_1 \triangleq \exp\{25x_1(k)/[x_1(k)+2]\}_i$ and $\{\ \}_i$ indicates that the quantity inside the braces is evaluated on the ith state-control history.

In the original problem the control values were assumed to be unbounded, but it is necessary that the gradient projection search be performed in a closed and bounded convex region; therefore, we introduce the artificial constraints

$$-2.0 \leq u(k) \leq 2.0, \qquad k = 0, 1, \ldots, N-1. \tag{6.6-75}$$

Since the control values are not really bounded, we may have to adjust these artificially imposed bounds so that the optimal control lies in the interior of the admissible region. The constraints specified by Eq. (6.6-75) must be expressed in the form $n_i u(k) - v_i \geq 0$. To accomplish this we write $u(k) \leq 2.0$ as

$$-u(k) + 2.0 \geq 0, \qquad k = 0, 1, \ldots, N-1, \tag{6.6-76a}$$

and $-2.0 \leq u(k)$ as

$$u(k) + 2.0 \geq 0, \qquad k = 0, 1, \ldots, N-1. \tag{6.6-76b}$$

Thus, implementation of the N constraints of Eq. (6.6-75) requires $2N$ constraint equations in the computational procedure.

Next, let us express the performance measure J_D entirely in terms of the control values. To achieve this, we substitute Eq. (6.6-60) into Eq. (6.6-72) with the result

$$J_D = \Delta t \sum_{k=0}^{N-1}\left[\left[\mathbf{x}_H(k) + \sum_{j=0}^{N-1}\mathbf{d}_j^k u^{(i+1)}(j)\right]^T\left[\mathbf{x}_H(k) + \sum_{j=0}^{N-1}\mathbf{d}_j^k u^{(i+1)}(j)\right] + 0.1\left[u^{(i+1)}(k)\right]^2\right]. \tag{6.6-77}$$

$\mathbf{x}_H(k)$ and $\mathbf{d}_j^k$ are evaluated from the ith state-control history. The $\mathbf{D}$ matrices of Eq. (6.6-60) are column vectors with two rows in this problem; hence, we write $\mathbf{d}_j^k$ rather than $\mathbf{D}_j^k$. Notice that $\mathbf{d}_j^0 = 0$, $j = 0, 1, \ldots, N-1$.

To obtain the lth component of the gradient of J_D with respect to $\mathcal{U}$ evaluated on the $(i+1)$st trajectory, we have

$$\left[\frac{\partial J_D}{\partial u(l)}\right]_{i+1} = \left\{\left[\frac{\partial J_D}{\partial \mathbf{x}(1)}\right]^T \frac{\partial \mathbf{x}(1)}{\partial u(l)} + \left[\frac{\partial J_D}{\partial \mathbf{x}(2)}\right]^T \frac{\partial \mathbf{x}(2)}{\partial u(l)} + \cdots + \left[\frac{\partial J_D}{\partial \mathbf{x}(N-1)}\right]^T \frac{\partial \mathbf{x}(N-1)}{\partial u(l)} + 0.2u(l)\right\}_{i+1},$$
$$l = 0, 1, \ldots, N-1.\dagger \tag{6.6-78}$$

Using the expression for J_D from (6.6-72) gives

$$\frac{\partial J_D}{\partial u(l)} = 2\,\Delta t\left\{\mathbf{x}^T(1)\frac{\partial \mathbf{x}(1)}{\partial u(l)} + \mathbf{x}^T(2)\frac{\partial \mathbf{x}(2)}{\partial u(l)} + \cdots + \mathbf{x}^T(N-1)\frac{\partial \mathbf{x}(N-1)}{\partial u(l)} + 0.1u(l)\right\}_{l+1}$$
$$= 2\,\Delta t\left[\sum_{k=1}^{N-1}\left[\mathbf{x}^T(k)\frac{\partial \mathbf{x}(k)}{\partial u(l)}\right] + 0.1u(l)\right]_{l+1}. \tag{6.6-79}$$

From the relationship for $\mathbf{x}(k)$ used in (6.6-77),

$$\frac{\partial \mathbf{x}(k)}{\partial u(l)} = \mathbf{d}_l^k; \tag{6.6-80}$$

hence,

$$\frac{\partial J_D}{\partial u(l)} = 2\,\Delta t\left[\sum_{k=1}^{N-1}\left\{\left[\mathbf{x}_H(k) + \sum_{j=0}^{N-1}\mathbf{d}_j^k u(j)\right]^T \mathbf{d}_l^k\right\} + 0.1u(l)\right]_{l+1},$$
$$l = 0, 1, \ldots, N-1. \tag{6.6-81}$$

The initial guess selected for the state-control history was

$$\mathbf{x}^{(0)}(k) = \mathbf{0}, \qquad k = 1, \ldots, N-1$$
$$u^{(0)}(k) = 0 \qquad k = 0, 1, \ldots, N-1.$$

A FORTRAN IV program was used to solve this problem on an IBM 360/67 digital computer. The time increment Δt used in the approximating difference equations was 0.01 unit. Calculations were performed using single-precision arithmetic, and the termination criterion was

$$\|u^{(i+1)} - u^{(i)}\| = \max_{\substack{k \\ k=0,\ldots,N-1}} |u^{(i+1)}(k) - u^{(i)}(k)| \leq 1.0 \times 10^{-3}$$
$$(N = 78). \tag{6.6-82}$$

When this stopping criterion was used, the algorithm required 15 iterations to converge to the control and trajectory shown in Figs. 6-20 and 6-21; the minimum found for the performance measure was $J^* = 0.02725$. After the third iteration, performance measure changes were in the fifth

† $\partial \mathbf{x}(k)/\partial u(l)$ is a column vector with components $\partial x_1(k)/\partial u(l)$, and $\partial x_2(k)/\partial u(l)$.

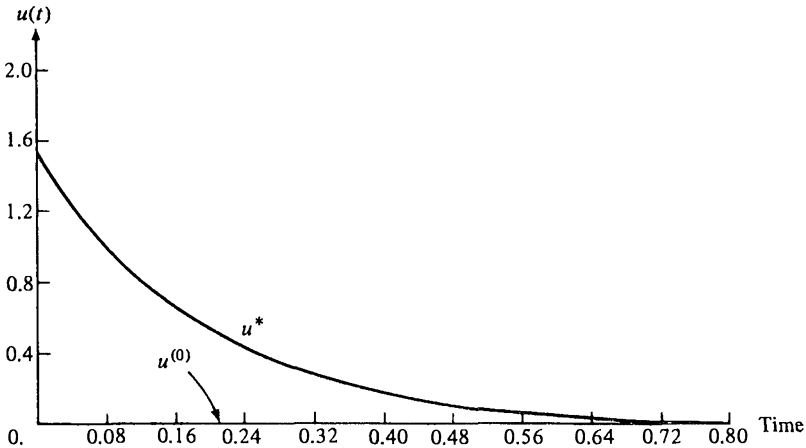


Figure 6-20 The optimal control for the stirred-tank reactor (gradient projection solution)

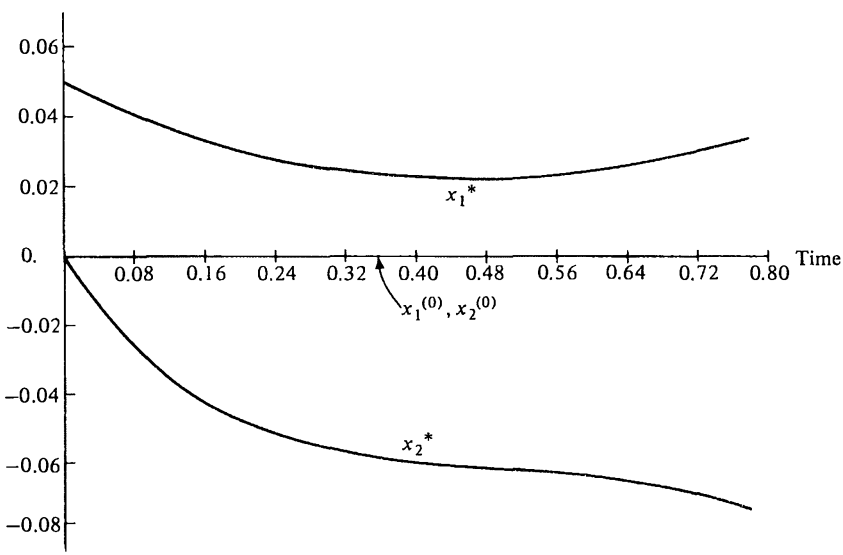


Figure 6-21 The optimal trajectory for the stirred-tank reactor (gradient projection solution)

significant figure, and the difference between J^* and $J^{(3)}$ (the performance measure after three iterations) was 0.00004.

Comparing the minimum cost, trajectory, and control found by gradient projection with the results obtained previously by using variational techniques, we observe slight numerical differences. These small

deviations are attributed primarily to the difference equations used to approximate the state equations. For a discussion of alternative difference equation approximations, see [R-6].

In the preceding example the performance measure contained a term that penalized the expenditure of control effort. It may be that control effort is to be conserved; however, since u represents the effect of the flow of a coolant, it is more likely that in this physical system the control is constrained by bounds and that the term containing u in the performance measure is a penalty function. If variational techniques are used, the penalty function approach may simplify numerical procedures, because the control is treated as if it were not bounded and the methods discussed in Sections 6.2 through 6.5 can be applied.† The cost of this simplification, however, is that the mathematical model is less accurate than if the control constraints were included.

The gradient projection method, on the other hand, allows routine incorporation of state and control inequality constraints in the problem solution. In fact, the gradient projection algorithm *requires* that the admissible controls lie in a bounded region; it was this requirement that caused us to introduce the artificial bounds $-2.0 \leq u(t) \leq 2.0$ in Example 6.6-2. To illustrate further the inclusion of state and control constraints, let us now consider a modified version of the continuous stirred-tank chemical reactor problem.

Example 6.6-3. The state equations are given by Eqs. (6.6-68), but the performance measure does not contain a penalty function involving the control; hence,

$$J = \int_0^{0.78} [x_1^2(t) + x_2^2(t)]\, dt. \tag{6.6-83}$$

The admissible controls are required to satisfy the constraints

$$-1.0 \leq u(t) \leq 1.0, \qquad t \in [0, 0.78]. \tag{6.6-84}$$

In addition, suppose that at $t = 0.78$, the state must be at the origin; that is,

$$\mathbf{x}(0.78) = \mathbf{x}(N) = \mathbf{0}. \tag{6.6-85}$$

This reformulation of the original stirred-tank reactor problem requires that only minor modifications be made to the computer program. Specifically, the control constraints become

† The techniques discussed in Sections 6.2 through 6.5 can be modified to handle problems that include inequality and equality constraints—see [S-3]. The constraints do complicate the algorithms, however.

$$-u(k) + 1.0 \geq 0, \qquad k = 0, 1, \ldots, N-1 \tag{6.6-86a}$$

$$u(k) + 1.0 \geq 0, \qquad k = 0, 1, \ldots, N-1, \tag{6.6-86b}$$

and the expressions for J_D and its gradient are

$$J_D = \Delta t \sum_{k=0}^{N-1} \left[\left[\mathbf{x}_H(k) + \sum_{j=0}^{N-1} \mathbf{d}_j^k u^{(i+1)}(j) \right]^T \left[\mathbf{x}_H(k) + \sum_{j=0}^{N-1} \mathbf{d}_j^k u^{(i+1)}(j) \right] \right] \tag{6.6-87}$$

$$\frac{\partial J_D}{\partial u(l)} = 2\,\Delta t \sum_{k=1}^{N-1} \left\{ \left[\mathbf{x}_H(k) + \sum_{j=0}^{N-1} \mathbf{d}_j^k u^{(i+1)}(j) \right]^T \mathbf{d}_l^k \right\}, \qquad l = 0, 1, \ldots, N-1. \tag{6.6-88}$$

The equality constraint on the terminal state values can be included by using the four inequality constraints

$$x_1(N) \geq 0 \tag{6.6-89a}$$

$$-x_1(N) \geq 0 \tag{6.6-89b}$$

$$x_2(N) \geq 0 \tag{6.6-90a}$$

$$-x_2(N) \geq 0. \tag{6.6-90b}$$

Taken together, the inequalities of (6.6-89) can be satisfied only by $x_1(N) = 0$. Similarly, the inequalities of (6.6-90) are satisfied only by $x_2(N) = 0$. In the computer program, these inequalities must be expressed in terms of the solution of the linearized state equations. The appropriate expressions are

$$\mathbf{x}_H(N) + \sum_{j=0}^{N-1} \mathbf{d}_j^N u^{(i+1)}(j) \geq \mathbf{0}, \tag{6.6-91}$$

which represents (6.6-89a) and (6.6-90a), and

$$-\left[\mathbf{x}_H(N) + \sum_{j=0}^{N-1} \mathbf{d}_j^N u^{(i+1)}(j) \right] \geq \mathbf{0}, \tag{6.6-92}$$

which represents (6.6-89b) and (6.6-90b).

With these additional constraints included in the program, the algorithm converged in four iterations to a minimum cost of $J^* = 0.00220$ with

$$\| u^{(i+1)} - u^{(i)} \| \triangleq \max_{\substack{k \\ k=0,1,\ldots,N-1}} |u^{(i+1)}(k) - u^{(i)}(k)| \leq 1.0 \times 10^{-4}. \tag{6.6-93}$$

The initial guess and the optimal control and trajectory are shown in Figs. 6-22 and 6-23. The final state values were

$$\mathbf{x}(0.78) = \begin{bmatrix} -6.167 \times 10^{-6} \\ -0.631 \times 10^{-6} \end{bmatrix}.$$

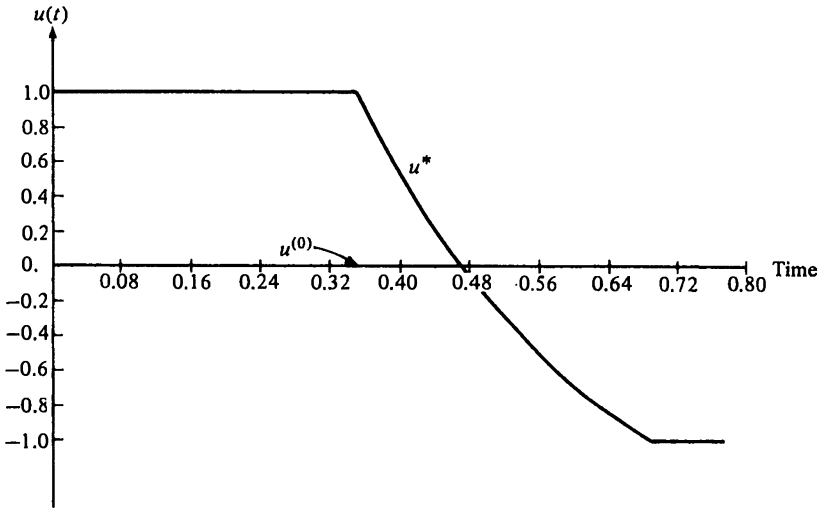


Figure 6-22 The optimal control for Example 6.6-3

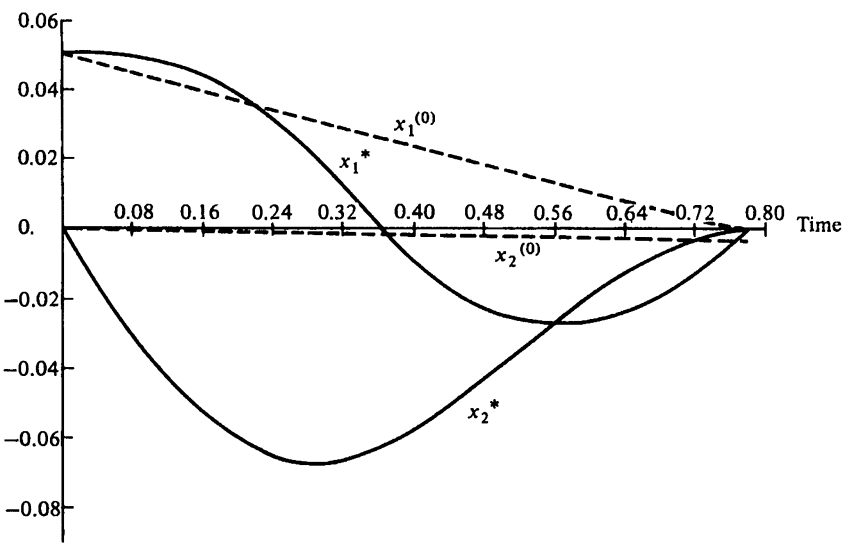


Figure 6-23 The optimal trajectory for Example 6.6-3

Summary

To summarize gradient projection solution of optimal control problems, let us enumerate the salient features of the computational procedure.

The Initial Guess. A state trajectory $\mathbf{x}^{(0)}(k)$ $(k = 0, 1, \ldots, N)$ and a control history $\mathbf{u}^{(0)}(k)$ $(k = 0, 1, \ldots, N-1)$ are required in order to begin the iterative procedure. In selecting these initial state and control histories, we use any available knowledge about the expected form of the optimal trajectory and control.

Storage Requirements. The current trial control $\mathscr{U}^{(i)}$, and $\mathscr{D}$ and $\mathscr{X}_H$ of Eq. (6.6-63a) must be stored. In addition, the gradient projection algorithm, which serves as a subroutine, also requires storage; the projection matrix $\mathbf{P}_q$ and the matrix $\mathbf{N}_L$ of Eq. (6.6-3b) account for most of this storage requirement.

Convergence. In several test examples it was observed that convergence of the algorithm occurred for a variety of initial guesses. The procedure generally converged in only a few iterations (often less than 10).

Computations Required. In each iteration a nonlinear programming problem is solved by using the gradient projection algorithm. To begin a new iteration, the trajectory $\mathbf{x}^{(i+1)}$ and the matrices $\mathscr{D}$ and $\mathscr{X}_H$ must be recomputed.

Stopping Criterion. The iterative procedure is terminated when a measure of the deviation of successive iterates becomes small. The stopping criterion used in the examples was

$$\|\mathbf{u}^{(i+1)} - \mathbf{u}^{(i)}\| \leq \gamma, \tag{6.6-94}$$

where γ is a preselected positive number, and

$$\|\mathbf{u}^{(i+1)} - \mathbf{u}^{(i)}\| \triangleq \sum_{j=1}^{m} \left\{ \max_{\substack{k \\ k=0,1,\ldots,N-1}} |u_j^{(i+1)}(k) - u_j^{(i)}(k)| \right\}. \tag{6.6-95}$$

Modifications for Fixed End Point Problems. As we have discussed and illustrated, fixed end point problems are a routine matter when gradient projection is the algorithm. In addition, state and control inequality and equality constraints at times throughout the interval $[t_0, t_f]$ are easily handled. This capability of solving problems with constraints is one of the strong selling points for the gradient projection algorithm.

REFERENCES

B-5 Bryson, A. E., Jr., and W. F. Denham, "Optimal Programming Problems with Inequality Constraints II: Solution by Steepest Ascent," *AIAA Journal* (1964), 25–34.

F-1 Fox, L., *Numerical Solution of Ordinary and Partial Differential Equations.* Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1962.

K-8 Kelley, H. J., "Method of Gradients," *Optimization Techniques with Applications To Aerospace Systems*, G. Leitmann, ed. New York: Academic Press Inc., 1962.

K-9 Kelley, H. J., "Gradient Theory of Optimal Flight Paths," *ARS Journal* (1960), 947–953.

L-5 Lapidus, L., and R. Luus, "The Control of Nonlinear Systems; Part II: Convergence by Combined First and Second Variations," *A. I. Ch. E. Journal* (1967), 108–113.

M-4 McGill, R., and P. Kenneth, "A Convergence Theorem on the Iterative Solution of Nonlinear Two-Point Boundary-Value Systems," reprint of paper presented at the XIVth International Astronautical Congress, Paris, France, Sept. 1963.

M-5 McGill, R., and P. Kenneth, "Solution of Variational Problems by Means of a Generalized Newton-Raphson Operator," *AIAA Journal* (1964), 1761–1766.

R-4 Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints," *J. SIAM* (1960), 181–217.

R-5 Rosen, J. B., "Iterative Solution of Nonlinear Optimal Control Problems," *J. SIAM Control* Series A (1966), 223–244.

R-6 Rosen, J. B., "Optimal Control and Convex Programming," *Nonlinear Programming*, J. Abadie, ed. New York: John Wiley & Sons, Inc., 1967.

S-3 Sage, A. P., *Optimum Systems Control*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1968.

PROBLEMS

6-1. The iteration equation for variation of extremals, Eq. (6.3-18), has been derived for the case where the final time t_f is fixed and the performance measure contains no terms that are explicitly dependent upon the final states ($h = 0$). The purpose of this problem is to extend Eq. (6.3-18) to the situation where

$$J = h(\mathbf{x}(t_f)) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt;$$

the final time is assumed to be fixed.

(a) Assume that $h(\mathbf{x}(t_f)) = \mathbf{d}^T\mathbf{x}(t_f)$, where $\mathbf{d} \neq \mathbf{0}$ is an $n \times 1$ matrix of constants. Derive the modified version of Eq. (6.3-18).

(b) Derive the equation analogous to (6.3-18) if $h(\mathbf{x}(t_f)) \triangleq \mathbf{x}^T(t_f)\mathbf{H}\mathbf{x}(t_f)$. $\mathbf{H}$ is a real symmetric positive semi-definite matrix.

(c) Repeat (b) for the case where $h(\mathbf{x}(t_f))$ is some general nonlinear twice-differentiable, scalar function. Compare with Eq. (6.3-20). *Hint.* Ap-

proximate $\partial h(\mathbf{x}^{(i+1)}(t_f))/\partial \mathbf{x}$ by the first two terms of a Taylor series about $\mathbf{x}^{(i)}(t_f)$.

(d) Show that the equation derived in part (c) yields the results obtained in parts (a) and (b) and Eq. (6.3-18) as special cases.

6-2. Show that variation of extremals converges in one iteration if the control problem is of the linear regulator type (see Section 5.2). Assume $\mathbf{H} = \mathbf{0}$.

6-3. Consider a two-point boundary-value problem in which the initial and final state values are specified. Find the iteration equation [corresponding to Eq. (6.3-20)] that would be used to solve this problem by the method of variation of extremals.

6-4. Repeat Problem 6-3 for the case where only some of the final states are specified.

6-5. In using variation of extremals it may be easier to make a reasonable guess for the missing values at the final time than it is to guess the missing initial costate values.

(a) If the missing final values are guessed, the iteration equation (6.3-20) must be modified. Find the modified version of (6.3-20) for the case where $\mathbf{x}(t_f)$ is free and $\mathbf{p}(t_f) = \mathbf{0}$.

(b) What would your initial guess be for the free final states if

$$J = \int_0^{t_f} [\mathbf{x}^T(t)\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{u}(t)]\, dt?$$

Why?

6-6. Consider the $2n$ linear, time-varying differential equations

$$\dot{\mathbf{z}}(t) = \mathbf{D}(t)\mathbf{z}(t) + \mathbf{f}(t); \tag{1}$$

$\mathbf{D}(t)$ is a $2n \times 2n$ matrix, and $\mathbf{f}(t)$ is a $2n \times 1$ vector of known functions of time. Show that if $\mathbf{z}^{H1}(t), \mathbf{z}^{H2}(t), \ldots, \mathbf{z}^{Hq}(t)$ are q solutions of the homogeneous equations

$$\dot{\mathbf{z}}(t) = \mathbf{D}(t)\mathbf{z}(t),$$

and $\mathbf{z}^p(t)$ is a solution of (1), then

$$\mathbf{y}(t) \triangleq c_1\mathbf{z}^{H1}(t) + \cdots + c_q\mathbf{z}^{Hq}(t) + \mathbf{z}^p(t)$$

is a solution of (1) for *arbitrary* values of the constants $c_1, \ldots, c_q$.

6-7. Solve Problem 6-3 for the method of quasilinearization.

6-8. Repeat Problem 6-7 for the case where only some of the final states are specified.

6-9. Show that quasilinearization converges in one iteration if used to solve a linear regulator problem.

6-10. The purpose of this problem is to extend the iteration equation of quasilinearization, Eq. (6.4-35), to the case where the final costate $\mathbf{p}(t_f)$ is not a known constant. The final time t_f is assumed to be specified.

(a) Derive the appropriate iteration equation if $h(\mathbf{x}(t_f)) = \mathbf{x}^T(t_f)\mathbf{H}\mathbf{x}(t_f)$. $\mathbf{H}$ is a real symmetric positive semi-definite matrix.

(b) Derive the iteration equation that applies when h is some general nonlinear twice-differentiable function of $\mathbf{x}(t_f)$. *Hint.* Approximate $\partial h(\mathbf{x}^{(i+1)}(t_f))/\partial \mathbf{x}$ by the first two terms of a Taylor series about $\mathbf{x}^{(i)}(t_f)$.

(c) The equation derived in part (b) should be the same as Eq. (6.4-38). Show that the equation derived in part (a) and Eq. (6.4-35) are special cases of Eq. (6.4-38).

6-11. Consider the similarities between the differential equations that arise in the linear tracking problem (see Section 5.2) and in the method of quasilinearization. Solution of the linear tracking problem requires no matrix inversion. Use the similarities observed to modify the method of quasilinearization so that no matrix inversion is required.

6-12. From Section 5.2 we know that the closed-loop solution to a linear regulator problem can be obtained by integrating the matrix Riccati equation. If $\mathbf{x}(t_f)$ is free, the boundary conditions for the matrix $\mathbf{K}$ are $\mathbf{K}(t_f) = \mathbf{H}$; however, if $\mathbf{x}(t_f)$ is fixed, this is not the case. Use the principles contained in the method of quasilinearization to devise a method for obtaining a set of boundary conditions for the $\mathbf{K}$ matrix for the case where $\mathbf{x}(t_f)$ is fixed.

The following problems are numerical solutions (requiring a digital computer) of optimization problems. In order that the student may gain familiarity with the numerical methods discussed in this chapter, Problems 6-19 through 6-33 are of the linear regulator type. The answers to these problems can, and should be, verified by integrating the Riccati equation as indicated in Section 5.2. It should be emphasized that the methods of Section 5.2 lead to the optimal control law, whereas the techniques discussed in this chapter yield an open-loop optimal control.

6-13. Using the gradient projection algorithm, determine the value of $\mathbf{y}$ that satisfies the constraints $y_1 \geq 0$, $y_2 \geq 0$, $2y_1 + 5y_2 \geq 6$, $y_1 + y_2 \geq 2$, and minimizes $f(\mathbf{y}) = y_1 + 2y_2$.

6-14. Use the gradient projection method to find the point $\mathbf{y}^*$ where

$$f(\mathbf{y}) = y_1 + 2y_2 + 3y_3$$

has its maximum value. The variables y_1, y_2, and y_3 are required to satisfy the constraints $y_1 \geq 0$, $y_2 \geq 0$, $y_3 \geq 0$, $-y_1 - y_2 - y_3 + 1 \geq 0$, $y_1 + y_2 - y_3 \geq 0$, $y_1 - 2y_2 \geq 0$.

6-15. Using gradient projection, determine the value of $\mathbf{y}$ that satisfies the constraints $y_1 \geq 0$, $y_2 \geq 0$, $y_1 - y_2 \geq -5$, $-0.2y_1 - y_2 \geq -8$, $-y_1 \geq -20$, and maximizes $f(\mathbf{y}) = y_1 - 10[y_2 - 1]^2$.

6-16. Use the gradient projection algorithm to find the maximum of the function $f(\mathbf{y}) = y_1 + y_3 + 1.25y_4 + 2y_5 + 2y_7 + 1.25y_8 - y_9$, where the variables

must satisfy the constraints $y_1, y_2, \ldots, y_9 \geq 0$, $y_9 \leq 1$, $0.5y_9 \geq y_1 + y_2$, $0.25y_9 \geq y_5 + y_6$, $0.5y_2 + 0.375y_6 \geq y_3 + y_4$, $0.3333y_2 + 0.625y_6 \geq y_7 + y_8$.

6-17. Use the gradient projection method to find the maximum of $f(\mathbf{y}) = 2y_1 - y_1^2 + y_2$, subject to the constraints $y_1 \geq 0$, $y_2 \geq 0$, $2y_1^2 + 3y_2^2 \leq 6$. *Hint.* Linearize the constraints, and solve a sequence of problems with linear constraints.

6-18. Determine, by using the gradient projection algorithm, the minimum value of $f(\mathbf{y}) = [y_1 - 5]^2 + [y_2 - 8]^2$, where the constraints $y_1^2 + y_2^2 \leq 6$, $y_1 \geq 0$, $y_2 \geq 0$ must be satisfied.

6-19. Find the optimal trajectory and control for the linear regulator problem

$$\dot{x}(t) = x(t) + u(t), \qquad x(0) = 4.0$$

$$J = \tfrac{1}{2} \int_0^1 [x^2(t) + u^2(t)]\, dt$$

by using the steepest descent method.

6-20. Repeat Problem 6-19, using variation of extremals.

6-21. Repeat Problem 6-19, using quasilinearization.

6-22. Repeat Problem 6-19, using the gradient projection method.

6-23. Use the steepest descent method to find an optimal trajectory and control for the system

$$\dot{x}(t) = -x(t) + u(t), \qquad x(0) = 4.0$$

and the performance measure

$$J = x^2(1) + \int_0^1 \tfrac{1}{2}u^2(t)\, dt.$$

6-24. Repeat Problem 6-23, using variation of extremals.

6-25. Repeat Problem 6-23, using quasilinearization.

6-26. Repeat Problem 6-23, using the gradient projection method.

6-27. Use the method of steepest descent to minimize the performance measure

$$J = \mathbf{x}^T(T)\mathbf{H}\mathbf{x}(T) + \int_0^T [\mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t) + \mathbf{R}u^2(t)]\, dt$$

subject to the differential equation constraints

$$\dot{x}_1(t) = x_2(t),$$
$$\dot{x}_2(t) = -x_1(t) + u(t), \qquad \mathbf{x}(0) = \begin{bmatrix} 10 \\ -5 \end{bmatrix}.$$

The final time T is 1.0, and $R = 1$. Consider the following cases for $\mathbf{Q}$ and $\mathbf{H}$:

(a) $\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$ $\mathbf{H} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

(b) $\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$ $\mathbf{H} = \begin{bmatrix} 10 & 0 \\ 0 & 20 \end{bmatrix}$

(c) $\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$ $\mathbf{H} = \begin{bmatrix} 10^6 & 0 \\ 0 & 10^6 \end{bmatrix}$.

6-28. Repeat Problem 6-27, using variation of extremals.

6-29. Repeat part (c) of Problem 6-27, using variation of extremals, but let

$$\mathbf{H} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{x}(1) = \mathbf{0}.$$

Compare with Problem 6-27(c).

6-30. Repeat Problem 6-27, using quasilinearization.

6-31. Repeat Problem 6-29, using quasilinearization.

6-32. Repeat Problem 6-27, using gradient projection.

6-33. Repeat Problem 6-29, using gradient projection.

Now that your numerical skills have been polished on linear regulator problems, you are ready to tackle the following nonlinear problems.

6-34. Solve the continuous stirred-tank chemical reactor problem introduced in Example 6.2-2 by the method of steepest descent. Investigate the effects of various initial guesses.

6-35. Repeat Problem 6-34, using variation of extremals.

6-36. Repeat Problem 6-34, using quasilinearization.

6-37. Repeat Problem 6-34, using the gradient projection technique.

V

Conclusion

7

Summation

Before reviewing our progress, let us investigate the relationship between dynamic programming and the minimum principle.

7.1 THE RELATIONSHIP BETWEEN DYNAMIC PROGRAMMING AND THE MINIMUM PRINCIPLE

We have considered the problem of finding a control $\mathbf{u}^* \in U$ that causes a system

$$\dot{\mathbf{x}}(t) = \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{7.1-1}$$

to respond in such a manner that a performance measure of the form

$$J = h(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} g(\mathbf{x}(t), \mathbf{u}(t), t)\, dt \tag{7.1-2}$$

is minimized.

In our discussion of dynamic programming we showed that the Hamilton-Jacobi-Bellman functional equation

$$0 = J_t^*(\mathbf{x}(t), t) + \min_{\mathbf{u}(t)} \{g(\mathbf{x}(t), \mathbf{u}(t), t) + [J_{\mathbf{x}}^*(\mathbf{x}(t), t)]^T \mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)\} \tag{7.1-3}$$

must be satisfied by an optimal control and its trajectory. $J^*(\mathbf{x}(t), t)$ is the minimum value of the performance measure for a process beginning at time t with an initial state $\mathbf{x}(t)$, and J_t^* and $J_{\mathbf{x}}^*$ are partial derivatives of $J^*(\mathbf{x}(t), t)$ with respect to t and $\mathbf{x}$. If $(\mathbf{x}^*(t), t)$ is a particular point in the state-time space, the H-J-B functional equation tells us that the optimal control value $\mathbf{u}^*(t)$, which corresponds to this point, satisfies the relationship

$$\begin{aligned} &g(\mathbf{x}^*(t), \mathbf{u}^*(t), t) + [J_{\mathbf{x}}^*(\mathbf{x}^*(t), t)]^T \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \\ &= \min_{\mathbf{u}(t)} \{g(\mathbf{x}^*(t), \mathbf{u}(t), t) + [J_{\mathbf{x}}^*(\mathbf{x}^*(t), t)]^T \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}(t), t)\}, \end{aligned} \tag{7.1-4}$$

for all $t \in [t_0, t_f]$. Thus, we can write Eq. (7.1-3) for the point $(\mathbf{x}^*(t), t)$ as

$$0 = J_t^*(\mathbf{x}^*(t), t) + g(\mathbf{x}^*(t), \mathbf{u}^*(t), t) + [J_{\mathbf{x}}^*(\mathbf{x}^*(t), t)]^T \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t). \tag{7.1-3a}$$

Equation (7.1-3a) is a first-order partial differential equation. If t_f is fixed and $\mathbf{x}(t_f)$ free, the boundary condition is

$$J^*(\mathbf{x}^*(t_f), t_f) = h(\mathbf{x}^*(t_f), t_f). \tag{7.1-5}$$

If Pontryagin's minimum principle is applied to the same optimal control problem, we find that

$$\dot{\mathbf{x}}^*(t) = \frac{\partial \mathscr{H}}{\partial \mathbf{p}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \tag{7.1-6}$$

$$\dot{\mathbf{p}}^*(t) = -\frac{\partial \mathscr{H}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \tag{7.1-7}$$

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) \leq \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t) \tag{7.1-8}$$

for all admissible $\mathbf{u}(t)$, and for all $t \in [t_0, t_f]$, are necessary conditions for $\mathbf{u}^*$ to be an optimal control and $\mathbf{x}^*$ an optimal trajectory. The boundary conditions for the $2n$ first-order state-costate differential equations (7.1-6) and (7.1-7) are

$$\mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{7.1-9}$$

and

$$\mathbf{p}^*(t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f). \tag{7.1-10}$$

Using the definition of the Hamiltonian

$$\mathscr{H}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t), t) \triangleq g(\mathbf{x}(t), \mathbf{u}(t), t) + \mathbf{p}^T(t)[\mathbf{a}(\mathbf{x}(t), \mathbf{u}(t), t)],$$

we can write Eqs. (7.1-6) and (7.1-7) as

$$\dot{\mathbf{x}}^*(t) = \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \tag{7.1-6a}$$

$$\dot{\mathbf{p}}^*(t) = -\left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T \mathbf{p}^*(t) - \left[\frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]. \tag{7.1-7a}$$

In addition, Eq. (7.1-8) implies that

$$\mathscr{H}(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}^*(t), t) = \min_{\mathbf{u}(t)} \mathscr{H}(\mathbf{x}^*(t), \mathbf{u}(t), \mathbf{p}^*(t), t), \tag{7.1-11}$$

or

$$\begin{aligned} &g(\mathbf{x}^*(t), \mathbf{u}^*(t), t) + \mathbf{p}^{*T}(t)[\mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)] \\ &= \min_{\mathbf{u}(t)} \{g(\mathbf{x}^*(t), \mathbf{u}(t), t) + \mathbf{p}^{*T}(t)[\mathbf{a}(\mathbf{x}^*(t), \mathbf{u}(t), t)]\}. \end{aligned} \tag{7.1-11a}$$

Comparing Eq. (7.1-11a) with Eq. (7.1-4), we observe an interesting similarity: These two equations have exactly the same form; in fact, *if*

$$J_{\mathbf{x}}^*(\mathbf{x}^*(t), t) = \mathbf{p}^*(t), \tag{7.1-12}$$

the two equations are identical.

As you suspect, this similarity is more than coincidental; in fact, let us now show that the equations that constitute Pontryagin's minimum principle can be derived from the Hamilton-Jacobi-Bellman functional equation.

First, notice that Eq. (7.1-3a) can be written as

$$0 = \min_{\mathbf{u}(t)} \{J_t^*(\mathbf{x}^*(t), t) + g(\mathbf{x}^*(t), \mathbf{u}(t), t) + [J_{\mathbf{x}}^*(\mathbf{x}^*(t), t)]^T \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}(t), t)\}, \tag{7.1-3b}$$

since J_t^* does not depend on $\mathbf{u}(t)$. In words, given the state value $\mathbf{x}^*(t)$, the control $\mathbf{u}^*(t)$ minimizes the right side of Eq. (7.1-3b), and *the minimum value is zero*. Now consider state values in a neighborhood of $\mathbf{x}^*(t)$; that is, let

$$\mathbf{x}(t) = \mathbf{x}^*(t) + \delta\mathbf{x}(t), \tag{7.1-13}$$

where $\|\delta\mathbf{x}\| < \epsilon$. We assert that

$$\begin{aligned} &J_t^*(\mathbf{x}^*(t), t) + g(\mathbf{x}^*(t), \mathbf{u}^*(t), t) + [J_{\mathbf{x}}^*(\mathbf{x}^*(t), t)]^T \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \\ &\leq J_t^*(\mathbf{x}(t), t) + g(\mathbf{x}(t), \mathbf{u}^*(t), t) + [J_{\mathbf{x}}^*(\mathbf{x}(t), t)]^T \mathbf{a}(\mathbf{x}(t), \mathbf{u}^*(t), t). \end{aligned} \tag{7.1-14}$$

Why is this the case? From Eq. (7.1-3b) we know that the left side of (7.1-14) is equal to zero. The minimum value of the right side (also zero) is attained only if $\mathbf{u}^*(t)$ is an extremal control for the state $\mathbf{x}(t)$ as well as for $\mathbf{x}^*(t)$. To cut through the maze of notation somewhat, let us write (7.1-14) as

$$v(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \leq v(\mathbf{x}(t), \mathbf{u}^*(t), t). \tag{7.1-15}$$

This equation tells us that for fixed $\mathbf{u}^*(t)$ and t the scalar function v has a local minimum at the point $\mathbf{x}(t) = \mathbf{x}^*(t)$; therefore,

$$\frac{\partial v}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) = \mathbf{0}, \tag{7.1-16}$$

if it is assumed that $\mathbf{x}(t)$ is not constrained by any boundaries. Writing out the terms in $v(\mathbf{x}^*(t), \mathbf{u}^*(t), t)$ yields

$$v = J_t^* + g + J_{x_1}^* a_1 + J_{x_2}^* a_2 + \cdots + J_{x_n}^* a_n; \tag{7.1-17}$$

for simplicity, the arguments $\mathbf{x}^*(t)$, $\mathbf{u}^*(t)$, and t have been omitted. Taking the gradient of v with respect to $\mathbf{x}$ and using (7.1-16) gives the equations (again we omit arguments)

$$\begin{aligned}
\frac{\partial v}{\partial x_1} &= J_{tx_1}^* + g_{x_1} + J_{x_1x_1}^* a_1 + J_{x_1}^* a_{1x_1} + J_{x_2x_1}^* a_2 + J_{x_2}^* a_{2x_1} \\
&\quad + \cdots + J_{x_nx_1}^* a_n + J_{x_n}^* a_{nx_1} = 0 \\
\frac{\partial v}{\partial x_2} &= J_{tx_2}^* + g_{x_2} + J_{x_1x_2}^* a_1 + J_{x_1}^* a_{1x_2} + J_{x_2x_2}^* a_2 + J_{x_2}^* a_{2x_2} \\
&\quad + \cdots + J_{x_nx_2}^* a_n + J_{x_n}^* a_{nx_2} = 0 \\
&\vdots \\
\frac{\partial v}{\partial x_n} &= J_{tx_n}^* + g_{x_n} + J_{x_1x_n}^* a_1 + J_{x_1}^* a_{1x_n} + J_{x_2x_n}^* a_2 + J_{x_2}^* a_{2x_n} \\
&\quad + \cdots + J_{x_nx_n}^* a_n + J_{x_n}^* a_{nx_n} = 0.\dagger
\end{aligned} \tag{7.1-18}$$

To simplify the notation in these equations, let

$$\psi_i(\mathbf{x}^*(t), t) \triangleq J_{x_i}^*(\mathbf{x}^*(t), t), \qquad i = 1, 2, \ldots, n. \tag{7.1-19}$$

Using these definitions, and the property

† $J_{tx_i}^*$ denotes $\partial^2 J^*/\partial t \partial x_i$; g_{x_i} denotes $\partial g/\partial x_i$; $J_{x_i}^*$ denotes $\partial J^*/\partial x_i$; a_{ix_j} denotes $\partial a_i/\partial x_j$; $J_{x_ix_j}^*$ denotes $\partial^2 J^*/\partial x_i \partial x_j$.

$$\begin{aligned}\frac{\partial}{\partial x_i}\left[\frac{\partial J^*}{\partial t}\right] &= \frac{\partial}{\partial t}\left[\frac{\partial J^*}{\partial x_i}\right], \qquad i = 1, 2, \ldots, n \\ \frac{\partial}{\partial x_i}\left[\frac{\partial J^*}{\partial x_j}\right] &= \frac{\partial}{\partial x_j}\left[\frac{\partial J^*}{\partial x_i}\right], \qquad i = 1, 2, \ldots, n,\end{aligned} \tag{7.1-20}$$

which holds, assuming that the mixed partial derivatives are continuous,† we find that Eqs. (7.1-18) are

$$\begin{aligned}\frac{\partial v}{\partial x_1} &= \frac{\partial \psi_1}{\partial t} + g_{x_1} + \frac{\partial \psi_1}{\partial x_1}a_1 + \psi_1 a_{1x_1} + \frac{\partial \psi_1}{\partial x_2}a_2 + \psi_2 a_{2x_1} \\ &\quad + \cdots + \frac{\partial \psi_1}{\partial x_n}a_n + \psi_n a_{nx_1} = 0 \\ \frac{\partial v}{\partial x_2} &= \frac{\partial \psi_2}{\partial t} + g_{x_2} + \frac{\partial \psi_2}{\partial x_1}a_1 + \psi_1 a_{1x_2} + \frac{\partial \psi_2}{\partial x_2}a_2 + \psi_2 a_{2x_2} \\ &\quad + \cdots + \frac{\partial \psi_2}{\partial x_n}a_n + \psi_n a_{nx_2} = 0 \\ &\vdots \\ \frac{\partial v}{\partial x_n} &= \frac{\partial \psi_n}{\partial t} + g_{x_n} + \frac{\partial \psi_n}{\partial x_1}a_1 + \psi_1 a_{1x_n} + \frac{\partial \psi_n}{\partial x_2}a_2 + \psi_2 a_{2x_n} \\ &\quad + \cdots + \frac{\partial \psi_n}{\partial x_n}a_n + \psi_n a_{nx_n} = 0.\end{aligned} \tag{7.1-21}$$

Moving terms that do not involve partial derivatives of the ψ's to the right side gives, for the ith equation,

$$\begin{aligned}&\frac{\partial \psi_i}{\partial t} + \frac{\partial \psi_i}{\partial x_1}a_1 + \frac{\partial \psi_i}{\partial x_2}a_2 + \cdots + \frac{\partial \psi_i}{\partial x_n}a_n \\ &= -\psi_1 a_{1x_i} - \psi_2 a_{2x_i} - \cdots - \psi_n a_{nx_i} - g_{x_i}, \qquad i = 1, 2, \ldots, n.\end{aligned} \tag{7.1-22}$$

We know that the state equations

$$\frac{dx_i^*(t)}{dt} = a_i(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \qquad i = 1, 2, \ldots, n, \tag{7.1-23}$$

must be satisfied; hence, (7.1-22) can be written

$$\begin{aligned}&\frac{\partial \psi_i}{\partial t} + \frac{\partial \psi_i}{\partial x_1}\frac{dx_1^*}{dt} + \frac{\partial \psi_i}{\partial x_2}\frac{dx_2^*}{dt} + \cdots + \frac{\partial \psi_i}{\partial x_n}\frac{dx_n^*}{dt} \\ &= -\psi_1 a_{1x_i} - \psi_2 a_{2x_i} - \cdots - \psi_n a_{nx_i} - g_{x_i}, \qquad i = 1, 2, \ldots, n.\end{aligned} \tag{7.1-24}$$

† See [O-2], p. 367.

Since $\psi_i \triangleq \psi_i(x_1^*(t), x_2^*(t), \ldots, x_n^*(t), t)$, the left side is the total derivative of ψ_i with respect to time; therefore,

$$\begin{aligned}
\frac{d\psi_1}{dt} &= -[a_{1x_1}\psi_1 + a_{2x_1}\psi_2 + \cdots + a_{nx_1}\psi_n] - g_{x_1} \\
\frac{d\psi_2}{dt} &= -[a_{1x_2}\psi_1 + a_{2x_2}\psi_2 + \cdots + a_{nx_2}\psi_n] - g_{x_2} \\
&\cdot \\
&\cdot \\
&\cdot \\
\frac{d\psi_n}{dt} &= -[a_{1x_n}\psi_1 + a_{2x_n}\psi_2 + \cdots + a_{nx_n}\psi_n] - g_{x_n},
\end{aligned} \tag{7.1-25}$$

or, in matrix form,

$$\frac{d\boldsymbol{\psi}}{dt} = -\left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}\right]^T \boldsymbol{\psi} - \frac{\partial g}{\partial \mathbf{x}}. \tag{7.1-25a}$$

To obtain the boundary conditions for $\boldsymbol{\psi}$, we use Eqs. (7.1-5) and (7.1-19) with the result

$$\boldsymbol{\psi}(\mathbf{x}^*(t_f), t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f). \tag{7.1-26}$$

To summarize, we have shown that

$$\frac{d\boldsymbol{\psi}}{dt}(\mathbf{x}^*(t), t) = -\left[\frac{\partial \mathbf{a}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t)\right]^T \boldsymbol{\psi}(\mathbf{x}^*(t), t) - \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x}^*(t), \mathbf{u}^*(t), t), \tag{7.1-25b}$$

where the optimal control $\mathbf{u}^*(t)$ satisfies

$$\begin{aligned}
&g(\mathbf{x}^*(t), \mathbf{u}^*(t), t) + [\boldsymbol{\psi}(\mathbf{x}^*(t), t)]^T \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \\
&= \min_{\mathbf{u}(t)} \{g(\mathbf{x}^*(t), \mathbf{u}(t), t) + [\boldsymbol{\psi}(\mathbf{x}^*(t), t)]^T \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}(t), t)\}.
\end{aligned} \tag{7.1-4a}$$

In addition, the state equations

$$\dot{\mathbf{x}}^*(t) = \mathbf{a}(\mathbf{x}^*(t), \mathbf{u}^*(t), t) \tag{7.1-27}$$

and the boundary conditions

$$\boldsymbol{\psi}(\mathbf{x}^*(t_f), t_f) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}^*(t_f), t_f) \tag{7.1-26}$$

must be satisfied. Comparing Eqs. (7.1-27), (7.1-25b), (7.1-4a), and (7.1-26) with Eqs. (7.1-6a), (7.1-7a), (7.1-11a), and (7.1-10), respectively, we observe

the anticipated correspondence: $\psi(\mathbf{x}^*(t), t)$ and $\mathbf{p}^*(t)$ satisfy the same differential equation and have the same boundary conditions; therefore, we conclude that

$$\psi(\mathbf{x}^*(t), t) = \mathbf{p}^*(t). \tag{7.1-28}$$

Although we have written many equations, the key point in our argument is contained in Eqs. (7.1-15) and (7.1-16). Thereafter, the steps we took were primarily manipulative.

We have succeeded in deriving the minimum principle from the Hamilton-Jacobi-Bellman functional equation; however, it is important to keep in mind the restrictions imposed by the derivation. In writing Eq. (7.1-16) we assumed that the states are not constrained by any boundaries. In addition, Eq. (7.1-18) is valid only if the indicated second partials exist {note that this was assumed to be the case in our derivation of the H-J-B equation [see Eq. (3.11-7)]}.†

The derivation we have performed also provides a useful interpretation of the costate variables. Let $\delta J^*(\mathbf{x}^*(t), t, \delta\mathbf{x}(t))$ denote the first-order approximation to the change in the minimum value of the performance measure that results when the state value at time t deviates from $\mathbf{x}^*(t)$ by an amount $\delta\mathbf{x}(t)$; then, from (7.1-12),

$$\begin{aligned}\delta J^*(\mathbf{x}^*(t), t, \delta\mathbf{x}(t)) &= [J_{\mathbf{x}}^*(\mathbf{x}^*(t), t)]^T \delta\mathbf{x}(t) \\ &= \mathbf{p}^{*T}(t)\, \delta\mathbf{x}(t).\end{aligned} \tag{7.1-29}$$

In other words, the extremal costate is the sensitivity of the *minimum value* of the performance measure to changes in the state value.

7.2 SUMMARY

Having discussed the relationship between dynamic programming and Pontryagin's minimum principle, let us now review the important features of these two techniques.

The Minimum Principle

Applying the minimum principle, or the calculus of variations, to determine optimal controls generally leads to a nonlinear two-point boundary-

† We shall not pursue the required mathematical properties of J^* and its derivatives any further. For examples of difficulties that may occur if certain properties fail to be satisfied, and for a discussion of how to handle such difficulties, see [S-5] and [D-2].

value problem that requires the use of iterative numerical techniques for solution. As noted previously, these iterative algorithms determine optimal controls in open-loop form.

If the state equations of a process are linear (or have been linearized), and the performance measure is a quadratic form, the optimal control *law* can be determined by numerically integrating a matrix differential equation of the Riccati type.

An important feature of the variational approach is that the form of optimal controls can be determined; hence, it is necessary only to consider the subset of controls having the appropriate form; this is a significant conceptual and computational advantage.

Dynamic Programming

Dynamic programming is essentially a clever way of examining all of the candidates for an optimal control law. To do this by direct enumeration of all the possibilities is a horrendous task, but by using the principle of optimality a multiple-stage decision process can be reduced to a sequence of single-stage decision processes, and a feasible computational algorithm is obtained. The algorithm consists of solving the functional recurrence equation

$$J^*_{N-K,N}(\mathbf{x}(N-K)) = \min_{\mathbf{u}(N-K)} \{g_D(\mathbf{x}(N-K), \mathbf{u}(N-K)) + J^*_{N-(K-1),N}(\mathbf{a}_D(\mathbf{x}(N-K), \mathbf{u}(N-K)))\} \quad (3.7\text{-}18)$$

by a direct search among the admissible control values. The presence of state and control constraints generally complicates the application of variational techniques; however, in dynamic programming, state and control constraints reduce the range of values to be searched and thereby simplify the solution. Another desirable feature of the dynamic programming approach is that the computational procedure determines the optimal control law. Moreover, since the algorithm makes a direct comparison of the performance measure values associated with all optimal control law candidates, it is ensured that the global, or absolute, optimal control law is obtained. The primary limitation of the dynamic programming approach is the need for large storage capacity in the digital computer when solving problems involving high-order systems—this is the "curse of dimensionality."

The Complementary Use of Several Optimization Techniques

Although a particular problem may perhaps be solved by applying only one of the techniques we have discussed, it is often beneficial to use the com-

plementary features of several different approaches. For example, suppose that the minimum principle indicates that the only values assumed by an optimal control are $+1$, 0, or -1. This knowledge can be used in designating the control values to be tried in obtaining a dynamic programming solution; instead of trying a finite set of controls that satisfy $-1 \leq u \leq +1$, we need use only $u = +1, 0$, and -1 as trial control values.

As another example, suppose that it is desired to find an optimal control law for a system whose initial state value is known to be in a specified region of the state space. One approach is to determine an optimal trajectory by employing iterative numerical techniques, and then to make use of this trajectory to define a region of the state space in which an optimal control law can be obtained by dynamic programming. By doing this, only a subset of state space values is searched in the dynamic programming solution, and thus the requirements for computer memory and computation time are eased.

As a third example, suppose that the variational approach indicates that a singular interval may occur. Nonlinear or dynamic programming may be helpful in determining whether or not singular controls are optimal.

7.3 CONTROLLER DESIGN

In most applications engineers are required to design a controller, that is, a device for generating control signals from observations of system outputs. This being the case, three alternatives are

1. An on-line digital computer that calculates optimal control signals as the process evolves, and additional hardware to synthesize the control signals.
2. A special-purpose digital controller to synthesize an optimal control law that has been precomputed off-line with a general-purpose digital computer.
3. A suboptimal, but easily implemented, controller whose configuration and parameters have been precalculated with an off-line computer.

Let us consider the implications of each of these alternatives.

For many applications it may be difficult to justify economically the presence of an on-line digital computer. In addition, such a controller must necessarily be suboptimal because of the finite time required for computation. In fact, if the system states change too quickly for the computer to keep up, serious difficulties may result. Slowly changing chemical processes exemplify the types of problems that are well suited to on-line control computation.

The second alternative has several advantages. All computing is done off-line; hence, the general-purpose computer required is available for solving many problems rather than being devoted exclusively to one system (or a few systems on a time-shared basis). The special-purpose digital controller will be much smaller, less expensive, and not as complex as a general-purpose digital computer. In addition, since the optimal control law is precomputed, the question of calculations having to keep up with the changing state of the controlled process does not arise. On the debit side, the control computer may require a large amount of storage; however, this need not always be rapid-access storage—a small magnetic tape arrangement may be quite acceptable. Notice that in this control scheme the optimal control *law* must be calculated. Presumably, this is accomplished by using dynamic programming; an alternative approach that relies on linearization of the state-costate differential equations about a nominal optimal trajectory is discussed in reference [B-6].

The concept of suboptimal, but easily implemented, controllers is very attractive from a practical point of view. Naturally, the system's performance with a suboptimal controller should be compared with optimal performance; such a comparison could be the basis for deciding on the acceptability of a proposed suboptimal design. Rejection of a controller design indicates that either the controller configuration needs to be altered, or that the controller parameters must be adjusted. Efforts to achieve acceptable suboptimal designs have met with limited success so far. The principal difficulty is that a controller may be nearly optimal for some initial conditions, but very poor for others.† A suggested method of alleviating this difficulty is to minimize the deviation from optimal response that results when the system assumes its worst possible initial state.‡ This leads to a minimax solution of the problem. An alternative approach is to assume a probability distribution for the initial state values and minimize the expected value of the performance measure.*

Each of these alternatives generally requires that *all* system states be available for generating the control signal; however, it may be necessary to generate control signals using estimates of the state values obtained from noisy observations of system outputs.§ One method of obtaining state estimates (which are optimal in a statistical sense) is to use a Kalman filter.‖

† See [F-2], [R-7], and [S-6].
‡ See [A-4], [K-10], [O-3] and [S-7].
* See [K-11] and [K-12].
§ For a discussion of the use of optimal estimates and optimal controllers to yield optimal stochastic systems, see [L-6], pp. 131ff.
‖ See [K-13] and [K-14].

7.4 CONCLUSION

Optimal control theory has been used to obtain solutions to a variety of aerospace engineering problems and holds great promise for other problem areas as well; however, much remains to be accomplished. Hopefully, the reader has been stimulated to learn more about optimal control theory and its applications, and now has a firm foundation on which to build his knowledge.

REFERENCES

A-4 Althouse, T. S., "A Piecewise-Linear Switching Function for Quasi-Minimum-Time Control," unpublished M. S. thesis, Naval Postgraduate School, 1968.

B-6 Breakwell, J. V., J. L. Speyer, and A. E. Bryson, "Optimization and Control of Nonlinear Systems Using the Second Variation," *J. SIAM Control*, series A (1963), 193–223.

D-2 Dreyfus, S. E., *Dynamic Programming and the Calculus of Variations.* New York: Academic Press Inc., 1965.

F-2 Frederick, D. K., and G. F. Franklin, "Design of Piecewise-Linear Switching Functions for Relay Control Systems," *IEEE Trans. Automatic Control* (1967), 380–387.

K-10 Koivuniemi, A. J., "A Computational Technique for the Design of a Specific Optimal Controller," *IEEE Trans. Automatic Control* (1967), 180–183.

K-11 Kleinman, D. L., and M. Athans, "The Design of Suboptimal Linear Time-Varying Systems," *IEEE Trans. Automatic Control* (1968), 150–159.

K-12 Kleinman, D. L., T. Fortmann, and M. Athans, "On the Design of Linear Systems with Piecewise-Constant Feedback Gains," *IEEE Trans. Automatic Control* (1968), 354–361.

K-13 Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," *ASME Journal of Basic Engineering* (1960), 35–45.

K-14 Kalman, R. E., and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," *ASME Journal of Basic Engineering* (1961), 95–108.

L-6 Lee, R. C. K., *Optimal Estimation, Identification, and Control.* Cambridge, Mass: Massachusetts Institute of Technology Press, 1964.

O-2 Olmstead, J. M. H., *Real Variables.* New York: Appleton-Century-Crofts, Inc., 1959.

O-3 Ozer, E., and D. E. Kirk, "Constrained Optimal Controls for Linear Regulators with Incomplete State Feedback," to appear.

R-7 Rekasius, Z. V., "Suboptimal Design of Intentionally Nonlinear Controllers," *IEEE Trans. Automatic Control* (1964), 380–386.

S-5 Sagan, H., "Dynamic Programming and Pontryagin's Minimum Principle," NASA Contractor Report CR-838 (1967).

S-6 Smith, F. W., "Design of Quasi-Optimal Minimum-Time Controllers," *IEEE Trans. Automatic Control* (1966), 71–77.

S-7 Salmon, D. M., "Minimax Controller Design," *IEEE Trans. Automatic Control* (1968), 369–376.

APPENDIX 1

Useful Matrix Properties and Definitions

1. The transpose of the product of two matrices equals the product of the transposed matrices in reverse order; that is,

$$[\mathbf{CD}]^T = \mathbf{D}^T\mathbf{C}^T. \tag{A.1-1}$$

2. The transpose of a scalar equals itself. For example, if $\mathbf{z}^T\mathbf{My}$ is a scalar, then

$$\mathbf{z}^T\mathbf{My} = [\mathbf{z}^T\mathbf{My}]^T. \tag{A.1-2}$$

Using Property 1 twice, we have

$$[\mathbf{z}^T\mathbf{My}]^T = [\mathbf{My}]^T\mathbf{z} = \mathbf{y}^T\mathbf{M}^T\mathbf{z}. \tag{A.1-3}$$

3. *Definition.* Let $\mathbf{P}$ and $\mathbf{S}$ be real symmetric matrices. If

$$\mathbf{y}^T\mathbf{Py} > 0 \qquad \text{for all } \mathbf{y} \neq \mathbf{0}, \tag{A.1-4}$$

$\mathbf{P}$ is called a *positive-definite matrix*; if

$$\mathbf{y}^T\mathbf{Sy} \geq 0 \qquad \text{for all } \mathbf{y}, \tag{A.1-5}$$

$\mathbf{S}$ is called a *positive semi-definite* matrix.

4. The sum of a positive definite matrix and a positive semi-definite matrix

is positive definite. Consider

$$\mathbf{y}^T[\mathbf{P} + \mathbf{S}]\mathbf{y} = \mathbf{y}^T[\mathbf{P}\mathbf{y} + \mathbf{S}\mathbf{y}] = \mathbf{y}^T\mathbf{P}\mathbf{y} + \mathbf{y}^T\mathbf{S}\mathbf{y}. \tag{A.1-6}$$

From Definition 3, $\mathbf{y}^T\mathbf{P}\mathbf{y} > 0$, and $\mathbf{y}^T\mathbf{S}\mathbf{y} \geq 0$ for all $\mathbf{y} \neq \mathbf{0}$; therefore,

$$\mathbf{y}^T[\mathbf{P} + \mathbf{S}]\mathbf{y} > 0 \qquad \text{for all } \mathbf{y} \neq \mathbf{0}, \tag{A.1-7}$$

so the matrix $[\mathbf{P} + \mathbf{S}]$ is positive definite.

5. If a matrix is positive definite, its inverse exists.

6. Let $s(\mathbf{y})$ be a scalar function of $\mathbf{y} = [y_1 \ldots y_m]^T$.
Definition. The gradient of s with respect to $\mathbf{y}$ is defined as

$$\frac{\partial s}{\partial \mathbf{y}}(\mathbf{y}) \triangleq \begin{bmatrix} \frac{\partial s}{\partial y_1}(\mathbf{y}) \\ \frac{\partial s}{\partial y_2}(\mathbf{y}) \\ \vdots \\ \frac{\partial s}{\partial y_m}(\mathbf{y}) \end{bmatrix}$$

The following properties follow from the definition in 6.

7. Let $\mathbf{y}$ be an $m \times 1$ column matrix, $\mathbf{z}$ an $m \times 1$ column matrix, and $\mathbf{M}$ an $m \times m$ matrix. To determine $\partial[\mathbf{y}^T\mathbf{M}\mathbf{z}]/\partial\mathbf{y}$ the $m \times 1$ column matrix $\mathbf{M}\mathbf{z}$ is treated as a matrix of constants. Letting $\mathbf{c} \triangleq \mathbf{M}\mathbf{z}$, we have

$$\frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{c}] = \frac{\partial}{\partial \mathbf{y}}[y_1c_1 + y_2c_2 + \cdots + y_mc_m], \tag{A.1-8}$$

which from the definition in 6 becomes

$$\frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{c}] = \mathbf{c} \triangleq \mathbf{M}\mathbf{z}. \tag{A.1-9}$$

8. The gradient of the quadratic form $\mathbf{y}^T\mathbf{M}\mathbf{y}$ is found by using Property 7 and the well-known rule for differentiation of a product; that is

$$\frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{M}\mathbf{y}] = \frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{c}^{(1)}] + \frac{\partial}{\partial \mathbf{y}}[\mathbf{c}^{(2)T}\mathbf{y}], \tag{A.1-10}$$

where $\mathbf{c}^{(1)} \triangleq \mathbf{M}\mathbf{y}$, and $\mathbf{c}^{(2)T} \triangleq \mathbf{y}^T\mathbf{M}$ are treated as constant matrices. Since $\mathbf{c}^{(2)T}\mathbf{y}$ is a scalar, and the transpose of a scalar equals itself, we have

$$\frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{M}\mathbf{y}] = \frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{c}^{(1)}] + \frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{c}^{(2)}]. \tag{A.1-11}$$

From Property 7, then,

$$\begin{aligned}\frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{M}\mathbf{y}] &= \mathbf{c}^{(1)} + \mathbf{c}^{(2)} \\ &= \mathbf{M}\mathbf{y} + \mathbf{M}^T\mathbf{y}.\end{aligned} \tag{A.1-12}$$

If $\mathbf{M}$ is a symmetric matrix, then from (A.1-12) it follows that

$$\frac{\partial}{\partial \mathbf{y}}[\mathbf{y}^T\mathbf{M}\mathbf{y}] = 2\mathbf{M}\mathbf{y}. \tag{A.1-13}$$

9. *Definition.* If $\mathbf{a}$ is an $n \times 1$ matrix function of $\mathbf{y}$ (an $m \times 1$ matrix), and $\mathbf{z}$ (an $n \times 1$ matrix), then

$$\frac{\partial}{\partial \mathbf{y}}[\mathbf{a}(\mathbf{y}, \mathbf{z})] \triangleq \begin{bmatrix} \frac{\partial a_1}{\partial y_1}(\mathbf{y}, \mathbf{z}) & \frac{\partial a_1}{\partial y_2}(\mathbf{y}, \mathbf{z}) & \cdots & \frac{\partial a_1}{\partial y_m}(\mathbf{y}, \mathbf{z}) \\ \frac{\partial a_2}{\partial y_1}(\mathbf{y}, \mathbf{z}) & \frac{\partial a_2}{\partial y_2}(\mathbf{y}, \mathbf{z}) & \cdots & \frac{\partial a_2}{\partial y_m}(\mathbf{y}, \mathbf{z}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial a_n}{\partial y_1}(\mathbf{y}, \mathbf{z}) & \frac{\partial a_n}{\partial y_2}(\mathbf{y}, \mathbf{z}) & \cdots & \frac{\partial a_n}{\partial y_m}(\mathbf{y}, \mathbf{z}) \end{bmatrix}$$

10. *Definition.* If s is a scalar function of an $m \times 1$ matrix $\mathbf{y}$, then

$$\frac{\partial^2 s}{\partial \mathbf{y}^2}(\mathbf{y}) \triangleq \begin{bmatrix} \frac{\partial^2 s}{\partial y_1^2}(\mathbf{y}) & \frac{\partial^2 s}{\partial y_1 \partial y_2}(\mathbf{y}) & \cdots & \frac{\partial^2 s}{\partial y_1 \partial y_m}(\mathbf{y}) \\ \frac{\partial^2 s}{\partial y_2 \partial y_1}(\mathbf{y}) & \frac{\partial^2 s}{\partial y_2^2}(\mathbf{y}) & \cdots & \frac{\partial^2 s}{\partial y_2 \partial y_m}(\mathbf{y}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 s}{\partial y_m \partial y_1}(\mathbf{y}) & \frac{\partial^2 s}{\partial y_m \partial y_2}(\mathbf{y}) & \cdots & \frac{\partial^2 s}{\partial y_m^2}(\mathbf{y}) \end{bmatrix}$$

Notice that if the order of differentiation is interchangeable, $\partial^2 s/\partial \mathbf{y}^2$ is symmetric.

11. From Definition 10, if $\mathbf{R}$ is a real symmetric $m \times$ m matrix and $\mathbf{u}$ is an $m \times 1$ matrix, then

$$\frac{\partial^2}{\partial \mathbf{u}^2}[\mathbf{u}^T\mathbf{R}\mathbf{u}] = 2\mathbf{R}. \tag{A.1-14}$$

APPENDIX 2

Difference Equation Representation of Linear Sampled-Data Systems

Consider a linear time-invariant system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t). \tag{A.2-1}$$

If the control is generated by a sample-and-zero-order-hold element, as shown in Fig. A-1, then $\mathbf{u}$ is a piecewise-constant signal. Assuming that the sampling rate is uniform and has a period T, we have

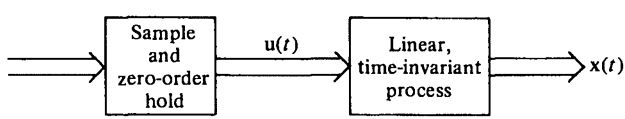


Figure A-1

$$\mathbf{u}(t) = \mathbf{u}(kT), \qquad kT \leq t < (k+1)T. \tag{A.2-2}$$

The solution of the state equations given in (A.2-1) is

$$\mathbf{x}(t) = \boldsymbol{\varphi}(t - t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \boldsymbol{\varphi}(t - \tau)\mathbf{B}\mathbf{u}(\tau)\, d\tau. \tag{A.2-3}$$

If the states are observed only at the sampling instants, then, letting $t_0 \triangleq kT$, and $t \triangleq (k+1)T$, we have

$$\mathbf{x}([k+1]T) = \boldsymbol{\varphi}(T)\mathbf{x}(kT) + \int_{kT}^{(k+1)T} \boldsymbol{\varphi}([k+1]T - \tau)\mathbf{B}\mathbf{u}(\tau)\, d\tau. \tag{A.2-4}$$

Since the control is constant during the interval $kT \leq t < (k+1)T$,

$$\mathbf{x}([k+1]T) = \boldsymbol{\varphi}(T)\mathbf{x}(kT) + \left[\int_{kT}^{(k+1)T} \boldsymbol{\varphi}([k+1]T - \tau)\mathbf{B}\, d\tau\right]\mathbf{u}(kT). \quad \text{(A.2-5)}$$

It can be verified that the integral has the same value for all k; thus,

$$\mathbf{x}([k+1]T) = \boldsymbol{\varphi}(T)\mathbf{x}(kT) + \left[\int_{0}^{T} \boldsymbol{\varphi}(T - \tau)\mathbf{B}\, d\tau\right]\mathbf{u}(kT). \quad \text{(A.2-6)}$$

Omitting the argument T, and defining the integral as the $n \times m$ matrix $\boldsymbol{\Delta}$, we obtain

$$\mathbf{x}(k+1) = \boldsymbol{\varphi}\mathbf{x}(k) + \boldsymbol{\Delta}\mathbf{u}(k), \quad \text{(A.2-7)}$$

a set of n first-order, linear, *difference equations*. The matrices $\boldsymbol{\varphi}$ and $\boldsymbol{\Delta}$ contain only constants (which depend on the value of T). If the process equations are time-varying, that is,

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \quad \text{(A.2-8)}$$

the linear difference equations

$$\mathbf{x}(k+1) = \boldsymbol{\varphi}(k)\mathbf{x}(k) + \boldsymbol{\Delta}(k)\mathbf{u}(k) \quad \text{(A.2-9)}$$

can be derived by following a similar procedure. Notice that $\boldsymbol{\varphi}$ and $\boldsymbol{\Delta}$ will be functions of k, however.

APPENDIX 3

Special Types of Euler Equations

CASE 1: g depending only on $\dot{x}(t)$.

The Euler equation

$$\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)\right] = 0 \tag{A.3-1}$$

reduces to

$$\frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(\dot{x}^*(t))\right] = 0, \tag{A.3-2}$$

or

$$\left[\frac{\partial^2 g}{\partial \dot{x}^2}(\dot{x}^*(t))\right]\ddot{x}^*(t) = 0; \tag{A.3-3}$$

thus, either $\ddot{x}^*(t) = 0$ or $\partial^2 g/\partial \dot{x}^2 = 0$. If $\ddot{x}^*(t) = 0$, $x^*(t) = c_1 t + c_2$, where c_1 and c_2 are constants of integration. If $\partial^2 g/\partial \dot{x}^2 = 0$ has a real root, that is, $\dot{x}^*(t) = c_3$, then we have $x^*(t) = c_3 t + c_4$. Consequently, if g depends only on $\dot{x}(t)$, the solution to the Euler equation is a family of straight lines.

CASE 2: g depending only on t and $\dot{x}(t)$.

Integrating both sides of the Euler equation

$$\frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(\dot{x}^*(t), t)\right] = 0 \tag{A.3-4}$$

gives

$$\frac{\partial g}{\partial \dot{x}}(\dot{x}^*(t), t) = c_1, \tag{A.3-5}$$

a first-order differential equation that does not involve $x^*(t)$. $x^*(t)$ can then be obtained by solving (A.3-5) for $\dot{x}^*(t)$ and integrating.

CASE 3: g depending on $x(t)$ and $\dot{x}(t)$.

The Euler equation is

$$\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t)) - \frac{d}{dt}\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t))\right], \tag{A.3-6}$$

or, if the chain rule is used,

$$\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t)) - \left[\frac{\partial^2 g}{\partial \dot{x}\partial x}(x^*(t), \dot{x}^*(t))\right]\dot{x}^*(t) - \left[\frac{\partial^2 g}{\partial \dot{x}^2}(x^*(t), \dot{x}^*(t))\right]\ddot{x}^*(t) = 0. \tag{A.3-7}$$

Multiplying this equation by $\dot{x}^*(t)$ makes the left side the exact derivative

$$\frac{d}{dt}\left[g(x^*(t), \dot{x}^*(t)) - \dot{x}^*(t)\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t))\right],$$

which implies that

$$g(x^*(t), \dot{x}^*(t)) - \dot{x}^*(t)\left[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t))\right] = c_1. \tag{A.3-8}$$

Since t does not appear explicitly, $x^*(t)$ can be obtained by solving for $\dot{x}^*(t)$ and using separation of variables.

CASE 4: g depending only on $x(t)$ and t.

In this case, the Euler equation becomes

$$\frac{\partial g}{\partial x}(x^*(t), t) = 0, \tag{A.3-9}$$

a nonlinear *algebraic* equation. Thus, there will be no constants of integration and there will be an extremal only if the curve that satisfies (A.3-9) passes through the specified boundary points.

CASE 5: g depending linearly on $\dot{x}(t)$.

We have the form

$$g(x(t), \dot{x}(t), t) = M(x(t), t) + [N(x(t), t)]\dot{x}(t). \tag{A.3-10}$$

The Euler equation reduces to

$$\frac{\partial M}{\partial x}(x^*(t), t) - \frac{\partial N}{\partial t}(x^*(t), t) = 0, \tag{A.3-11}$$

which is, as in Case 4, an algebraic equation. Thus, an extremal will exist only if the curve(s) that satisfies (satisfy) (A.3-11) happens to pass through the specified boundary points.

APPENDIX 4

Answers to Selected Problems

CHAPTER 1

1-1. (a) $dq(t)/dt = -.16q(t)$; $dp(t)/dt = .16q(t) - .16p(t)$

(c) $\varphi_{11}(t) = \epsilon^{-.16t}$; $\varphi_{12}(t) = 0.$; $\varphi_{21}(t) = .16t\epsilon^{-.16t}$; $\varphi_{22}(t) = \epsilon^{-.16t}$

(d) $q(t) = 60\epsilon^{-.16t}$; $p(t) = 9.6t\epsilon^{-.16t}$, $t \geq 0$.

1-3. (a) $dy(t)/dt = \dot{y}(t)$; $d\dot{y}(t)/dt = -Ky(t)/M - B\dot{y}(t)/M + f(t)/M$

(c) $\varphi_{11}(t) = \sqrt{2}\epsilon^{-t}\cos(t - \pi/4)$; $\varphi_{12}(t) = \epsilon^{-t}\sin t$; $\varphi_{21}(t) = -2\epsilon^{-t}\sin t$; $\varphi_{22}(t) = \sqrt{2}\epsilon^{-t}\cos(t + \pi/4)$

(d) $y(t) = .2\sqrt{2}\epsilon^{-t}\cos(t - \pi/4) + \epsilon^{-2t} + \sqrt{2}\epsilon^{-t}\cos(t - 3\pi/4)$

$\dot{y}(t) = -.4\epsilon^{-t}\sin t - 2\epsilon^{-2t} + 2\epsilon^{-t}\cos t.$

1-5. $d\theta(t)/dt = \dot{\theta}(t)$; $d\dot{\theta}(t)/dt = -K\theta(t)/I - B\dot{\theta}(t)/I + \lambda(t)/I.$

1-7. $di_f(t)/dt = -R_f i_f(t)/L_f + K_a e(t)/L_f$; $d\omega(t)/dt = K_i i_f(t)/I - B\omega(t)/I.$

1-9. $dv_c(t)/dt = i_{L_1}(t)/C$; $di_{L_1}(t)/dt = -L_2 v_c(t)/k^2 - R_1 L_2 i_{L_1}(t)/k^2 + MR_2 i_{L_2}(t)/k^2 + L_2 e(t)/k^2$; $di_{L_2}(t)/dt = Mv_c(t)/k^2 + MR_1 i_{L_1}(t)/k^2 - [R_2/L_2 + M^2 R_2/L_2 k^2] i_{L_2}(t) - Me(t)/k^2$; $k^2 \triangleq L_1 L_2 - M^2$.

1-12. (a) $\dot{x}_1(t) = x_2(t)$; $\dot{x}_2(t) = -x_2(t) + 5u(t)$; $y(t) = x_1(t)$

(c) $\dot{x}_1(t) = x_2(t)$; $\dot{x}_2(t) = x_3(t)$; $\dot{x}_3(t) = -3x_1(t) - 6x_2(t) - 5x_3(t) + 10u(t)$; $y(t) = x_1(t)$

(e) $\dot{x}_1(t) = x_2(t)$; $\dot{x}_2(t) = -x_2(t) + 5u(t)$; $y(t) = 2x_1(t) + x_2(t)$

(g) $\dot{x}_1(t) = x_2(t)$; $\dot{x}_2(t) = x_3(t)$; $\dot{x}_3(t) = -3x_1(t) - 6x_2(t) - 5x_3(t) + 10u(t)$; $y(t) = 3x_1(t) + 2x_2(t) + x_3(t)$

(i) $\dot{x}_1(t) = u(t)$; $\dot{x}_2(t) = -x_2(t) + u(t)$; $\dot{x}_3(t) = -2x_3(t) + u(t)$; $y(t) = 6x_1(t) - 6x_2(t) + x_3(t)$.

NOTE: There are other correct answers for different selections of the state variables.

1-14. Key: C ≙ controllable, NC ≙ not controllable, O ≙ observable, NO ≙ not observable

(a) C, O; (b) C, NO; (c) NC, O; (d) C, O; (e) NC, O; (f) C, NO; (g) C, O.

1-15. C if $b_i \neq 0$, $i = 1, 2, 3, 4$; O if $c_i \neq 0$, $i = 1, 2, 3, 4$.

CHAPTER 2

2-1. (a) $J = \int_0^{1 \text{ day}} [v_2(t) - M]^2 \, dt$

(b) $0 \leq h_1(t) \leq H_{1 \max}$, $0 \leq h_2(t) \leq H_{2 \max}$, $0 \leq v_1(t) \leq V_{1 \max}$, $0 \leq v_2(t) \leq V_{2 \max}$; $0 \leq w_1(t) \leq W_{1 \max}$, $0 \leq w_2(t) \leq W_{2 \max}$, $0 \leq m(t) \leq M_{\max}$.

2-3. (a) See 1-7 with $K_a = 1$; add $-\lambda_L(t)/I$ to right side of $d\omega(t)/dt$ equation.

(b) $|i_f(t)| \leq I_{f \max}$, $|\omega(t)| \leq \Omega_{\max}$, $|e(t)| \leq E_{\max}$, $|\lambda_L(t)| \leq \lambda_{\max}$.

(c) (i) $J = \int_0^{t_f} [[k\omega(t) - 5]^2 + \mu e^2(t)] \, dt$, k is a constant

(ii) $J = \int_0^{t_f} [[k\omega(t) - 5]^2 + \mu e(t) i_f(t)] \, dt$, μ is a weighting factor.

2-4. (a) $|u(t)| \leq U_{\max}$, $14.9° \leq \theta(30) \leq 15.1°$

(b) $J = \int_0^{30} |u(t)| \, dt$.

2-6. (a) $0 \leq x_1(t)$, $M_{\min} \leq x_5(t) \leq M_{\max}$, $0 \leq u_1(t) \leq T_{\max}$, $-\pi \leq u_2(t) \leq \pi$

(b) $x_3(t_f) = 3.0$, $J = -x_1(t_f)$

(c) $x_1(2.5) = 500$, $x_3(2.5) = 3.0$, $J = \int_0^{2.5} u_1(t) \, dt$, or $J = -x_5(t_f)$.

CHAPTER 3

3-1. (a) $x_1(k+1) = x_1(k) + .01x_2(k)$; $x_2(k+1) = -.01x_1(k) + [1 + .01[1 - x_1^2(k)]]x_2(k) + .01u(k)$; $J = [x_1(N) - 5]^2 + .01 \sum_{k=0}^{N-1} [x_2^2(k) + 20[x_1(k) - 5]^2 + u^2(k)]$, $N = 10/.01 = 1000$

(b) No computational adjustments required.

3-3. (a)

$k = 1$			$k = 0$		
$x(k)$	$u^*(x(k), k)$		$x(k)$	$u^*(x(k), k)$	
3.	None		3.	−0.5	
2.	−1.	;	2.	0.	
1.	−0.5		1.	−0.5	
0.	0.		0.	0.	

(b) $x(0) = -2 \longrightarrow u^*(0) = 0 \longrightarrow x(1) = -1. \longrightarrow u^*(1) = .5 \longrightarrow x(2) = 0$

These answers assume no control interpolation, i.e., the quantized values are the only admissible ones.

3-5. (a)

$k = 1$			$k = 0$		
$x(k)$	$u^*(x(k), k)$		$x(k)$	$u^*(x(k), k)$	
6.	None		6.	−0.5	
4.	−1.	;	4.	0., −0.5	
2.	0.		2.	0.	
0.	0.		0.	0.	

(b) $x(0) = 6. \longrightarrow u^*(0) = -.5 \longrightarrow x(1) = 4. \longrightarrow u^*(1) = -1. \longrightarrow x(2) = 2.$; $J^*(6.) = 1.25$

3-7.

$$u^*(\mathbf{x}(t), t) = \begin{cases} -1 & \text{for} \quad J^*_{x_2} > 1 \\ -J^*_{x_2} & \text{for} \quad |J^*_{x_2}| \leq 1 \\ +1 & \text{for} \quad J^*_{x_2} < -1 \end{cases}$$

3-11. One way is to define a new state, $\dot{x}_{n+1}(t) = u^2(t)$ and add the constraint $0 \leq x_{n+1}(t) \leq M$.

3-12. (a) $c_{ij}^{(k+1)} = \min_{\substack{l \\ l \neq i}} \{c_{il}^{(0)} + c_{lj}^{(k)}\}$

(b)

$$\mathbf{C}^{(1)} = \begin{bmatrix} 0 & 1 & 5 & 4 & 2 \\ & 0 & 5 & 3 & 3 \\ & & 0 & 2 & 6 \\ & & & 0 & 4 \\ & & & & 0 \end{bmatrix} = \mathbf{C}^{(2)} \text{ symmetric}$$

(c) No optimal path among five nodes can contain more than three intermediate nodes.

(d) $c_{ij}^{(3)} \leq c_{ij}^{(2)} \leq c_{ij}^{(1)} \leq c_{ij}^{(0)}$ for all i, j

(e) Must compute lower diagonal part of $\mathbf{C}^{(k+1)}$ as well as upper diagonal part.

CHAPTER 4

4-3. (a) $df = [12t^2 - 5/t^2]\Delta t$

(b) $df = [10q_1 + 6q_2]\Delta q_1 + [6q_1 + 4q_2]\Delta q_2$

(c) $df = [2q_1 + 5q_2q_3 + 2q_2]\Delta q_1 + [2q_2 + 5q_1q_3 + 2q_1]\Delta q_2 + [5q_1q_2 + 3]\Delta q_3$.

4-4. (a) $\delta J = \int_{t_0}^{t_f} [[3x^2(t) - 2x(t)\dot{x}(t)]\, \delta x(t) - x^2(t)\delta\dot{x}(t)]\, dt$

(b) $\delta J = \int_{t_0}^{t_f} [[2x_1(t) + x_2(t)]\delta x_1(t) + [x_1(t) + 2x_2(t)]\delta x_2(t) + [2\dot{x}_2(t)]\delta\dot{x}_1(t) + [2\dot{x}_1(t)]\delta\dot{x}_2(t)]\, dt$

(c) $\delta J = \int_{t_0}^{t_f} \epsilon^{x(t)}\delta x(t)\, dt.$

4-7. (a) $t^* = -1, -2$; (b) $t^* = 1/2$; (c) $q_1^* = -37/7, q_2^* = 11/7$.

4-8. (a) $x^*(t) = [\epsilon^{-t} - \epsilon^t]/[\epsilon^{-1} - \epsilon^1]$

(b) $x^*(t) = c_1\epsilon^t + c_2\epsilon^{-t}$, $c_1 = [-3 - \epsilon^{-2}]/[\epsilon^2 - \epsilon^{-2}]$, $c_2 = [\epsilon^2 + 3]/[\epsilon^2 - \epsilon^{-2}]$

(c) $x_1^*(t) = x_2^*(t) = \sinh(t)/\sinh(\pi/2)$.

4-9. $x^*(t) = c_1\epsilon^{2t} + c_2\epsilon^{-2t} + c_3$; $c_2 = [5 - 2\epsilon^2]/[\epsilon^{-2} - \epsilon^2]$, $c_1 = 2 - c_2$, $c_3 = -1$.

4-10. (a) $x^*(t) = [\epsilon^1\epsilon^{-t} + \epsilon^{-1}\epsilon^t]/[\epsilon^{-1} + \epsilon^1]$

(b) $x^*(t) = t^2/2 - 3t/2 + 1/2$

(c) $x_1^*(t) = -x_2^*(t) = -\sin(t)$.

4-12. (a) $x^*(t) = t^3 + t$

(b) $x^*(t) = \epsilon^{-t} + 3t\epsilon^{-t}$.

4-13. $x^*(t) = t/4 + 2$.

4-14. $x^{*2}(t) = -t^2 + 10t$.

4-15. $x^*(t) = -t + 5$.

4-16. $x^*(t) = \pm\sqrt{8t - t^2}$.

4-18. $x^*(t) = \begin{cases} 1.039t + 2.077, & t \in [-2.0, -0.0696] \\ -1.874t + 1.874, & t \in [-0.0696, 1.]. \end{cases}$

4-20. $x^*(t) = \begin{cases} -t & \text{for } 0 \le t \le 1 \\ t - 2 & \text{for } 1 \le t \le 4 \end{cases}$; or $x^*(t) = \begin{cases} t & \text{for } 0 \le t \le 3 \\ -t + 6 & \text{for } 3 \le t \le 4. \end{cases}$

4-21. $y_1^* = \pm 2$, $y_2^* = -1/2$.

4-22. $\mathbf{y}^* = [2, 2, 1]^T$ global min.; $\mathbf{y}^* = [-1, -1, 7]^T$ is a local max. or local min.

4-24. (a) $\dot{p}_1^*(t) = -[2w_1^*(t) + w_2^*(t) - p_2^*(t) - 2w_1^*(t)w_2^*(t)p_2^*(t)]$;

$\dot{p}_2^*(t) = -[w_1^*(t) + 2w_2^*(t) + p_1^*(t) + p_2^*(t)[1 - w_1^{*2}(t)]]$;

$0 = 2w_3^*(t) + p_2^*(t)$

(b) $\dot{p}_1^*(t) = 0$; $\dot{p}_2^*(t) = -p_1^*(t)$; $2w_3^*(t) + p_2^*(t) = 0$

(c) $\dot{p}_1^*(t) = 0$; $\dot{p}_2^*(t) = -p_1^*(t) + 2p_2^*(t)|w_2^*(t)|$; $0 = 2w_3^*(t) + p_2^*(t)$.

4-25. $x^*(t) = \pm 2 \sin(n\pi t)$, $n = 1, 2, 3, \ldots$.

CHAPTER 5

5-2. (a) $\dot{p}_1^*(t) = -[2x_1^*(t) - p_2^*(t) - 2p_2^*(t)x_2^*(t)x_1^*(t)]$;

$\dot{p}_2^*(t) = -[x_2^*(t) + p_1^*(t) + p_2^*(t)[1 - x_1^{*2}(t)]]$

(b) (i) $u^*(t) = -p_2^*(t)$

(ii) $u^*(t) = \begin{cases} -1 & \text{for } p_2^*(t) > 1 \\ -p_2^*(t) & \text{for } |p_2^*(t)| \le 1 \\ +1 & \text{for } p_2^*(t) < -1 \end{cases}$

5-4. (a) $\dot{p}_1^*(t) = p_2^*(t) + 2x_1^*(t)x_2^*(t)p_2^*(t)$; $\dot{p}_2^*(t) = -p_1^*(t) + [x_1^{*2}(t) - 1]p_2^*(t)$

(b)

$$u^*(t) = \begin{cases} -1 & \text{for } 1 < p_2^*(t) \\ 0 & \text{for } -1 < p_2^*(t) < 1 \\ +1 & \text{for } p_2^*(t) < -1 \\ \text{undetermined} & \text{for } p_2^*(t) = \pm 1 \end{cases}$$

(c) $[x_1^*(t_f) - 4]^2 + [x_2^*(t_f) - 5]^2 + [t_f - 2]^2 = 9$

$-\mathbf{p}^*(t_f) = d[2[x_1^*(t_f) - 4], 2[x_2^*(t_f) - 5]]^T$

$\mathscr{H}(\mathbf{x}^*(t_f), u^*(t_f), \mathbf{p}^*(t_f)) = 2d[t_f - 2]$.

5-6. (a) $u^*(t) = 3[\epsilon^{-2[T-t]} - \epsilon^{2[T-t]}]x^*(t)/[3\epsilon^{-2[T-t]} + \epsilon^{2[T-t]}]$

(b) $u^*(t) = -3x^*(t)$.

5-7. (a) $x^*(t) = x(0)[\epsilon^{-a[T-t]} - \epsilon^{a[T-t]}]/[\epsilon^{-aT} - \epsilon^{aT}]$

(b) $u^*(t) = 2ax(0)\epsilon^{-a[T-t]}/[\epsilon^{-aT} - \epsilon^{aT}]$

(c) $F(t, T, a) = 2a\epsilon^{-a[T-t]}/[\epsilon^{-a[T-t]} - \epsilon^{a[T-t]}]$

(d) As $t \longrightarrow T$: $F \longrightarrow \infty$, $u^*(T) \longrightarrow 2ax(0)/[\epsilon^{-aT} - \epsilon^{aT}]$

As $T \longrightarrow \infty$: $F \longrightarrow 0$, $u^*(t) \longrightarrow 0$ for all $t \in [t_0, t_f)$.

5-11. (a) $u^*(t) = [3\alpha_1(t) + \alpha_2(t)]x(t)/[\alpha_1(t) - \alpha_2(t)]$; $\alpha_1(t) \triangleq \epsilon^{-2[1-t]}$, $\alpha_2(t) \triangleq \epsilon^{2[1-t]}$

(b) $u^*(t) = 3[\alpha_1(t) - \alpha_2(t)]x(t)/[\alpha_1(t) + 3\alpha_2(t)]$; $\alpha_1(t)$, $\alpha_2(t)$ as in part (a).

5-12. (a) $u^*(t) = 2.6 + .04\epsilon^t$; (b) 7.42.

5-13. (a) $u^*(t) = M$, $t \in [0, 100]$, for all $x(t)$

(b) $u^*(t) = \begin{cases} M, & t \in [0, K/M] \\ 0, & t \in (K/M, 100] \end{cases}$ for all $x(t)$

(c) $u^*(t) = \begin{cases} 0, & t \in [0, 100 - K/M) \\ M, & t \in [100 - K/M, 100] \end{cases}$ for all $x(t)$.

5-14. (a) Pole at $s = -\sqrt{a^2 + [q/r]}$

(b) poles of closed-loop system at -2, -3.

5-15. $u^*(t) = -i(t) - \omega(t)$.

5-21. Switching curve described by $x_1(t) = x_2^2(t)/2$ for $x_2(t) \leq 2$, $x_1(t) = -x_2^2(t)/2 + 4$ for $x_2(t) \geq 2$; to the right of the switching curve $u^* = -1$, to the left $u^* = +1$.

5-22. Switching curve $x_2(t) = \mp 1 \pm [x_1(t) \pm 1]^{a_2/a_1}$; for part (a) $x_2(t) = \mp 1 \pm [x_1(t) \pm 1]^2$

5-24. (a) The initial states satisfying $-0.5 < x_0 < 0.5$ can be transferred to the origin.

(b) An optimal control does not exist if $|x_{i_0}| \geq |b_i|/a_i$, for any $i = 1, 2, \ldots, n$. This implies that an optimal control exists for all initial states if $a_i \leq 0$ for all i.

5-27. (a) The state equations (5.1-80) have the same form but with $M = x_5(t)$, $u(t) = u_1(t)$ and $T = -ku_2(t)$; in addition, there is the fifth state equation $\dot{x}_5(t) = u_2(t)$.

(b) $\dot{p}_1^*(t)$ through $\dot{p}_4^*(t)$ are as before; $\dot{p}_5^*(t) = -ku_2^*(t)[p_3^*(t) \sin u_1^*(t) + p_4^*(t) \cos u_1^*(t)]/x_5^{*2}(t)$

(c) $s \triangleq -k[p_3^*(t) \sin u_1^*(t) + p_4^*(t) \cos u_1^*(t)]/x_5^*(t) + p_5^*(t)$

$$u_2^*(t) = \begin{cases} -\eta & \text{for } s > 0 \\ 0 & \text{for } s < 0 \\ \text{undetermined for } s = 0 \end{cases};$$

Minimization of $\mathscr{H}$ with respect to u_1 yields same results as in Example 5.1-2.

(d) Boundary conditions same as in Example 5.1-2, except $p_5^*(t_f) = 0$ because $x_5(t_f) = M(t_f)$ is unspecified.

5-35. (a) $\dot{p}_1^*(t) = 0$; $\dot{p}_2^*(t) = -p_1^*(t) + \alpha\{p_2^*(t)[2x_2^{*2}(t) + x_4^{*2}(t)] + p_4^*(t)x_2^*(t)x_4^*(t)\}/[x_2^{*2}(t) + x_4^{*2}(t)]^{1/2}$, etc.

(b) $x_1^*(t_f) = e_1$; $p_2^*(t_f) = 0$; $x_3^*(t_f) = e_3$; $p_4^*(t_f) = 0$; $\mathscr{H}(\mathbf{x}^*(t_f), \mathbf{u}^*(t_f), \mathbf{p}^*(t_f), t_f) = 0$

(c) $u_1^*(t) = -p_2^*(t)/[p_2^{*2}(t) + p_4^{*2}(t)]^{1/2}$; $u_2^*(t) = -p_4^*(t)/[p_2^{*2}(t) + p_4^{*2}(t)]^{1/2}$

(d) $\mathbf{u}^*(t) = \begin{cases} [-1 \quad 0]^T & \text{for } |p_4^*(t)/p_2^*(t)| < 1, p_2^*(t) > 0 \\ [1 \quad 0]^T & \text{for } |p_4^*(t)/p_2^*(t)| < 1, p_2^*(t) < 0 \\ [0 \quad -1]^T & \text{for } |p_4^*(t)/p_2^*(t)| > 1, p_4^*(t) > 0 \end{cases}$

$[0 \quad 1]^T$ for $|p_4^*(t)/p_2^*(t)| > 1, p_4^*(t) < 0$

(e) Part (c) controls.

CHAPTER 6

6-13. $y_1^* = 4/3, y_2^* = 2/3, f(\mathbf{y}^*) = 8/3.$

6-14. $y_1^* = 1/3, y_2^* = 1/6, y_3^* = 1/2, f(\mathbf{y}^*) = 2.167.$

6-15. $y_1^* = 20, y_2^* = 1.0, f(\mathbf{y}^*) = 20.$

6-16. $y_1^* = 0, y_2^* = 1/2, y_3^* = 0, y_4^* = 1/4, y_5^* = 1/4, y_6^* = 0, y_7^* = 1/6, y_8^* = 0, y_9^* = 1, f(\mathbf{y}^*) = 7/48.$

6-17. $y_1^* = 0.7907, y_2^* = 1.2585, f(\mathbf{y}^*) = 2.2147.$

6-18. $y_1^* = 1.2982, y_2^* = 2.0772, f(\mathbf{y}^*) = 48.7831.$

6-19–6-22. Using 100 intervals for variational solution: $x^*(1.0) = 4.9389$, $u^*(0.0) = -6.7580 = -p^*(0.0), J^* = 13.5160.$

6-23–6-26. Using 100 intervals for variational solution: $x^*(1.0) = 0.7892$, $u^*(0.0) = -0.5806 = -p^*(0.0), J^* = 1.1613.$

Index

N

O

P

A CATALOG OF SELECTED

DOVER BOOKS

IN SCIENCE AND MATHEMATICS

Astronomy

BURNHAM'S CELESTIAL HANDBOOK, Robert Burnham, Jr. Thorough guide to the stars beyond our solar system. Exhaustive treatment. Alphabetical by constellation: Andromeda to Cetus in Vol. 1; Chamaeleon to Orion in Vol. 2; and Pavo to Vulpecula in Vol. 3. Hundreds of illustrations. Index in Vol. 3. 2,000pp. 6⅛ x 9¼.
Vol. I: 23567-X
Vol. II: 23568-8
Vol. III: 23673-0

EXPLORING THE MOON THROUGH BINOCULARS AND SMALL TELESCOPES, Ernest H. Cherrington, Jr. Informative, profusely illustrated guide to locating and identifying craters, rills, seas, mountains, other lunar features. Newly revised and updated with special section of new photos. Over 100 photos and diagrams. 240pp. 8¼ x 11. 24491-1

THE EXTRATERRESTRIAL LIFE DEBATE, 1750–1900, Michael J. Crowe. First detailed, scholarly study in English of the many ideas that developed from 1750 to 1900 regarding the existence of intelligent extraterrestrial life. Examines ideas of Kant, Herschel, Voltaire, Percival Lowell, many other scientists and thinkers. 16 illustrations. 704pp. 5⅜ x 8½. 40675-X

THEORIES OF THE WORLD FROM ANTIQUITY TO THE COPERNICAN REVOLUTION, Michael J. Crowe. Newly revised edition of an accessible, enlightening book recreates the change from an earth-centered to a sun-centered conception of the solar system. 242pp. 5⅜ x 8½. 41444-2

A HISTORY OF ASTRONOMY, A. Pannekoek. Well-balanced, carefully reasoned study covers such topics as Ptolemaic theory, work of Copernicus, Kepler, Newton, Eddington's work on stars, much more. Illustrated. References. 521pp. 5⅜ x 8½. 65994-1

A COMPLETE MANUAL OF AMATEUR ASTRONOMY: Tools and Techniques for Astronomical Observations, P. Clay Sherrod with Thomas L. Koed. Concise, highly readable book discusses: selecting, setting up and maintaining a telescope; amateur studies of the sun; lunar topography and occultations; observations of Mars, Jupiter, Saturn, the minor planets and the stars; an introduction to photoelectric photometry; more. 1981 ed. 124 figures. 26 halftones. 37 tables. 335pp. 6½ x 9¼. 42820-6

AMATEUR ASTRONOMER'S HANDBOOK, J. B. Sidgwick. Timeless, comprehensive coverage of telescopes, mirrors, lenses, mountings, telescope drives, micrometers, spectroscopes, more. 189 illustrations. 576pp. 5⅜ x 8¼. (Available in U.S. only.) 24034-7

STARS AND RELATIVITY, Ya. B. Zel'dovich and I. D. Novikov. Vol. 1 of *Relativistic Astrophysics* by famed Russian scientists. General relativity, properties of matter under astrophysical conditions, stars, and stellar systems. Deep physical insights, clear presentation. 1971 edition. References. 544pp. 5⅜ x 8¼. 69424-0

Chemistry

THE SCEPTICAL CHYMIST: The Classic 1661 Text, Robert Boyle. Boyle defines the term "element," asserting that all natural phenomena can be explained by the motion and organization of primary particles. 1911 ed. viii+232pp. 5⅜ x 8½.
42825-7

RADIOACTIVE SUBSTANCES, Marie Curie. Here is the celebrated scientist's doctoral thesis, the prelude to her receipt of the 1903 Nobel Prize. Curie discusses establishing atomic character of radioactivity found in compounds of uranium and thorium; extraction from pitchblende of polonium and radium; isolation of pure radium chloride; determination of atomic weight of radium; plus electric, photographic, luminous, heat, color effects of radioactivity. ii+94pp. 5⅜ x 8½. 42550-9

CHEMICAL MAGIC, Leonard A. Ford. Second Edition, Revised by E. Winston Grundmeier. Over 100 unusual stunts demonstrating cold fire, dust explosions, much more. Text explains scientific principles and stresses safety precautions. 128pp. 5⅜ x 8½. 67628-5

THE DEVELOPMENT OF MODERN CHEMISTRY, Aaron J. Ihde. Authoritative history of chemistry from ancient Greek theory to 20th-century innovation. Covers major chemists and their discoveries. 209 illustrations. 14 tables. Bibliographies. Indices. Appendices. 851pp. 5⅜ x 8½. 64235-6

CATALYSIS IN CHEMISTRY AND ENZYMOLOGY, William P. Jencks. Exceptionally clear coverage of mechanisms for catalysis, forces in aqueous solution, carbonyl- and acyl-group reactions, practical kinetics, more. 864pp. 5⅜ x 8½.
65460-5

ELEMENTS OF CHEMISTRY, Antoine Lavoisier. Monumental classic by founder of modern chemistry in remarkable reprint of rare 1790 Kerr translation. A must for every student of chemistry or the history of science. 539pp. 5⅜ x 8½. 64624-6

THE HISTORICAL BACKGROUND OF CHEMISTRY, Henry M. Leicester. Evolution of ideas, not individual biography. Concentrates on formulation of a coherent set of chemical laws. 260pp. 5⅜ x 8½. 61053-5

A SHORT HISTORY OF CHEMISTRY, J. R. Partington. Classic exposition explores origins of chemistry, alchemy, early medical chemistry, nature of atmosphere, theory of valency, laws and structure of atomic theory, much more. 428pp. 5⅜ x 8½. (Available in U.S. only.) 65977-1

GENERAL CHEMISTRY, Linus Pauling. Revised 3rd edition of classic first-year text by Nobel laureate. Atomic and molecular structure, quantum mechanics, statistical mechanics, thermodynamics correlated with descriptive chemistry. Problems. 992pp. 5⅜ x 8½. 65622-5

FROM ALCHEMY TO CHEMISTRY, John Read. Broad, humanistic treatment focuses on great figures of chemistry and ideas that revolutionized the science. 50 illustrations. 240pp. 5⅜ x 8½. 28690-8

Engineering

DE RE METALLICA, Georgius Agricola. The famous Hoover translation of greatest treatise on technological chemistry, engineering, geology, mining of early modern times (1556). All 289 original woodcuts. 638pp. 6¾ x 11. 60006-8

FUNDAMENTALS OF ASTRODYNAMICS, Roger Bate et al. Modern approach developed by U.S. Air Force Academy. Designed as a first course. Problems, exercises. Numerous illustrations. 455pp. 5⅜ x 8½. 60061-0

DYNAMICS OF FLUIDS IN POROUS MEDIA, Jacob Bear. For advanced students of ground water hydrology, soil mechanics and physics, drainage and irrigation engineering, and more. 335 illustrations. Exercises, with answers. 784pp. 6⅛ x 9¼. 65675-6

THEORY OF VISCOELASTICITY (Second Edition), Richard M. Christensen. Complete, consistent description of the linear theory of the viscoelastic behavior of materials. Problem-solving techniques discussed. 1982 edition. 29 figures. xiv+364pp. 6⅛ x 9¼. 42880-X

MECHANICS, J. P. Den Hartog. A classic introductory text or refresher. Hundreds of applications and design problems illuminate fundamentals of trusses, loaded beams and cables, etc. 334 answered problems. 462pp. 5⅜ x 8½. 60754-2

MECHANICAL VIBRATIONS, J. P. Den Hartog. Classic textbook offers lucid explanations and illustrative models, applying theories of vibrations to a variety of practical industrial engineering problems. Numerous figures. 233 problems, solutions. Appendix. Index. Preface. 436pp. 5⅜ x 8½. 64785-4

STRENGTH OF MATERIALS, J. P. Den Hartog. Full, clear treatment of basic material (tension, torsion, bending, etc.) plus advanced material on engineering methods, applications. 350 answered problems. 323pp. 5⅜ x 8½. 60755-0

A HISTORY OF MECHANICS, René Dugas. Monumental study of mechanical principles from antiquity to quantum mechanics. Contributions of ancient Greeks, Galileo, Leonardo, Kepler, Lagrange, many others. 671pp. 5⅜ x 8½. 65632-2

STABILITY THEORY AND ITS APPLICATIONS TO STRUCTURAL MECHANICS, Clive L. Dym. Self-contained text focuses on Koiter postbuckling analyses, with mathematical notions of stability of motion. Basing minimum energy principles for static stability upon dynamic concepts of stability of motion, it develops asymptotic buckling and postbuckling analyses from potential energy considerations, with applications to columns, plates, and arches. 1974 ed. 208pp. 5⅜ x 8½. 42541-X

METAL FATIGUE, N. E. Frost, K. J. Marsh, and L. P. Pook. Definitive, clearly written, and well-illustrated volume addresses all aspects of the subject, from the historical development of understanding metal fatigue to vital concepts of the cyclic stress that causes a crack to grow. Includes 7 appendixes. 544pp. 5⅜ x 8½. 40927-9

ROCKETS, Robert Goddard. Two of the most significant publications in the history of rocketry and jet propulsion: "A Method of Reaching Extreme Altitudes" (1919) and "Liquid Propellant Rocket Development" (1936). 128pp. 5⅜ x 8½. 42537-1

STATISTICAL MECHANICS: Principles and Applications, Terrell L. Hill. Standard text covers fundamentals of statistical mechanics, applications to fluctuation theory, imperfect gases, distribution functions, more. 448pp. 5⅜ x 8½. 65390-0

ENGINEERING AND TECHNOLOGY 1650–1750: Illustrations and Texts from Original Sources, Martin Jensen. Highly readable text with more than 200 contemporary drawings and detailed engravings of engineering projects dealing with surveying, leveling, materials, hand tools, lifting equipment, transport and erection, piling, bailing, water supply, hydraulic engineering, and more. Among the specific projects outlined–transporting a 50-ton stone to the Louvre, erecting an obelisk, building timber locks, and dredging canals. 207pp. 8⅜ x 11¼. 42232-1

THE VARIATIONAL PRINCIPLES OF MECHANICS, Cornelius Lanczos. Graduate level coverage of calculus of variations, equations of motion, relativistic mechanics, more. First inexpensive paperbound edition of classic treatise. Index. Bibliography. 418pp. 5⅜ x 8½. 65067-7

PROTECTION OF ELECTRONIC CIRCUITS FROM OVERVOLTAGES, Ronald B. Standler. Five-part treatment presents practical rules and strategies for circuits designed to protect electronic systems from damage by transient overvoltages. 1989 ed. xxiv+434pp. 6⅛ x 9¼. 42552-5

ROTARY WING AERODYNAMICS, W. Z. Stepniewski. Clear, concise text covers aerodynamic phenomena of the rotor and offers guidelines for helicopter performance evaluation. Originally prepared for NASA. 537 figures. 640pp. 6⅛ x 9¼. 64647-5

INTRODUCTION TO SPACE DYNAMICS, William Tyrrell Thomson. Comprehensive, classic introduction to space-flight engineering for advanced undergraduate and graduate students. Includes vector algebra, kinematics, transformation of coordinates. Bibliography. Index. 352pp. 5⅜ x 8½. 65113-4

HISTORY OF STRENGTH OF MATERIALS, Stephen P. Timoshenko. Excellent historical survey of the strength of materials with many references to the theories of elasticity and structure. 245 figures. 452pp. 5⅜ x 8½. 61187-6

ANALYTICAL FRACTURE MECHANICS, David J. Unger. Self-contained text supplements standard fracture mechanics texts by focusing on analytical methods for determining crack-tip stress and strain fields. 336pp. 6⅛ x 9¼. 41737-9

STATISTICAL MECHANICS OF ELASTICITY, J. H. Weiner. Advanced, self-contained treatment illustrates general principles and elastic behavior of solids. Part 1, based on classical mechanics, studies thermoelastic behavior of crystalline and polymeric solids. Part 2, based on quantum mechanics, focuses on interatomic force laws, behavior of solids, and thermally activated processes. For students of physics and chemistry and for polymer physicists. 1983 ed. 96 figures. 496pp. 5⅜ x 8½. 42260-7

Mathematics

FUNCTIONAL ANALYSIS (Second Corrected Edition), George Bachman and Lawrence Narici. Excellent treatment of subject geared toward students with background in linear algebra, advanced calculus, physics, and engineering. Text covers introduction to inner-product spaces, normed, metric spaces, and topological spaces; complete orthonormal sets, the Hahn-Banach Theorem and its consequences, and many other related subjects. 1966 ed. 544pp. 6¼ x 9¼. 40251-7

ASYMPTOTIC EXPANSIONS OF INTEGRALS, Norman Bleistein & Richard A. Handelsman. Best introduction to important field with applications in a variety of scientific disciplines. New preface. Problems. Diagrams. Tables. Bibliography. Index. 448pp. 5⅜ x 8½. 65082-0

VECTOR AND TENSOR ANALYSIS WITH APPLICATIONS, A. I. Borisenko and I. E. Tarapov. Concise introduction. Worked-out problems, solutions, exercises. 257pp. 5⅜ x 8¼. 63833-2

THE ABSOLUTE DIFFERENTIAL CALCULUS (CALCULUS OF TENSORS), Tullio Levi-Civita. Great 20th-century mathematician's classic work on material necessary for mathematical grasp of theory of relativity. 452pp. 5⅜ x 8¼. 63401-9

AN INTRODUCTION TO ORDINARY DIFFERENTIAL EQUATIONS, Earl A. Coddington. A thorough and systematic first course in elementary differential equations for undergraduates in mathematics and science, with many exercises and problems (with answers). Index. 304pp. 5⅜ x 8½. 65942-9

FOURIER SERIES AND ORTHOGONAL FUNCTIONS, Harry F. Davis. An incisive text combining theory and practical example to introduce Fourier series, orthogonal functions and applications of the Fourier method to boundary-value problems. 570 exercises. Answers and notes. 416pp. 5⅜ x 8½. 65973-9

COMPUTABILITY AND UNSOLVABILITY, Martin Davis. Classic graduate-level introduction to theory of computability, usually referred to as theory of recurrent functions. New preface and appendix. 288pp. 5⅜ x 8½. 61471-9

ASYMPTOTIC METHODS IN ANALYSIS, N. G. de Bruijn. An inexpensive, comprehensive guide to asymptotic methods–the pioneering work that teaches by explaining worked examples in detail. Index. 224pp. 5⅜ x 8½ 64221-6

APPLIED COMPLEX VARIABLES, John W. Dettman. Step-by-step coverage of fundamentals of analytic function theory–plus lucid exposition of five important applications: Potential Theory; Ordinary Differential Equations; Fourier Transforms; Laplace Transforms; Asymptotic Expansions. 66 figures. Exercises at chapter ends. 512pp. 5⅜ x 8½. 64670-X

INTRODUCTION TO LINEAR ALGEBRA AND DIFFERENTIAL EQUATIONS, John W. Dettman. Excellent text covers complex numbers, determinants, orthonormal bases, Laplace transforms, much more. Exercises with solutions. Undergraduate level. 416pp. 5⅜ x 8½. 65191-6

CALCULUS OF VARIATIONS WITH APPLICATIONS, George M. Ewing. Applications-oriented introduction to variational theory develops insight and promotes understanding of specialized books, research papers. Suitable for advanced undergraduate/graduate students as primary, supplementary text. 352pp. 5⅜ x 8½.
64856-7

COMPLEX VARIABLES, Francis J. Flanigan. Unusual approach, delaying complex algebra till harmonic functions have been analyzed from real variable viewpoint. Includes problems with answers. 364pp. 5⅜ x 8½. 61388-7

AN INTRODUCTION TO THE CALCULUS OF VARIATIONS, Charles Fox. Graduate-level text covers variations of an integral, isoperimetrical problems, least action, special relativity, approximations, more. References. 279pp. 5⅜ x 8½.
65499-0

COUNTEREXAMPLES IN ANALYSIS, Bernard R. Gelbaum and John M. H. Olmsted. These counterexamples deal mostly with the part of analysis known as "real variables." The first half covers the real number system, and the second half encompasses higher dimensions. 1962 edition. xxiv+198pp. 5⅜ x 8½. 42875-3

CATASTROPHE THEORY FOR SCIENTISTS AND ENGINEERS, Robert Gilmore. Advanced-level treatment describes mathematics of theory grounded in the work of Poincaré, R. Thom, other mathematicians. Also important applications to problems in mathematics, physics, chemistry, and engineering. 1981 edition. References. 28 tables. 397 black-and-white illustrations. xvii+666pp. 6⅛ x 9¼.
67539-4

INTRODUCTION TO DIFFERENCE EQUATIONS, Samuel Goldberg. Exceptionally clear exposition of important discipline with applications to sociology, psychology, economics. Many illustrative examples; over 250 problems. 260pp. 5⅜ x 8½.
65084-7

NUMERICAL METHODS FOR SCIENTISTS AND ENGINEERS, Richard Hamming. Classic text stresses frequency approach in coverage of algorithms, polynomial approximation, Fourier approximation, exponential approximation, other topics. Revised and enlarged 2nd edition. 721pp. 5⅜ x 8½. 65241-6

INTRODUCTION TO NUMERICAL ANALYSIS (2nd Edition), F. B. Hildebrand. Classic, fundamental treatment covers computation, approximation, interpolation, numerical differentiation and integration, other topics. 150 new problems. 669pp. 5⅜ x 8½. 65363-3

THREE PEARLS OF NUMBER THEORY, A. Y. Khinchin. Three compelling puzzles require proof of a basic law governing the world of numbers. Challenges concern van der Waerden's theorem, the Landau-Schnirelmann hypothesis and Mann's theorem, and a solution to Waring's problem. Solutions included. 64pp. 5⅜ x 8½.
40026-3

THE PHILOSOPHY OF MATHEMATICS: An Introductory Essay, Stephan Körner. Surveys the views of Plato, Aristotle, Leibniz & Kant concerning propositions and theories of applied and pure mathematics. Introduction. Two appendices. Index. 198pp. 5⅜ x 8½. 25048-2

INTRODUCTORY REAL ANALYSIS, A.N. Kolmogorov, S. V. Fomin. Translated by Richard A. Silverman. Self-contained, evenly paced introduction to real and functional analysis. Some 350 problems. 403pp. 5⅜ x 8½. 61226-0

APPLIED ANALYSIS, Cornelius Lanczos. Classic work on analysis and design of finite processes for approximating solution of analytical problems. Algebraic equations, matrices, harmonic analysis, quadrature methods, more. 559pp. 5⅜ x 8½. 65656-X

AN INTRODUCTION TO ALGEBRAIC STRUCTURES, Joseph Landin. Superb self-contained text covers "abstract algebra": sets and numbers, theory of groups, theory of rings, much more. Numerous well-chosen examples, exercises. 247pp. 5⅜ x 8½. 65940-2

QUALITATIVE THEORY OF DIFFERENTIAL EQUATIONS, V. V. Nemytskii and V.V. Stepanov. Classic graduate-level text by two prominent Soviet mathematicians covers classical differential equations as well as topological dynamics and ergodic theory. Bibliographies. 523pp. 5⅜ x 8½. 65954-2

THEORY OF MATRICES, Sam Perlis. Outstanding text covering rank, nonsingularity and inverses in connection with the development of canonical matrices under the relation of equivalence, and without the intervention of determinants. Includes exercises. 237pp. 5⅜ x 8½. 66810-X

INTRODUCTION TO ANALYSIS, Maxwell Rosenlicht. Unusually clear, accessible coverage of set theory, real number system, metric spaces, continuous functions, Riemann integration, multiple integrals, more. Wide range of problems. Undergraduate level. Bibliography. 254pp. 5⅜ x 8½. 65038-3

MODERN NONLINEAR EQUATIONS, Thomas L. Saaty. Emphasizes practical solution of problems; covers seven types of equations. ". . . a welcome contribution to the existing literature. . . . "—*Math Reviews*. 490pp. 5⅜ x 8½. 64232-1

MATRICES AND LINEAR ALGEBRA, Hans Schneider and George Phillip Barker. Basic textbook covers theory of matrices and its applications to systems of linear equations and related topics such as determinants, eigenvalues, and differential equations. Numerous exercises. 432pp. 5⅜ x 8½. 66014-1

MATHEMATICS APPLIED TO CONTINUUM MECHANICS, Lee A. Segel. Analyzes models of fluid flow and solid deformation. For upper-level math, science, and engineering students. 608pp. 5⅜ x 8½. 65369-2

ELEMENTS OF REAL ANALYSIS, David A. Sprecher. Classic text covers fundamental concepts, real number system, point sets, functions of a real variable, Fourier series, much more. Over 500 exercises. 352pp. 5⅜ x 8½. 65385-4

SET THEORY AND LOGIC, Robert R. Stoll. Lucid introduction to unified theory of mathematical concepts. Set theory and logic seen as tools for conceptual understanding of real number system. 496pp. 5⅜ x 8¼. 63829-4

TENSOR CALCULUS, J.L. Synge and A. Schild. Widely used introductory text covers spaces and tensors, basic operations in Riemannian space, non-Riemannian spaces, etc. 324pp. 5⅜ x 8¼. 63612-7

ORDINARY DIFFERENTIAL EQUATIONS, Morris Tenenbaum and Harry Pollard. Exhaustive survey of ordinary differential equations for undergraduates in mathematics, engineering, science. Thorough analysis of theorems. Diagrams. Bibliography. Index. 818pp. 5⅜ x 8½. 64940-7

INTEGRAL EQUATIONS, F. G. Tricomi. Authoritative, well-written treatment of extremely useful mathematical tool with wide applications. Volterra Equations, Fredholm Equations, much more. Advanced undergraduate to graduate level. Exercises. Bibliography. 238pp. 5⅜ x 8½. 64828-1

FOURIER SERIES, Georgi P. Tolstov. Translated by Richard A. Silverman. A valuable addition to the literature on the subject, moving clearly from subject to subject and theorem to theorem. 107 problems, answers. 336pp. 5⅜ x 8½. 63317-9

INTRODUCTION TO MATHEMATICAL THINKING, Friedrich Waismann. Examinations of arithmetic, geometry, and theory of integers; rational and natural numbers; complete induction; limit and point of accumulation; remarkable curves; complex and hypercomplex numbers, more. 1959 ed. 27 figures. xii+260pp. 5⅜ x 8½. 42804-4

POPULAR LECTURES ON MATHEMATICAL LOGIC, Hao Wang. Noted logician's lucid treatment of historical developments, set theory, model theory, recursion theory and constructivism, proof theory, more. 3 appendixes. Bibliography. 1981 ed. ix+283pp. 5⅜ x 8½. 67632-3

CALCULUS OF VARIATIONS, Robert Weinstock. Basic introduction covering isoperimetric problems, theory of elasticity, quantum mechanics, electrostatics, etc. Exercises throughout. 326pp. 5⅜ x 8½. 63069-2

THE CONTINUUM: A Critical Examination of the Foundation of Analysis, Hermann Weyl. Classic of 20th-century foundational research deals with the conceptual problem posed by the continuum. 156pp. 5⅜ x 8½. 67982-9

CHALLENGING MATHEMATICAL PROBLEMS WITH ELEMENTARY SOLUTIONS, A. M. Yaglom and I. M. Yaglom. Over 170 challenging problems on probability theory, combinatorial analysis, points and lines, topology, convex polygons, many other topics. Solutions. Total of 445pp. 5⅜ x 8½. Two-vol. set.
Vol. I: 65536-9 Vol. II: 65537-7

INTRODUCTION TO PARTIAL DIFFERENTIAL EQUATIONS WITH APPLICATIONS, E. C. Zachmanoglou and Dale W. Thoe. Essentials of partial differential equations applied to common problems in engineering and the physical sciences. Problems and answers. 416pp. 5⅜ x 8½. 65251-3

THE THEORY OF GROUPS, Hans J. Zassenhaus. Well-written graduate-level text acquaints reader with group-theoretic methods and demonstrates their usefulness in mathematics. Axioms, the calculus of complexes, homomorphic mapping, p-group theory, more. 276pp. 5⅜ x 8½. 40922-8

Math–Decision Theory, Statistics, Probability

ELEMENTARY DECISION THEORY, Herman Chernoff and Lincoln E. Moses. Clear introduction to statistics and statistical theory covers data processing, probability and random variables, testing hypotheses, much more. Exercises. 364pp. 5⅜ x 8½. 65218-1

STATISTICS MANUAL, Edwin L. Crow et al. Comprehensive, practical collection of classical and modern methods prepared by U.S. Naval Ordnance Test Station. Stress on use. Basics of statistics assumed. 288pp. 5⅜ x 8½. 60599-X

SOME THEORY OF SAMPLING, William Edwards Deming. Analysis of the problems, theory, and design of sampling techniques for social scientists, industrial managers, and others who find statistics important at work. 61 tables. 90 figures. xvii +602pp. 5⅜ x 8½. 64684-X

LINEAR PROGRAMMING AND ECONOMIC ANALYSIS, Robert Dorfman, Paul A. Samuelson and Robert M. Solow. First comprehensive treatment of linear programming in standard economic analysis. Game theory, modern welfare economics, Leontief input-output, more. 525pp. 5⅜ x 8½. 65491-5

PROBABILITY: An Introduction, Samuel Goldberg. Excellent basic text covers set theory, probability theory for finite sample spaces, binomial theorem, much more. 360 problems. Bibliographies. 322pp. 5⅜ x 8½. 65252-1

GAMES AND DECISIONS: Introduction and Critical Survey, R. Duncan Luce and Howard Raiffa. Superb nontechnical introduction to game theory, primarily applied to social sciences. Utility theory, zero-sum games, n-person games, decision-making, much more. Bibliography. 509pp. 5⅜ x 8½. 65943-7

INTRODUCTION TO THE THEORY OF GAMES, J. C. C. McKinsey. This comprehensive overview of the mathematical theory of games illustrates applications to situations involving conflicts of interest, including economic, social, political, and military contexts. Appropriate for advanced undergraduate and graduate courses; advanced calculus a prerequisite. 1952 ed. x+372pp. 5⅜ x 8½. 42811-7

FIFTY CHALLENGING PROBLEMS IN PROBABILITY WITH SOLUTIONS, Frederick Mosteller. Remarkable puzzlers, graded in difficulty, illustrate elementary and advanced aspects of probability. Detailed solutions. 88pp. 5⅜ x 8½. 65355-2

PROBABILITY THEORY: A Concise Course, Y. A. Rozanov. Highly readable, self-contained introduction covers combination of events, dependent events, Bernoulli trials, etc. 148pp. 5⅜ x 8¼. 63544-9

STATISTICAL METHOD FROM THE VIEWPOINT OF QUALITY CONTROL, Walter A. Shewhart. Important text explains regulation of variables, uses of statistical control to achieve quality control in industry, agriculture, other areas. 192pp. 5⅜ x 8½. 65232-7

Math–Geometry and Topology

ELEMENTARY CONCEPTS OF TOPOLOGY, Paul Alexandroff. Elegant, intuitive approach to topology from set-theoretic topology to Betti groups; how concepts of topology are useful in math and physics. 25 figures. 57pp. 5⅜ x 8½. 60747-X

COMBINATORIAL TOPOLOGY, P. S. Alexandrov. Clearly written, well-organized, three-part text begins by dealing with certain classic problems without using the formal techniques of homology theory and advances to the central concept, the Betti groups. Numerous detailed examples. 654pp. 5⅜ x 8½. 40179-0

EXPERIMENTS IN TOPOLOGY, Stephen Barr. Classic, lively explanation of one of the byways of mathematics. Klein bottles, Moebius strips, projective planes, map coloring, problem of the Koenigsberg bridges, much more, described with clarity and wit. 43 figures. 210pp. 5⅜ x 8½. 25933-1

CONFORMAL MAPPING ON RIEMANN SURFACES, Harvey Cohn. Lucid, insightful book presents ideal coverage of subject. 334 exercises make book perfect for self-study. 55 figures. 352pp. 5⅜ x 8¼. 64025-6

THE GEOMETRY OF RENÉ DESCARTES, René Descartes. The great work founded analytical geometry. Original French text, Descartes's own diagrams, together with definitive Smith-Latham translation. 244pp. 5⅜ x 8½. 60068-8

PRACTICAL CONIC SECTIONS: The Geometric Properties of Ellipses, Parabolas and Hyperbolas, J. W. Downs. This text shows how to create ellipses, parabolas, and hyperbolas. It also presents historical background on their ancient origins and describes the reflective properties and roles of curves in design applications. 1993 ed. 98 figures. xii+100pp. 6½ x 9¼. 42876-1

THE THIRTEEN BOOKS OF EUCLID'S ELEMENTS, translated with introduction and commentary by Thomas L. Heath. Definitive edition. Textual and linguistic notes, mathematical analysis. 2,500 years of critical commentary. Unabridged. 1,414pp. 5⅜ x 8½. Three-vol. set. Vol. I: 60088-2 Vol. II: 60089-0 Vol. III: 60090-4

GEOMETRY OF COMPLEX NUMBERS, Hans Schwerdtfeger. Illuminating, widely praised book on analytic geometry of circles, the Moebius transformation, and two-dimensional non-Euclidean geometries. 200pp. 5⅜ x 8¼. 63830-8

DIFFERENTIAL GEOMETRY, Heinrich W. Guggenheimer. Local differential geometry as an application of advanced calculus and linear algebra. Curvature, transformation groups, surfaces, more. Exercises. 62 figures. 378pp. 5⅜ x 8¼. 63433-7

CURVATURE AND HOMOLOGY: Enlarged Edition, Samuel I. Goldberg. Revised edition examines topology of differentiable manifolds; curvature, homology of Riemannian manifolds; compact Lie groups; complex manifolds; curvature, homology of Kaehler manifolds. New Preface. Four new appendixes. 416pp. 5⅜ x 8½. 40207-X

History of Math

THE WORKS OF ARCHIMEDES, Archimedes (T. L. Heath, ed.). Topics include the famous problems of the ratio of the areas of a cylinder and an inscribed sphere; the measurement of a circle; the properties of conoids, spheroids, and spirals; and the quadrature of the parabola. Informative introduction. clxxxvi+326pp; supplement, 52pp. 5⅜ x 8½. 42084-1

A SHORT ACCOUNT OF THE HISTORY OF MATHEMATICS, W. W. Rouse Ball. One of clearest, most authoritative surveys from the Egyptians and Phoenicians through 19th-century figures such as Grassman, Galois, Riemann. Fourth edition. 522pp. 5⅜ x 8½. 20630-0

THE HISTORY OF THE CALCULUS AND ITS CONCEPTUAL DEVELOPMENT, Carl B. Boyer. Origins in antiquity, medieval contributions, work of Newton, Leibniz, rigorous formulation. Treatment is verbal. 346pp. 5⅜ x 8½. 60509-4

THE HISTORICAL ROOTS OF ELEMENTARY MATHEMATICS, Lucas N. H. Bunt, Phillip S. Jones, and Jack D. Bedient. Fundamental underpinnings of modern arithmetic, algebra, geometry, and number systems derived from ancient civilizations. 320pp. 5⅜ x 8½. 25563-8

A HISTORY OF MATHEMATICAL NOTATIONS, Florian Cajori. This classic study notes the first appearance of a mathematical symbol and its origin, the competition it encountered, its spread among writers in different countries, its rise to popularity, its eventual decline or ultimate survival. Original 1929 two-volume edition presented here in one volume. xxviii+820pp. 5⅜ x 8½. 67766-4

GAMES, GODS & GAMBLING: A History of Probability and Statistical Ideas, F. N. David. Episodes from the lives of Galileo, Fermat, Pascal, and others illustrate this fascinating account of the roots of mathematics. Features thought-provoking references to classics, archaeology, biography, poetry. 1962 edition. 304pp. 5⅜ x 8½. (Available in U.S. only.) 40023-9

OF MEN AND NUMBERS: The Story of the Great Mathematicians, Jane Muir. Fascinating accounts of the lives and accomplishments of history's greatest mathematical minds–Pythagoras, Descartes, Euler, Pascal, Cantor, many more. Anecdotal, illuminating. 30 diagrams. Bibliography. 256pp. 5⅜ x 8½. 28973-7

HISTORY OF MATHEMATICS, David E. Smith. Nontechnical survey from ancient Greece and Orient to late 19th century; evolution of arithmetic, geometry, trigonometry, calculating devices, algebra, the calculus. 362 illustrations. 1,355pp. 5⅜ x 8½. Two-vol. set. Vol. I: 20429-4 Vol. II: 20430-8

A CONCISE HISTORY OF MATHEMATICS, Dirk J. Struik. The best brief history of mathematics. Stresses origins and covers every major figure from ancient Near East to 19th century. 41 illustrations. 195pp. 5⅜ x 8½. 60255-9

Physics

OPTICAL RESONANCE AND TWO-LEVEL ATOMS, L. Allen and J. H. Eberly. Clear, comprehensive introduction to basic principles behind all quantum optical resonance phenomena. 53 illustrations. Preface. Index. 256pp. 5⅜ x 8½. 65533-4

QUANTUM THEORY, David Bohm. This advanced undergraduate-level text presents the quantum theory in terms of qualitative and imaginative concepts, followed by specific applications worked out in mathematical detail. Preface. Index. 655pp. 5⅜ x 8½. 65969-0

ATOMIC PHYSICS: 8th edition, Max Born. Nobel laureate's lucid treatment of kinetic theory of gases, elementary particles, nuclear atom, wave-corpuscles, atomic structure and spectral lines, much more. Over 40 appendices, bibliography. 495pp. 5⅜ x 8½. 65984-4

A SOPHISTICATE'S PRIMER OF RELATIVITY, P. W. Bridgman. Geared toward readers already acquainted with special relativity, this book transcends the view of theory as a working tool to answer natural questions: What is a frame of reference? What is a "law of nature"? What is the role of the "observer"? Extensive treatment, written in terms accessible to those without a scientific background. 1983 ed. xlviii+172pp. 5⅜ x 8½. 42549-5

AN INTRODUCTION TO HAMILTONIAN OPTICS, H. A. Buchdahl. Detailed account of the Hamiltonian treatment of aberration theory in geometrical optics. Many classes of optical systems defined in terms of the symmetries they possess. Problems with detailed solutions. 1970 edition. xv+360pp. 5⅜ x 8½. 67597-1

PRIMER OF QUANTUM MECHANICS, Marvin Chester. Introductory text examines the classical quantum bead on a track: its state and representations; operator eigenvalues; harmonic oscillator and bound bead in a symmetric force field; and bead in a spherical shell. Other topics include spin, matrices, and the structure of quantum mechanics; the simplest atom; indistinguishable particles; and stationary-state perturbation theory. 1992 ed. xiv+314pp. 6⅛ x 9¼. 42878-8

LECTURES ON QUANTUM MECHANICS, Paul A. M. Dirac. Four concise, brilliant lectures on mathematical methods in quantum mechanics from Nobel Prize–winning quantum pioneer build on idea of visualizing quantum theory through the use of classical mechanics. 96pp. 5⅜ x 8½. 41713-1

THIRTY YEARS THAT SHOOK PHYSICS: The Story of Quantum Theory, George Gamow. Lucid, accessible introduction to influential theory of energy and matter. Careful explanations of Dirac's anti-particles, Bohr's model of the atom, much more. 12 plates. Numerous drawings. 240pp. 5⅜ x 8½. 24895-X

ELECTRONIC STRUCTURE AND THE PROPERTIES OF SOLIDS: The Physics of the Chemical Bond, Walter A. Harrison. Innovative text offers basic understanding of the electronic structure of covalent and ionic solids, simple metals, transition metals and their compounds. Problems. 1980 edition. 582pp. 6⅛ x 9¼. 66021-4

HYDRODYNAMIC AND HYDROMAGNETIC STABILITY, S. Chandrasekhar. Lucid examination of the Rayleigh-Benard problem; clear coverage of the theory of instabilities causing convection. 704pp. 5⅜ x 8¼. 64071-X

INVESTIGATIONS ON THE THEORY OF THE BROWNIAN MOVEMENT, Albert Einstein. Five papers (1905–8) investigating dynamics of Brownian motion and evolving elementary theory. Notes by R. Fürth. 122pp. 5⅜ x 8½. 60304-0

THE PHYSICS OF WAVES, William C. Elmore and Mark A. Heald. Unique overview of classical wave theory. Acoustics, optics, electromagnetic radiation, more. Ideal as classroom text or for self-study. Problems. 477pp. 5⅜ x 8½. 64926-1

PHYSICAL PRINCIPLES OF THE QUANTUM THEORY, Werner Heisenberg. Nobel Laureate discusses quantum theory, uncertainty, wave mechanics, work of Dirac, Schroedinger, Compton, Wilson, Einstein, etc. 184pp. 5⅜ x 8½. 60113-7

ATOMIC SPECTRA AND ATOMIC STRUCTURE, Gerhard Herzberg. One of best introductions; especially for specialist in other fields. Treatment is physical rather than mathematical. 80 illustrations. 257pp. 5⅜ x 8½. 60115-3

AN INTRODUCTION TO STATISTICAL THERMODYNAMICS, Terrell L. Hill. Excellent basic text offers wide-ranging coverage of quantum statistical mechanics, systems of interacting molecules, quantum statistics, more. 523pp. 5⅜ x 8½. 65242-4

THEORETICAL PHYSICS, Georg Joos, with Ira M. Freeman. Classic overview covers essential math, mechanics, electromagnetic theory, thermodynamics, quantum mechanics, nuclear physics, other topics. xxiii+885pp. 5⅜ x 8½. 65227-0

PROBLEMS AND SOLUTIONS IN QUANTUM CHEMISTRY AND PHYSICS, Charles S. Johnson, Jr. and Lee G. Pedersen. Unusually varied problems, detailed solutions in coverage of quantum mechanics, wave mechanics, angular momentum, molecular spectroscopy, more. 280 problems, 139 supplementary exercises. 430pp. 6½ x 9¼. 65236-X

THEORETICAL SOLID STATE PHYSICS, Vol. I: Perfect Lattices in Equilibrium; Vol. II: Non-Equilibrium and Disorder, William Jones and Norman H. March. Monumental reference work covers fundamental theory of equilibrium properties of perfect crystalline solids, non-equilibrium properties, defects and disordered systems. Total of 1,301pp. 5⅜ x 8½. Vol. I: 65015-4 Vol. II: 65016-2

WHAT IS RELATIVITY? L. D. Landau and G. B. Rumer. Written by a Nobel Prize physicist and his distinguished colleague, this compelling book explains the special theory of relativity to readers with no scientific background, using such familiar objects as trains, rulers, and clocks. 1960 ed. vi+72pp. 23 b/w illustrations. 5⅜ x 8½. 42806-0 $6.95

A TREATISE ON ELECTRICITY AND MAGNETISM, James Clerk Maxwell. Important foundation work of modern physics. Brings to final form Maxwell's theory of electromagnetism and rigorously derives his general equations of field theory. 1,084pp. 5⅜ x 8½. Two-vol. set. Vol. I: 60636-8 Vol. II: 60637-6

QUANTUM MECHANICS: Principles and Formalism, Roy McWeeny. Graduate student-oriented volume develops subject as fundamental discipline, opening with review of origins of Schrödinger's equations and vector spaces. Focusing on main principles of quantum mechanics and their immediate consequences, it concludes with final generalizations covering alternative "languages" or representations. 1972 ed. 15 figures. xi+155pp. 5⅜ x 8½. 42829-X

INTRODUCTION TO QUANTUM MECHANICS WITH APPLICATIONS TO CHEMISTRY, Linus Pauling & E. Bright Wilson, Jr. Classic undergraduate text by Nobel Prize winner applies quantum mechanics to chemical and physical problems. Numerous tables and figures enhance the text. Chapter bibliographies. Appendices. Index. 468pp. 5⅜ x 8½. 64871-0

METHODS OF THERMODYNAMICS, Howard Reiss. Outstanding text focuses on physical technique of thermodynamics, typical problem areas of understanding, and significance and use of thermodynamic potential. 1965 edition. 238pp. 5⅜ x 8½. 69445-3

TENSOR ANALYSIS FOR PHYSICISTS, J. A. Schouten. Concise exposition of the mathematical basis of tensor analysis, integrated with well-chosen physical examples of the theory. Exercises. Index. Bibliography. 289pp. 5⅜ x 8½. 65582-2

THE ELECTROMAGNETIC FIELD, Albert Shadowitz. Comprehensive undergraduate text covers basics of electric and magnetic fields, builds up to electromagnetic theory. Also related topics, including relativity. Over 900 problems. 768pp. 5⅜ x 8¼. 65660-8

GREAT EXPERIMENTS IN PHYSICS: Firsthand Accounts from Galileo to Einstein, Morris H. Shamos (ed.). 25 crucial discoveries: Newton's laws of motion, Chadwick's study of the neutron, Hertz on electromagnetic waves, more. Original accounts clearly annotated. 370pp. 5⅜ x 8½. 25346-5

RELATIVITY, THERMODYNAMICS AND COSMOLOGY, Richard C. Tolman. Landmark study extends thermodynamics to special, general relativity; also applications of relativistic mechanics, thermodynamics to cosmological models. 501pp. 5⅜ x 8½. 65383-8

STATISTICAL PHYSICS, Gregory H. Wannier. Classic text combines thermodynamics, statistical mechanics, and kinetic theory in one unified presentation of thermal physics. Problems with solutions. Bibliography. 532pp. 5⅜ x 8½. 65401-X

DOVER BOOKS ON ENGINEERING

THEORY OF WING SECTIONS: INCLUDING A SUMMARY OF AIRFOIL DATA, Ira H. Abbott and A. E. von Doenhoff. (60586-8)

VECTORS, TENSORS AND THE BASIC EQUATIONS OF FLUID MECHANICS, Rutherford Aris. (66110-5)

AERODYNAMICS OF WINGS AND BODIES, Holt Ashley and Marten Landahl. (64899-0)

DYNAMICS OF FLUIDS IN POROUS MEDIA, Jacob Bear. (65675-6)

AEROELASTICITY, Raymond L. Bisplinghoff, Holt Ashley and Robert L. Halfman. (69189-6)

THE PHENOMENA OF FLUID MOTIONS, Robert S. Brodkey. (68605-1)

ANALYTICAL MECHANICS OF GEARS, Earle Buckingham. (65712-4)

HYDRODYNAMIC AND HYDROMAGNETIC STABILITY, S. Chandrasekhar. (64071-X)

A HISTORY OF MECHANICS, René Dugas. (65632-2)

PROBABILISTIC THEORY OF STRUCTURES, SECOND EDITION. Isaac Elishakoff. (40691-1)

CREEP AND RELAXATION OF NONLINEAR VISCOELASTIC MATERIALS, William N. Findley, James S. Lai, and Kasif Onaran. (66016-8)

METAL FATIGUE, N. E. Frost, K. J. Marsh, and L. P. Pook. (40927-9)

WAVE MOTION IN ELASTIC SOLIDS, Karl F. Graff. (66745-6)

FLUID MECHANICS, Robert A. Granger. (68356-7)

DIGITAL FILTERS, R. W. Hamming. (65088-X)

GROUNDWATER AND SEEPAGE, Milton E. Harr. (66881-9)

RELIABILITY-BASED DESIGN IN CIVIL ENGINEERING, Milton E. Harr. (69429-1)

MECHANICAL VIBRATIONS, J. P. Den Hartog. (64785-4)

THE FINITE ELEMENT METHOD: LINEAR STATIC AND DYNAMIC FINITE ELEMENT ANALYSIS, Thomas J. R. Hughes. (41181-8)

STRESS WAVES IN SOLIDS, H. Kolsky. (61098-5)

MATHEMATICAL HANDBOOK FOR SCIENTISTS AND ENGINEERS: DEFINITIONS, THEOREMS, AND FORMULAS FOR REFERENCE AND REVIEW, Granino A. Korn and Theresa M. Korn. (41147-8)

COMPLEX VARIABLES AND THE LAPLACE TRANSFORM FOR ENGINEERS, Wilbur R. LePage. (63926-6)

TELECOMMUNICATION SYSTEMS ENGINEERING, William C. Lindsey and Marvin K. Simon. (66838-X)

AERODYNAMICS OF V/STOL FLIGHT, Barnes W. McCormick, Jr. (40460-9)

THEORETICAL AERODYNAMICS, L. M. Milne-Thomson. (61980-X)

NON-LINEAR ELASTIC DEFORMATIONS, R. W. Ogden. (69648-0)

UNIFIED ANALYSIS AND SOLUTIONS OF HEAT AND MASS DIFFUSION, M. D. Mikhailov and M. Necati Ozisik. (67876-8)

APPLIED HYDRO- AND AEROMECHANICS, Ludwig Prandtl and O. G. Tietjens. (60375-X)

(continued on back flap)

(continued from front flap)