TIME-OPTIMAL CONTROL OF MULTIPLE INPUT LINEAR SAMPLED-DATA SYSTEMS

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

RICHARD A. BEDNAR

1971

This is to certify that the

thesis entitled

TIME-OPTIMAL CONTROL OF MULTIPLE INPUT
LINEAR SAMPLED-DATA SYSTEMS

presented by

Richard A. Bednar

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Electrical Engineering

John B. Kreer
Major professor

Date August 13, 1971

O-7639

ABSTRACT

TIME-OPTIMAL CONTROL OF MULTIPLE INPUT LINEAR SAMPLED-DATA SYSTEMS

By

Richard A. Bednar

This thesis is concerned with two problems from the theory of time-optimal control of multiple input discrete-time linear systems. The first and second problems differ mainly in the target set and the admissible controls.

The subject of the first problem is the time-optimal control of [illegible] given by G = {x(NT) [illegible] the state vector, T [illegible] number and N is the [illegible] x(NT) ε G. The amplitude [illegible] to be unconstrained [illegible] the minimum number of [illegible], and whether the sequence [illegible] not unique, one is [illegible] to reach the [illegible]

ABSTRACT

TIME-OPTIMAL CONTROL OF MULTIPLE INPUT LINEAR SAMPLED-DATA SYSTEMS

By

Richard A. Bednar

This thesis is concerned with two problems from the theory of time-optimal control of multiple input discrete-time linear systems. The first and second problems differ mainly in the target set and the admissible controls.

The subject of the first problem is the time-optimal control of a linear system whose target set is given by $G = \{x(NT): x^T(NT)x(NT) \leq R^2\}$ where $x(NT)$ is the state vector, T is the sampling period, R is a real number and N is the fewest number of samples such that $x(NT) \in G$. The amplitude of the controller is assumed to be unconstrained. Results are derived for determining the minimum number of samples required to reach the target, and whether the sequence is unique. If the sequence is not unique, one is chosen which requires minimum energy to reach the target. It is shown that this problem reduces

to finding the roots of a polynomial of order less than or equal to 2n where n is the order of the system. An alternate formulation obtains the sequence of time-optimal controls in the form of a feedback control law. The results for the open-loop controller are then extended to include a system with a delay in the input and a stochastic system.

In the second problem the initial state of the sampled-data system is to be driven in the fewest number of samples to a target set described by $\{x_i(NT): -M_i \leq x_i(NT) \leq M_i \ , i=1,2,\ldots,n\}, M_i \geq 0$, where $x_i(NT)$ are the components of the state vector $x(NT)$. The amplitude of the controller may or may not be constrained. If the sequence of controls is not unique, a sequence is chosen to satisfy a minimum fuel criterion. This problem is formulated as a linear programming problem. The theory of the Simplex Method is used to determine whether a solution exists and if it is unique. A corresponding open-loop stochastic system is considered, and the procedure for obtaining a solution is given.

Several numerical examples are solved to illustrate the theory related to each of the above problems.

TIME-OPTIMAL CONTROL OF MULTIPLE INPUT

LINEAR SAMPLED-DATA SYSTEMS

By

Richard A. Bednar

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering and
Systems Science

1971

ACKNOWLEDGMENTS

The author would like to express his appreciation to his major professors, Dr. John Kreer and Dr. Israel Korn for their interest and valuable comments related to this thesis. Gratitude is expressed toward the other guidance committee members, Dr. J. Sutherland Frame, Dr. H. E. Koenig, and Dr. R. O. Hart. A special note should be made of the joint effort provided by Dr. Frame and the author in developing Section 3.3 of this thesis.

Financial support was granted by Michigan State University for one year in the form of a teaching assistantship and for two years by a National Science Foundation Fellowship. This support is gratefully acknowledged.

ACKNOWLEDGMENTS

The author would like to express his appreciation to his major professors, Dr. John Kreer and Dr. Israel Korn for their interest and valuable comments related to this thesis. Gratitude is expressed toward the other guidance committee members, Dr. J. Sutherland Frame, Dr. H. E. Koenig, and Dr. R. O. Barr. A special note should be made of the joint effort provided by Dr. Frame and the author in developing Section 3.3 of this thesis.

Financial support was granted by Michigan State University for one year in the form of a teaching assistantship and for two years by a National Science Foundation Fellowship. This support is gratefully acknowledged.

TABLE OF CONTENTS

Chapter		Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Among the many types of optimal control design problems is the problem of time-optimal control. In general terms, this problem is concerned with determining the system inputs which will take the system from an initial state to a terminal state or collection of states in minimum time, subject to possible constraints. Beginning in the 1950's, modern control theory has provided insight and methods of solving this type of problem. The growth in optimal control theory was paralleled by that of computer technology which led to the establishment of computer controlled processes. This in turn led to the study of the time-optimal control problem as a problem in sampled-data or discrete-time control theory. Starting with the original paper by Kalman [KAL1], this problem has been studied by many different authors [KAL2], [CAD1], [DES1], [HO1], [TOU1], [TOU2], [GRA1], [FAR1], [SAR1]. One of the properties related to this problem is nonuniqueness of the solution. Suppose we are given a single-input n-th order linear discrete-time system. If the number of samples N required to reach the origin is greater than n, the sequence of time-optimal controls is in general not unique. This led to optimization according to an additional criterion such as minimum fuel [TOR1], [FEG1] or minimum energy [HO1], [CAN1].

An additional property of the solution of the single input discrete-time optimal control problem is that if $N \leq n$, the solution is unique. With the structure of the system fixed, this in turn implies that we are unable to optimize the system according to any additional criteria. On the other hand, it may be acceptable to operate at any of a collection of operating points; for example, in a small region about the desired state. It is shown in this thesis that in general the solution is now no longer unique when $N \leq n$. This in turn provides the opportunity to optimize the system according to additional criteria. The additional criteria studied here are minimum fuel and minimum energy.

In this thesis several discrete-time optimal control problems will be studied with two types of target sets: a hypersphere and a hyperrectangle. The problem of determining the time-optimal control for a continuous-time system (unsampled) with hyperspherical target set has been studied previously [PLA1], [PLA2] using an iterative technique. Chapter 3 is concerned with determining a sequence of discrete time-optimal controls for the case of a hyperspherical target set. Both open and closed loop formulations are obtained. A test is given to determine if the sequence is unique; if not, a sequence is chosen that minimizes the energy required to drive the system to the target set. A linear system with a single input delay is also considered as well as an open-loop stochastic system.

In Chapter IV the discrete time-optimal control problem with a hyperrectangular target set is considered. This problem is reduced to a linear programming problem. Using the properties of the Simplex Method [HAD1] of linear programming, it is possible to determine if the solution is unique. If not, a sequence is chosen which minimizes the total fuel required to drive the initial state to the target. An open-loop stochastic version of this problem is also considered; the method of solution is quite similar to that of the deterministic system.

The theory related to the problems discussed in Chapters III and IV is illustrated by several examples.

CHAPTER II

PRELIMINARY ANALYSIS AND THEORETICAL BACKGROUND

In this chapter some basic definitions and theorems that will be used in the sequel are given. Section 2.1 is concerned with certain theorems and definitions from matrix theory. Section 2.2 is devoted to the study of the degenerate system of equations $Ax = y$, and Section 2.3 discusses the concepts of controllability and observability of linear discrete-time systems. In Section 2.4 the linear programming problem is stated and the Simplex Method of solution is outlined. Some theorems of linear programming pertinent to the control problems to be discussed later are given.

2.1 Matrix Theory

Basic definitions and well-known theorems from the theory of matrices are given in this section [FRA1], [FRA2], [GAN1], [PED1], [HIL1], [AYR1]. It shall always be assumed that all matrices and vectors have only real elements.

<u>Definition</u> The <u>rank</u> of an mxn matrix A, written r(A), is the maximum number of linearly independent columns of A.

<u>Theorem 2.1.1</u> Let A and B be conformal matrices. Then $r(AB) \leq \min[r(A), r(B)]$.

<u>Theorem 2.1.2</u> If A is of order pxq and B is of order qxr, then $r(AB) \geq r(A) + r(B) - q$.

<u>Theorem 2.1.3</u> Let A be an nxn nonsingular matrix denoted by $A = (a_1, a_2, \ldots, a_n)$. If we remove s columns from A, $1 \leq s < n$, then the remaining matrix is of maximum rank.

<u>Proof</u> The proof is by induction. Remove one column a_k from A. The remaining matrix is of order nx(n-1). We need to show this matrix has rank r=n-1, that is, there exists at least one (n-1)x(n-1) minor different from zero. Suppose the opposite is true, that is, there are no (n-1)x(n-1) minors different from zero. Expanding $|A|$ along its k th column, we get

$$|A| = \sum_{i=1}^{n} a_{ik}(-1)^{i+k}|A_{ik}|$$

where a_{ik} is the i th element of the k th column of A and $|A_{ik}|$ is the corresponding (n-1)x(n-1) minor of A. By assumption we have $|A_{ik}| = 0$ for i=1,2,...,n. Hence $|A| = 0$ which is a contradiction since it was assumed that A was nonsingular.

Assume now that s columns of A have been removed and that the remaining matrix A' is of full rank, that is, A' contains an (n-s)x(n-s) nonsingular matrix M'. Let us remove the j th column from A'. We assume that there is no (n-s-1)x(n-s-1) minor that is different from zero, that is, the resulting matrix is not of full rank. Expanding the matrix M' along the j th column, we have

$$|M'| = \sum_{i=1}^{n-s} m'_{ij} (-1)^{i+j} |p'_{ij}|$$

where m'_{ij} is the i th element of the j th column of M', and $|p'_{ij}|$ is the corresponding (n-s-1)x(n-s-1) minor of M'. By assumption, all the p'_{ij} have rank less than n-s-1. Thus $|p'_{ij}| = 0$, i=1,2,...,n-s and therefore, $|M'| = 0$ which is a contradiction of the assumption that M' is nonsingular. Thus the matrix obtained by removing s+1 columns from A is also of maximum rank. By induction, the proof is complete.

QED

<u>Definition</u> The matrix A is <u>positive definite</u> if $x^TAx > 0$ for all $x \neq 0$. It is <u>positive semidefinite</u> if $x^TAx \geq 0$.

<u>Theorem 2.1.4</u> Let A be a symmetric matrix. Then A is positive definite if and only if all the eigenvalues of A are positive. A is positive semidefinite if and only if all the eigenvalues of A are nonnegative.

<u>Theorem 2.1.5</u> Let F be a nxm matrix of rank m greater than zero and A be a nxn positive definite symmetric matrix. Then $C = F^TAF$ is positive definite and symmetric.

<u>Theorem 2.1.6</u> Let A be a nxm matrix of rank n greater than zero. Then $C = AA^T$ is a symmetric and positive definite matrix.

Theorem 2.1.7 The rank of a real symmetric matrix is equal to its number of nonzero eigenvalues.

Theorem 2.1.8 Every real symmetric matrix A is orthogonally similar to a diagonal matrix whose diagonal elements are the eigenvalues of A. That is, there exists a nonsingular matrix P such that $\Lambda = P^TAP$ where $P^TP = I$, $\Lambda = \mathrm{diag}(\lambda_1 I_{r_1}, \lambda_2 I_{r_2}, \ldots, \lambda_s I_{r_s})$ and I_{r_i} is an $r_i x r_i$ identity matrix. The r_i and λ_i can be found from the expression

$$|\lambda I - A| = (\lambda - \lambda_1)^{r_1}(\lambda - \lambda_2)^{r_2}\ldots(\lambda - \lambda_s)^{r_s}$$

Theorem 2.1.9 A real symmetric matrix A of rank r is positive semidefinite if and only if there exists a matrix C of rank r such that $A = C^TC$. Similarly, A is positive definite and symmetric if and only if there exists a nonsingular matrix C such that $A = C^TC$.

Theorem 2.1.10 [KOE1], [FRA1] Let the characteristic polynomial and adjoint equation for the kxk matrix P be written as

$$c(\gamma) = |\gamma I - P| = s_0\gamma^k + s_1\gamma^{k-1} + \ldots + s_{k-1}\gamma + s_k$$

$$\mathrm{adj}(\gamma I - P) = G_0\gamma^{k-1} + G_1\gamma^{k-2} + \ldots + G_{k-2}\gamma + G_{k-1}$$

where $s_0 = 1$, $G_0 = I$. Then the scalar coefficients s_i and the matrix coefficients G_i are given recursively by the following equations:

$$\begin{array}{ll} G_0 = I & s_0 = 1 \\ G_1 = PG_0 + s_1 I & s_1 = -\,tr(PG_0) \\ G_2 = PG_1 + s_2 I & s_2 = -\frac{1}{2}tr(PG_1) \\ \vdots & \vdots \\ G_{k-1} = PG_{k-2} + s_{k-1}I & s_k = -1/k\ tr(PG_{k-1}) \\ G_k = PG_{k-1} + s_k I = 0 & \end{array}$$

Theorem 2.1.11 Let A be a real nxn symmetric matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Define $f(A) = A^k + \alpha_1 A^{k-1} + \alpha_2 A^{k-2} + \ldots + \alpha_{k-1}A + \alpha_k I$ where k is a positive integer and the α_i are scalars. The eigenvalues θ_i of f(A) are $\theta_i = f(\lambda_i) = \lambda_i^k + \alpha_1\lambda_i^{k-1} + \alpha_2\lambda_i^{k-2} + \ldots + \alpha_{k-1}\lambda_i + \alpha_k$.

Theorem 2.1.12 If A is a real symmetric matrix then A^k is also symmetric for any positive integer k.

Definition Let b be an nx1 vector and c be a mx1 vector. The outer product bc^T of b and c is defined to be the nxm matrix given by

$$bc^T = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} [c_1\ c_2 \ldots c_m] = \begin{bmatrix} b_1c_1 & b_1c_2 & \ldots & b_1c_m \\ b_2c_1 & \ldots & & \vdots \\ \vdots & & & \\ b_nc_1 & b_nc_2 & & b_nc_m \end{bmatrix}$$

That is, if $bc^T = \{a_{ij}\}$ then $a_{ij} = b_i c_j$ for i=1,2,...,n and j=1,2,...,m.

Theorem 2.1.13 The outer product of a nonzero vector y with itself is a symmetric matrix of rank 1.

Proof By definition, we have $yy^T = \{a_{ij}\}$ i=1,2,...,n and j=1,2,...,n where $a_{ij} = y_i y_j$. Therefore, $a_{ji} = y_j y_i =$

$y_i y_j = a_{ij}$ and the matrix is symmetric. To show that the matrix is of rank 1, we use the above definition of outer product to write

$$yy^T = \begin{bmatrix} y_1y_1 & y_1y_2 & \cdots & y_1y_n \\ y_2y_1 & . & . & . \\ \vdots & & & \vdots \\ y_ny_1 & y_ny_2 & \cdot\cdot & y_ny_n \end{bmatrix} = AB$$

where

$$A = \begin{bmatrix} y_1 & y_1 & \cdots & y_1 \\ y_2 & y_2 & \cdots & y_2 \\ \vdots & & & \\ y_n & y_n & \cdots & y_n \end{bmatrix} \qquad B = \begin{bmatrix} y_1 & 0 & \cdots & & 0 \\ 0 & y_2 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ & & & & 0 \\ 0 & 0 & \cdots & 0 & y_n \end{bmatrix}$$

Then by Theorem 2.1.1, rank $(yy^T) \leq$ min [rank(A), rank(B)]. By definition the rank of a matrix is the maximum number of linear independent columns in the matrix. Thus rank (A) = 1 and rank$(yy^T) \leq 1$. By assumption $y \neq 0$, so rank$(yy^T) \geq 1$, and the result follows.

QED

<u>Definition</u> Given a polynomial in λ. If the polynomial vanishes when λ is replaced by a matrix A, the polynomial is called an <u>annihilating polynomial</u> of A.

<u>Theorem 2.1.14</u> The characteristic polynomial $D(\lambda) = |\lambda I - A|$ is an annihilating polynomial of A, that is, $D(A) = 0$.

<u>Definition</u> A <u>monic polynomial</u> is a polynomial with unity as the coefficient of the highest power of λ.

<u>Definition</u> The monic annihalating polynomial $m(\lambda)$ of least degree in λ is called the <u>minimal polynomial</u> of the matrix A.

<u>Theorem 2.1.15</u> The minimal polynomial $m(\lambda)$ of a nxn matrix is given by

$$m(\lambda) = \frac{D(\lambda)}{D_{n-1}(\lambda)}$$

where $D(\lambda) = |\lambda I - A|$ and $D_{n-1}(\lambda)$ is the greatest common divisor of all the minors of order n-1 of $\lambda I - A$.

<u>Theorem 2.1.16</u> [KOE1] Let the minimal polynomial of the matrix A be given by $m(\lambda) = (\lambda - \lambda_1)^{r_1}(\lambda - \lambda_2)^{r_2}\ldots(\lambda - \lambda_k)^{r_k}$ If $f(\lambda)$ is an analytic function, then any analytic matrix function $f(A)$ can be written as

$$\begin{aligned} f(A) = {} & Z_{11}f(\lambda_1) + Z_{12}\frac{f^{(1)}(\lambda_1)}{1!} + \ldots + Z_{1r_1}\frac{f^{(r_1-1)}(\lambda_1)}{(r_1-1)!} \\ & + Z_{21}f(\lambda_2) + Z_{22}\frac{f^{(1)}(\lambda_2)}{1!} + \ldots + Z_{2r_2}\frac{f^{(r_2-1)}(\lambda_2)}{(r_2-1)!} \\ & + \ldots \\ & + Z_{k1}f(\lambda_k) + Z_{k2}\frac{f^{(1)}(\lambda_k)}{1!} + \ldots + Z_{kr_k}\frac{f^{(r_k-1)}(\lambda_k)}{(r_k-1)!} \end{aligned}$$

where the matrices Z_{ij}, i=1,2,...,k and j-1,2,...,r, called the constituent matrices, are independent of the function $f(\lambda)$.

<u>Theorem 2.1.17</u> The constituent matrices Z_{i1} (i=1,2,...,k) in Theorem 2.1.16 are idempotent and sum to unity, that is,

$$Z_{i1}^2 = Z_{i1}$$

$$Z_{11} + Z_{21} + \ldots + Z_{k1} = I$$

Furthermore, $Z_{it}Z_{jq}$ for $i \neq j$. If A is a symmetric matrix, then $Z_{ij} = 0$ for $i=1,2,\ldots,k$ and $j \neq 1$.

Theorem 2.1.18 [DER1] Let A be a square symmetric matrix with repeated eigenvalue λ_i. If the root is repeated k times, then k linearly independent columns of the P matrix in Theorem 2.1.8 can be obtained from the nonzero columns of

$$\frac{d^{k-1}}{d\lambda^{k-1}}\,[\mathrm{adj}(\ \ I - A)]\bigg|_{\lambda = \lambda_i}$$

Theorem 2.1.19 The trace of a square matrix is equal to the sum of its eigenvalues.

Definition An nxk matrix A is said to have a right inverse A^R if $AA^R = I_n$ where I_n is an nxn identity matrix.

Theorem 2.1.20 An nxk matrix A has a right inverse if and only if A is of rank n.

Theorem 2.1.21 Given the matrices A,B and C. Then

$$(A-BCB^T)^{-1} = A^{-1} + A^{-1}B(C^{-1}-B^TA^{-1}B)^{-1}B^TA^{-1}$$

where the indicated inverses are assumed to exist.

This result can be verified by post multiplying both sides by $(A-BCB^T)$ and collecting terms.

2.2 Properties of Degenerate Linear Equations [FRA1], [FRA2]

In this section the equation $y = Ax$ where A is of order mxn will be studied. This equation is called degenerate if A is either not square or not invertible. If A and y are given then either many or no solutions exist. In the case where no solution exists we may still be interested in finding a vector x_1 such that

$$|| y - Ax ||^2 = \text{minimum for } x = x_1$$

The determination of the vector x_1 is facilitated by the study of the generalized inverse of a matrix [PEN1], [GRE1], [GRE2], [DEU1], [DEU2], [ZAD1], [ACK1]. It is assumed in the following that all matrices are real.

<u>Definition</u> A <u>generalized inverse</u> of an mxn matrix A is any nxm matrix A^I which satisfies $AA^IA = A$.

<u>Theorem 2.2.1</u> If A = BC is any rank factorization of the mxn matrix A of rank r > 0, then A^I is a generalized inverse of A if and only if $CA^IB = I_r$ where I_r is a rxr identity matrix.

<u>Proof</u> [FRA1], [FRA2] If $CA^IB = I_r$, then $AA^IA = BCA^IBC = BI_rC = A$.

On the other hand, assume $AA^IA = A = BC$. The equations $B^+B = CC^+ = I_r$ are satisfied by $B^+ = (B^TB)^{-1}B^T$ and $C^+ = C^T(CC^T)^{-1}$. Thus, $CA^IB = I_rCA^IBI_r = B^+BCA^IBCC^+ = B^+AA^IAC^+ = B^+AC^+ = (B^+B)(CC^+) = I_r$.

QED

<u>Definition</u> A <u>semi-inverse</u> A^S of an mxn matrix A of rank r is any generalized inverse of the same rank r. That is, a semi-inverse is defined by the conditions $AA^SA = A$ and $\text{rank}(A^S) = \text{rank}(A)$.

<u>Definition</u> The matrix A is called idempotent if $A^2 = A$.

<u>Theorem 2.2.2</u> The matrices A^sA and AA^s are idempotent.

<u>Proof</u> $(A^sA)(A^sA) = A^s(AA^sA) = A^sA$

$(AA^s)(AA^s) = (AA^sA)A^s = AA^s$

QED

<u>Definition</u> A semi-inverse A^s of an mxn matrix A of rank r is called a <u>right pseudoinverse</u> of A if the left idempotent AA^s is symmetric, that is, $(AA^s)^T = AA^s$. Similarly, a semi-inverse is called a <u>left pseudoinverse</u> of A if the right idempotent A^sA is symmetric, that is, $(A^sA)^T = A^sA$.

<u>Theorem 2.2.3</u> There is a unique matrix A^+ called the (Moore-Penrose) <u>pseudoinverse</u> of A that satisfies the following equations

$$AA^+A = A$$

$$A^+AA^+ = A^+$$

$$(AA^+)^T = AA^+$$

$$(A^+A)^T = A^+A$$

<u>Theorem 2.2.4</u> The mxn matrix A with rank factorization $A = BC$ has for its pseudoinverse $A^+ = C^T(CC^T)^{-1}(B^TB)^{-1}B^T$ for $A \neq 0$ and $0^+ = 0^T$.

Some useful identities are given by the following theorem.

<u>Theorem 2.2.5</u> The matrix pseudoinverse has the following properties [DEU1], [DEU2]:

1) If A is nonsingular, then $A^+ = A^{-1}$ (2.2.1)

2) $(A^+)^+ = A$ (2.2.2)

3) If A = BC is a rank factorization of A, then

$$A^+ = (BC)^+ = C^+B^+ \qquad (2.2.3)$$

4) $(A^+)^TA^T = AA^+$ (2.2.4)

5) $(A^T)^+ = (A^+)^T$ (2.2.5)

Proof

1) Since $AA^+A = A$ and A^{-1} exists, then $A^+A = I$. Similarly, $AA^+ = I$. Therefore, $A^+ = A^{-1}$ by definition of A^{-1}.

2) Let $E = C^T(CC^T)^{-1}$ and $D = (B^TB)^{-1}B^T$. Then $A^+ = ED$ and

$$\begin{aligned}(A^+)^+ &= D^T(DD^T)^{-1}(E^TE)^{-1}E^T \\ &= B(B^TB)^{-1}[(B^TB)^{-1}B^TB(B^TB)^{-1}]^{-1} \\ &\quad [(CC^T)^{-1}CC^T(CC^T(CC^T)^{-1}]^{-1}\ (CC^T)^{-1}C \\ &= B(B^TB)^{-1}(B^TB)(CC^T)(CC^T)^{-1}C \\ &= BC \\ &= A\end{aligned}$$

3) Since by definition B has rank r, we can perform a rank factorization on B as in Theorem 2.2.4: B = DI where B is mxr, D is mxr and I is an rxr identity matrix. Then by Theorem 2.2.4,

$$\begin{aligned}B^+ &= I^T(II^T)^{-1}(D^TD)^{-1}D^T \\ &= (D^TD)^{-1}D^T \\ &= (B^TB)^{-1}B^T\end{aligned}$$

Similarly, $C^+ = C^T(CC^T)^{-1}$. Therefore,

$$C^+B^+ = C^T(CC^T)^{-1}(B^TB)^{-1}B^T = (BC)^+ = A^+.$$

4) Substituting the expression for A^+ given in Theorem 2.2.4 into the expression $(A^+)^TA^T$, we get

$$(A^+)^T A^T = \left[C^T(CC^T)^{-1}(B^TB)^{-1}B^T\right]^T C^TB^T$$
$$= B(B^TB)^{-1}(CC^T)^{-1}CC^TB^T$$
$$= B(B^TB)^{-1}B^T$$
$$= B\,CC^T\,(CC^T)^{-1}(B^TB)^{-1}B^T$$
$$= AA^+$$

5) $(A^T)^+ = \left[(BC)^T\right]^+ = (C^TB^T)^+ = B(B^TB)^{-1}(CC^T)^{-1}C$
$$= \left[C^T(CC^T)^{-1}(B^TB)^{-1}B^T\right]^T = (A^+)^T$$

QED

<u>Theorem 2.2.6</u> If the equation $y = Ax$ has a solution vector x_1, then every solution x has the form $x = A^Iy + x_0$ where $AA^IA = A$ and $Ax_0 = 0$.

<u>Theorem 2.2.7</u> The vectors x that minimize (for given y and A) the quantity $\|y - Ax\|^2$ have the form $x = A^Sy + x_0$ where A^S is any right pseudoinverse of A and $Ax_0 = 0$. (By definition, $\|y\|^2 = y^Ty$)

<u>Theorem 2.2.8</u> The unique vector x with $\|x\|^2$ minimum that minimizes the quantity $\|y - Ax\|^2$ is $x = A^+y$ where A^+ is the (Moore-Penrose) pseudoinverse of A.

2.3 Controllability and Observability of Sampled-Data Systems

<u>Definition</u> [BER1], [KAL2] A system is defined to be <u>completely controllable</u> if it is possible to find a sequence of controls which will drive the system from an arbitrary initial state to any desired state in a finite number of sampling periods.

Theorem 2.3.1 [BER1] Given the linear system $x[(k+1)T] = Cx(kT) + Du(kT)$, $k=0,1,\ldots,N-1$, where C is of order nxn and nonsingular and D is of order nxm. The system is completely controllable if and only if rank(G) = n where $G = [C^{n-1}D, C^{n-2}D, \ldots, D]$.

Corollary The system is completely controllable if and only if $\text{rank}(U_n) = n$, where $U_n = [C^{-1}D, C^{-2}D, \ldots, C^{-n}D]$.

Proof of Corollary By assumption, C is nonsingular. The rank of a matrix is not changed by multiplying it by a nonsingular matrix. Premultiplying G by C^{-n} gives the desired result.

QED

Definition Given a system described by $x[(k+1)T] = Cx(kT) + Du(kT)$, $y(kT) = Bx(kT)$. The state $x_i(kT)$ is said to be observable if the input $u(kT)$ and output $y(kT)$, $k=0,1,\ldots,N-1$ are sufficient to determine $x_i(0)$. If every state of the system is observable, we say that the system is completely observable.

Theorem 2.3.2 A necessary and sufficient condition for a linear n-th order system to be completely observable is that rank(H) = n where $H = [C^{N-1}B, C^{N-2}B, \ldots, B]$.

2.4 Linear Programming Theory

The Simplex Method [HAD1] of solving linear programming problems is outlined in this section. Theorems concerned with existence and uniqueness of solutions that are pertinent to the problem to be solved in Chapter IV are given here.

The basic problem to be solved in a linear programming problem is to maximize a linear objective function

$$z = \sum_{j=1}^{\ell} c_j x_j \qquad (2.4.1)$$

subject to m linear inequalities (or equalities) of the form

$$\sum_{j=1}^{\ell} a_{ij} x_j \{\leqslant = \geqslant\} b_i \quad i=1,\ldots,m \qquad (2.4.2)$$

and the non-negativity restrictions

$$x_j \geqslant 0 \qquad j=1,2,\ldots,\ell \qquad (2.4.3)$$

All the b_i must be non-negative. If required, we can multiply an inequality by -1 to obtain $b_i \geqslant 0$. The next step in solving the problem is to add slack or surplus variables to convert an inequality to an equality. Thus

$$\sum_{j=1}^{\ell} a_{ij} x_j \leqslant b_i$$

becomes

$$\sum_{j=1}^{\ell} a_{ij} x_j + x_{r+i} = b_i \qquad x_{r+i} \geqslant 0$$

and

$$\sum_{j=1}^{\ell} a_{ij} x_j \geqslant b_i$$

becomes

$$\sum_{j=1}^{\ell} a_{ij} x_j - x_{r+i} = b_i \qquad x_{r+i} \geqslant 0$$

By defining $A = \{a_{ij}\}$, $b = (b_1, b_2 \ldots, b_m)^T$, and $c = (c_1, c_2, \ldots, c_\ell)$, the above problem can be written as

$$\max z = cX$$

subject to $AX = b$

$$X \geq 0$$

where $X = (x_1, x_2 \ldots, x_k)^T$.

<u>Definition</u> Given a system of m simultaneous linear equations in n unknowns, $AX = b$ $(m<n)$ and rank of A equal to m. If any mxm nonsingular matrix is chosen from A, and if all the n-m variables not associated with the columns of this matrix are set equal to zero, the solution to the resulting system of equations is called a <u>basic solution</u>. The m variables which can be different from zero are called <u>basic variables</u>.

<u>Definition</u> Any set of x_j which satisfy the set of constraints given by (2.4.2) is called a <u>solution</u> to the linear programming problem. Any solution which satisfies the non-negativity restrictions is called a <u>feasible solution</u>. Any feasible solution which maximizes the value of z is called an <u>optimal feasible solution</u>.

The problem is to determine an optimal feasible solution. The steps in solving the problem by means of the Simplex Method are now outlined [HAD1].

1) Construct an initial tableau such as that given in reference [HAD1] or as shown in Section 4.3. The basis vector x_B in this case is given by $x_B = b$. Except for the last row the other columns y_j in the tableau are given by $y_j = a_j$ where the a_j are the columns of the A

matrix. The last row of the tableau gives z and $z_j - c_j$ where

$$z = c_B b \text{ and } z_j - c_j = c_B a_j - c_j \qquad (2.4.4)$$

c_B is a row vector composed of those c_j corresponding to basis variables. The c_j are sometimes called "prices."

2) The <u>optimality criterion</u> is then used. This criterion states that if all $z_j - c_j \geq 0$ then a basic feasible solution is optimal. If one or more $z_j - c_j < 0$ then the problem is not solved and we proceed to the next step.

3) Compute $z_k - c_k = \min_j (z_j - c_j)$ for $z_j - c_j < 0$ (2.4.5)
The criterion implies that we add a_k to the basis. (If there is a tie we may choose either one.) Once a_k has been chosen there are two possibilities depending on y_{ik} where y_{ik} are the elements of the y_k vector:

a) If $y_{ik} \leq 0$ for all i then there is an unbounded solution involving the vectors in the basis and a_k.

b) If $y_{ik} > 0$ for at least one i then a new basis feasible solution can be found having $\hat{z} \geq z$.

4) If at least one $y_{ik} > 0$, then determine which vector is to leave the basis. This vector is chosen by the following criterion. Compute

$$\frac{x_{Br}}{y_{rk}} = \min_i \left\{ \frac{x_{Bi}}{y_{ik}} \;,\; y_{ik} > 0 \right\} \qquad (2.4.6)$$

The vector in column r of the basis is removed and replaced by a_k. If there is a tie, any one of the tied columns can be removed and replaced by a_k.

5) The next step is to construct a new tableau. The elements $\hat{y}_{ij}$ of the new tableau are related to those y_{ij} of the old tableau by the relations

$$\hat{y}_{ij} = y_{ij} - \frac{y_{ik}}{y_{rk}} y_{rj} \quad \text{all } j,\ i=1,\dots,m+1\ i \neq r \qquad (2.4.7)$$

$$\hat{y}_{rj} = \frac{y_{rj}}{y_{rk}} \quad \text{all } j \qquad (2.4.8)$$

$$x_B = y_0,\ z=y_{m+1,0},\ z_j - c_j = y_{m+1,j}\ \ j=1,\dots,n \qquad (2.4.9)$$

The price in the rth position of the column headed "c_B" should be replaced by c_k and the vector in the rth position under "vector in basis" should be replaced by a_k.

6) Return to Step 3. The Simplex Method requires a finite number of steps to reach an optimal (or unbounded) solution. In general, the number of iterations required to reach an optimal solution lies between m and 2m where m is the number of constraints.

Let x_{B_i} be the component of x_B corresponding to column b_i of the basis. If $x_{B_i} > 0$, we say that b_i is in the basis at a positive level and if $x_{B_i} = 0$, b_i is in the basis at a zero level. The following theorems will be used in Sections 4.3 and 4.4.

<u>Theorem 2.4.1</u> [HAD1]

a) If no artificial vectors appear in the basis and the optimality criterion is satisfied, then the solution is an optimal basic feasible solution to the given problem. The constraint equations are consistent, and there is no redundancy in the constraint equations.

b) If one or more artificial vectors appear in the basis at a zero level and the optimality criterion is satisfied, then the solution is an optimal solution to the given problem. The constraint equations are consistent, but in this case redundancy may exist in the constraints.

c) If one or more artificial vectors appear in the basis at a positive level and the optimality criterion is satisfied, the original problem has no feasible solution. There may be no feasible solution either because the constraint equations are inconsistent, or because there are solutions, but none is feasible.

<u>Definition</u> A basic solution to $AX = b$ is <u>degenerate</u> if one or more of the basic variables vanish.

<u>Theorem 2.4.2</u> [HAD1]

a) If the optimal solution represented by the last tableau is not degenerate and if $z_j - c_j > 0$ for each a_j not in the basis, then the optimal basic feasible solution is unique. No vector can be inserted into the basis without decreasing the value of the objective function.

b) When $z_j - c_j = 0$ for one or more a_j not in the basis, any such vector a_j can be inserted to yield a different solution if $y_{ij} > 0$ for at least one i and $\min(x_B/y_{ij})$, $y_{ij} > 0$ is positive. If a_j enters at a zero level we do not obtain a different solution; the result is only a different representation of the same degenerate extreme point.

CHAPTER III

TIME-OPTIMAL CONTROL WITH HYPERSPHERICAL TARGET SET

3.1 Statement of Control Problem and Basic Assumptions

Given a linear time-invariant system described by

$$\dot{x} = Ax + Bu \qquad (3.1.1)$$

$$x(0) = x_0$$

where $x(t)$ is a nx1 vector
$u(t)$ is a mx1 vector
A is a nxn matrix
B is a nxm matrix

It is assumed that $u(t)$ is of the sampled-data type so that

$$u(t) = u(kT) \text{ for } kT \leqslant t < (k+1)T \qquad (3.1.2)$$

where T is the sampling period. We wish to determine the fewest number of samples $\bar{N}$ and a corresponding sequence of controls $u(kT)$, $k=0,1,\ldots,\bar{N}-1$ which drive the system from the initial state x_0 to the target set described by

$$G = \{x(\bar{N}T) : x^T(\bar{N}T)x(\bar{N}T) \leqslant R^2\} \qquad (3.1.3)$$

where R is a real number and $x_0 \notin G$. If the solution is not unique, we want to determine a sequence of controls that drives the system to the set G with minimum energy. That is, we want to minimize J where

$$J = \sum_{k=0}^{\bar{N}-1} u^T(kT)u(kT) \qquad (3.1.4)$$

The problem shall be solved under the following two assumptions wherein N denotes a running variable.

Assumption 1 It is assumed that the discrete-time system is completely controllable. From Theorem 2.3.1 this means that $\text{rank}[D,CD,\ldots,C^{n-1}D] = n$ where C and D are defined in equation 3.2.1.

Assumption 2 It is assumed that $\text{rank}[D,CD,\ldots,C^{N-1}D]$ = maximum for all $N > 0$.

The second assumption does not follow automatically from the first. This is shown by the following example. Let

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad D = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$$

Then n=3, m=2, and n/m = 3/2. We then have

$$[D,CD,C^2D] = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 4 & 4 & 8 & 8 \\ 3 & 9 & 9 & 9 & 27 & 27 \end{bmatrix}$$

The determinant of the matrix formed by the first, third, and fifth columns of the above matrix is twelve. Thus, $\text{rank}[D,CD,C^2D] = 3$, and the system is completely controllable. If we choose $N = 1 < n/m$, then $\text{rank}[D,CD,\ldots,C^{N-1}D] = \text{rank}(D) = 1 \neq Nm$. Thus the first assumption is satisfied but not the second.

If we let Assumption 1 hold, then Assumption 2 is the same as Assumption 1 under any of the following conditions:

1) The system has a single input, that is, $u(t)$ is a scalar.

2) $N = n/m$ is an integer and $|D,CD,\ldots,C^{N-1}D| \neq 0$. In other words, the first n columns of the controllability matrix $[D,CD,\ldots,C^{n-1}D]$ are linearly independent.

3) $n/m < 2$ and D is of full rank.

Condition 1) is a special case of condition 2). This follows from the fact that for a scalar controller, $[D,CD,\ldots,C^{n-1}D]$ is a nxn matrix and must be nonsingular in order that the system be completely controllable. To show that condition 2) implies that Assumption 2 is the same as Assumption 1, we use Theorem 2.1.3. Under condition 2 the matrix $[D,CD,\ldots,C^{N-1}D]$ is nonsingular. From Theorem 2.1.3, if we remove columns from the nxNm matrix $[D,CD,\ldots,C^{N-1}D]$, the remaining matrix is of maximum rank, that is, rank $[D,CD,\ldots,C^{N-1}D] = Nm$ when $N \leq n/m$. Let condition 3) hold. If we want $N \leq n/m$, we must have $N=1$, and $\text{rank}[D,CD,\ldots,C^{N-1}D] = \text{rank}(D) = Nm$ because $n > m$ if $N < n/m$.

3.2 Solution of Control Problem

The general solution to equation (3.1.1) is

$$x(t) = \phi(t-t_0)x(t_0) + \int_{t_0}^{t} \phi(t-\tau)Bu(\tau)d\tau$$

where $\phi(t)$ is the transition matrix of the system described by equation (3.1.1) [DER1]. Let $t=(k+1)T$ and $t_0 = kT$. Then

$$x[(k+1)T] = \phi(T)x(kT) + \int_{kT}^{(k+1)T} \phi[(k+1)T-\tau]Bu(\tau)\, d\tau$$

From equation (3.1.2) this becomes

$$x[(k+1)T] = \phi(T)x(kT) + \int_{kT}^{(K+1)T} \phi[(k+1)T-\tau]Bd\tau u(kT)$$

If we let $\gamma = \tau - kT$, then

$$x[(k+1)T] = \phi(T)x(kT) + \int_0^T \phi(T-\gamma)Bd\gamma u(kT)$$

or

$$x[(k+1)T] = Cx(kT) + Du(kT) \qquad (3.2.1)$$

where

$$D = \int_0^T \phi(T-\tau)Bd\tau \qquad \text{(nxm matrix)} \qquad (3.2.2)$$

and $C = \phi(T)$ is nonsingular by the properties of the transition matrix [ATH1].

By assuming k=0 and then k=1, we obtain respectively, $x(T) = Cx(0) + Du(0)$, $x(2T) = Cx(T) + Du(T)$. Thus, $x(2T) = C^2x(0) + CDu(0) + Du(T)$. Repeating this argument for k=3,4,...,N-1, we arrive at the general solution to equation (3.2.1):

$$x(kT) = C^k x(0) + \sum_{i=0}^{k-1} C^{k-1-i}Du(iT) \qquad (3.2.3)$$

If we define

$$F_i = - C^{-(i+1)}D \qquad i=0,1,\ldots,k-1 \qquad (3.2.4)$$

and

$$u_i = u(iT) \qquad i=0,1,\ldots,k-1$$

then equation (3.2.3) becomes

$$x(kT) = C^k x(0) - C^k \sum_{i=0}^{k-1} F_i u_i \tag{3.2.5}$$

It will be shown that if there exists a sequence of controls which drives the initial state to the interior of the target, there exists a sequence of controls which drives the initial state to the boundary of the target in the same number of samples or fewer. The problem will then be restricted to determining a sequence of controls such that

$$x^T(NT)x(NT) = R^2 \tag{3.2.6}$$

Substituting equation (3.2.5) into (3.2.6), we get

$$x^T(NT)x(NT) = x_0^T \psi_N x_0 - 2x_0^T \psi_N \sum_{j=0}^{N-1} F_j u_j$$

$$+ \left(\sum_{j=0}^{N-1} F_j u_j\right)^T \psi_N \left(\sum_{i=0}^{N-1} F_i u_i\right)$$

where

$$\psi_N = (C^N)^T C^N \tag{3.2.7}$$

This can be rewritten as

$$x^T(NT)x(NT) = U^T Q_N U - 2d_N^T U + x_0^T \psi_N x_0 \tag{3.2.8}$$

where $\quad U = (u_0, u_1, \ldots, u_{N-1})^T \quad$ (mNx1 vector) $\quad$ (3.2.9)

$$Q_N = \overline{F}_N^T \psi_N \overline{F}_N \qquad \text{(mNxmN matrix)} \tag{3.2.10}$$

$$\overline{F}_N = [F_0, F_1, \ldots, F_{N-1}] \qquad \text{(nxmN matrix)} \tag{3.2.11}$$

$$d_N^T = x_0^T \psi_N \overline{F}_N \qquad \text{(1xmN vector)} \tag{3.2.12}$$

By defining

$$e_N = x_0^T \psi_N x_0 - R^2 \qquad (3.2.13)$$

equations (3.2.6) and (3.2.8) combine to give

$$f(U) = U^T Q_N U - 2d_N^T U + e_N = 0 \qquad (3.2.14)$$

Theorem 3.2.1 Q_N is a symmetric matrix with the following properties:

1) If $N \leqslant n/m$, Q_N is positive definite.

2) If $N > n/m$, Q_N is positive semidefinite and singular.

Proof By Theorem 2.1.6 we know that ψ_N is symmetric and positive definite. Thus by equation (3.2.10), $Q_N^T = (\overline{F}_N^T \psi_N \overline{F}_N)^T = \overline{F}_N^T \psi_N^T \overline{F}_N = \overline{F}_N^T \psi_N \overline{F}_N = Q_N$, and Q_N is symmetric.

1) Assume that $n \geqslant Nm$. Since $\overline{F}_N$ is of order nxNm, it follows from Assumption 2 of Section 3.1 that the rank of $\overline{F}_N$ is Nm. Thus $\overline{F}_N$ is of maximum rank and $Q_N = \overline{F}_N^T \psi_N \overline{F}_N$ is positive definite by Theorem 2.1.5.

2) If $n < Nm$, it follows that Q_N is positive semidefinite since $U^T Q_N U = U^T \overline{F}_N^T \psi_N \overline{F}_N U = Y^T \psi_N Y$ where $Y = \overline{F}_N U$. Because ψ_N is positive definite, $U^T Q_N U = Y^T \psi_N Y \geqslant 0$. Q_N is not positive definite because it is singular. This follows from Theorem 2.1.1, that is, $\text{rank}(Q_N) \leqslant \min [\text{rank}(\overline{F}_N), \text{rank}(\psi_N)] < mN$. On the other hand, Q_N is of order mNxmN. Hence Q_N is singular.

QED

Theorem 3.2.2 a) If $N \leqslant n/m$, the expression for f(U) given in equation (3.2.14) can be written as

$$Y^T\Lambda_N Y + g_N = 0 \qquad (3.2.15)$$

where

$$g_N = e_N - d_N^T Q_N^{-1} d_N \qquad (3.2.16)$$

by the transformation

$$U = P_N Y + Q_N^{-1} d_N \qquad (3.2.17)$$

where P_N is such that $P_N^T P_N = I$, and

$$\Lambda_N = P_N^T Q_N P_N \qquad (3.2.18)$$

is a diagonal matrix with diagonal elements equal to the eigenvalues of Q_N.

b) If N=n/m = integer, equation (3.2.15) reduces to

$$Y^T\Lambda_N Y - R^2 = 0 \qquad (3.2.19)$$

<u>Proof</u> a) By Theorem 3.2.1, Q_N^{-1} always exists for $N \leq n/m$. Substituting equation (3.2.17) into (3.2.14), we obtain

$$(P_N Y + Q_N^{-1} d_N)^T Q_N (P_N Y + Q_N^{-1} d_N) - 2d_N^T (P_N Y + Q_N^{-1} d_N) + e_N$$

$$= Y^T P_N^T Q_N P_N Y - d_N^T Q_N^{-1} d_N + e_N = 0 \qquad (3.2.20)$$

By assumption P_N is chosen to diagonalize Q_N. Such a matrix exists by Theorem 2.1.8. Thus $\Lambda_N = P_N^T Q_N P_N$ is a diagonal matrix with the eigenvalues of Q_N along the diagonal, and equation (3.2.20) becomes

$$Y^T\Lambda_N Y - d_N^T Q_N^{-1} d_N + e_N = 0 \qquad (3.2.21)$$

By equations (3.2.16) and (3.2.21) the first part of the theorem follows.

b) If $N = n/m$ = integer, then $\bar{F}_N$ given by equation (3.2.11) is nonsingular by Assumption 2. Then by equations (3.2.10)-(3.2.14)

$$g_N = e_N - d_N^T Q_N^{-1} d_N = e_N - x_0^T \psi_N \bar{F}_N (\bar{F}_N^T \psi_N \bar{F}_N)^{-1} \bar{F}_N^T \psi_N x_0$$

$$= e_N - x_0^T \psi_N \psi_N^{-1} \psi_N x_0 = x_0^T \psi_N x_0 - R^2 - x_0^T \psi_N x_0 = -R^2 \qquad (3.2.22)$$

Substituting equation (3.2.22) into (3.2.15) gives the second part of the theorem.

QED

<u>Theorem 3.2.3</u> Let $N \leq n/m$. Then the following is true.

a) If $e_N = d_N^T Q_N^{-1} d_N$, the unique sequence of controls such that $x(NT) \varepsilon \partial G$ (the boundary of G) is given by $U = Q_N^{-1} d_N$.

b) If $e_N < d_N^T Q_N^{-1} d_N$, a nonunique sequence of controls exists such that $x(NT) \varepsilon G$.

c) If $e_N > d_N^T Q_N^{-1} d_N$, no sequence of controls exists such that $x(NT) \varepsilon G$.

<u>Proof</u> a) If $e_N = d_N^T Q_N^{-1} d_N$, then by equations (3.2.15) and (3.2.16)

$$Y^T \Lambda_N Y = 0 \qquad (3.2.23)$$

The matrix Λ_N is positive definite since it contains the eigenvalues of Q_N (which must be positive by Theorem 2.1.4) along its diagonal. By definition of positive definiteness, equation (3.2.23) holds only if $Y = 0$. Thus, by equation (3.2.17) it follows that $U = Q_N^{-1} d_N$ is the unique sequence of controls.

b) If $e_N < d_N^T Q_N^{-1} d_N$, then by equations (3.2.15) and (3.2.16)

$$Y^T \Lambda_N Y = d_N^T Q_N^{-1} d_N - e_N > 0 \qquad (3.2.24)$$

If N > 1, equation (3.2.24) has an infinite number of solutions. If N=1, there are two solutions. In either case a solution exists but is not unique.

c) If $e_N > d_N^T Q_N^{-1} d_N$, then be equations (3.2.15) and (3.2.16)

$$Y^T \Lambda_N Y = d_N^T Q_N^{-1} d_N - e_N < 0 \qquad (3.2.25)$$

Since Λ_N is positive definite, there is no value of Y which satisfies (3.2.25). Thus, no solution exists for this value of N.

QED

The previous two theorems were for the case when $N \leqslant n/m$. We now consider the case when $N > n/m$.

Theorem 3.2.4 a) If $N > n/m$, a nonunique sequence of controls exists such that $x(NT) \in \partial G$ (the boundary of G). A sequence of controls can be found from

$$U = P_N Y + \bar{F}_N^T (\bar{F}_N \bar{F}_N^T)^{-1} x_0 \qquad (3.2.26)$$

where

$$Y^T \Lambda_N Y = R^2 \qquad (3.2.27)$$

and P_N is chosen such that $P_N^T P_N = I$, and $\Lambda_N = P_N^T Q_N P_N$ is a diagonal matrix with diagonal elements equal to the eigenvalues of Q_N.

b) If R=0, $U = \bar{F}_N^T(\bar{F}_N\bar{F}_N^T)^{-1}x_0$ is the sequence of controls which require minimum energy to reach the origin.

<u>Proof</u> From equations (3.2.10), (3.2.12) and (3.2.14)

$$f(U) = U^T\bar{F}_N^T(C^N)^T C^N\bar{F}_N U - 2x_0^T\psi_N\bar{F}_N U + e_N$$

$$= \|C^N\bar{F}_N U - C^N x_0\|^2 - x_0^T\psi_N x_0 + e_N$$

where $\|y\|^2 = y^T y$. From equation (3.2.13) the expression for f(U) becomes

$$f(U) = \|C^N\bar{F}_N U - C^N x_0\|^2 - R^2 \qquad (3.2.28)$$

Minimizing f(U) with respect to U is equivalent to minimizing h(U) where

$$h(U) = \|C^N\bar{F}_N U - C^N x_0\|^2 \qquad (3.2.29)$$

By Theorems 2.2.7 and 2.2.8, we know that a value of U that minimizes h(U) is given by $U = (C^N\bar{F}_N)^+C^N x_0$. The matrix C^N is of order nxn and of rank n while $\bar{F}_N$ is of order nxNm and of rank n by Assumption 2 and the fact that $n < Nm$. Thus C^N and $\bar{F}_N$ represent a rank factorization of $C^N\bar{F}_N$. By equation (2.2.3) we then have

$$U = \bar{F}_N^+(C^N)^+C^N x_0 = \bar{F}_N^+ x_0 \qquad (3.2.30)$$

Returning to the expression for f(U) given by equation (3.2.14), if we set

$$U = P_N Y + \bar{F}_N^+ x_0 \qquad (3.2.31)$$

then the expression for f(U) becomes

$$L(Y) = (P_N Y + \overline{F}_N^+ x_0)^T Q_N (P_N Y + \overline{F}_N^+ x_0) - 2d_N^T (P_N Y + \overline{F}_N^+ x_0) + e_N$$

$$= Y^T P_N^T Q_N P_N Y + Y^T P_N^T Q_N \overline{F}_N^+ x_0 + x_0^T (\overline{F}_N^+)^T Q_N P_N Y$$

$$+ x_0^T (\overline{F}_N^+)^T Q_N \overline{F}_N^+ x_0 - 2d_N^T P_N Y - 2d_N^T \overline{F}_N^+ x_0 + e_N$$

By using equations (2.2.5) and (3.2.12), this becomes

$$L(Y) = Y^T P_N^T Q_N P_N Y + 2x_0^T (\overline{F}_N^+)^T \overline{F}_N^+ \psi_N \overline{F}_N P_N Y - 2x_0^T \psi_N \overline{F}_N P_N Y$$

$$- 2x_0^T \psi_N \overline{F}_N \overline{F}_N^+ x_0 + x_0^T (\overline{F}_N^+)^T Q_N \overline{F}_N^+ x_0 + e_N \qquad (3.2.32)$$

L(Y) can be simplified. For notational convenience, let

$$S = C^N \qquad (3.2.33)$$

Using equation (3.2.10), the second to last term on the right side of equation (3.2.32) then becomes

$$x_0^T (\overline{F}_N^+)^T Q_N \overline{F}_N^+ x_0 = x_0^T (\overline{F}_N^+)^T \overline{F}_N^T S^T S^T S \overline{F}_N \overline{F}_N^+ x_0$$

$$= x_0^T S^T (\overline{F}_N^+ S^{-1})^T (S\overline{F}_N)^T S\overline{F}_N \overline{F}_N^+ S^{-1} S\, x_0$$

$$= x_0^T S^T (\overline{F}_N^+ S^{-1})^T (S\overline{F}_N)^T S\overline{F}_N (S\overline{F}_N)^+ S x_0$$

$$= x_0^T S^T [(S\overline{F}_N)^+]^T (S\overline{F}_N)^T S\overline{F}_N (S\overline{F}_N)^+ S x_0$$

$$= x_0^T S^T S\overline{F}_N (S\overline{F}_N)^+ S\overline{F}_N (S\overline{F}_N)^+ S x_0 \quad \text{by equation (2.2.4)}$$

$$= x_0^T S^T S\overline{F}_N (S\overline{F}_N)^+ S x_0 \quad \text{by Theorem 2.2.3} \qquad (3.2.34)$$

By Theorem 2.2.4, equation (3.2.34) becomes

$$x_0^T (F_N^+)^T Q_N F_N^+ x_0 = x_0^T S^T S\overline{F}_N \overline{F}_N^T (\overline{F}_N \overline{F}_N^T)^{-1} (S^T S)^{-1} S^T S x_0$$

$$= x_0^T S^T S x_0$$

$$= x_0^T\psi_N x_0 \quad \text{by equations (3.2.33) and (3.2.7)}$$

(3.2.35)

The second term on the right side of equation (3.2.32) can be written as

$$2x_0^T(\bar{F}_N^+)^T\bar{F}_N^T\psi_N\bar{F}_N P_N Y = 2x_0^T S^T(S^T)^{-1}(\bar{F}_N^T)^+\bar{F}_N^T\psi_N\bar{F}_N P_N Y$$

$$= 2x_0^T S^T[(S\bar{F}_N)^+]^T\bar{F}_N^T S^T S\bar{F}_N P_N Y$$

$$= 2x_0^T S^T[(S\bar{F}_N)^T]^+(S\bar{F}_N)^T S\bar{F}_N P_N Y$$

$$= 2x_0^T S^T(S\bar{F}_N)(S\bar{F}_N)^+(S\bar{F}_N)P_N Y$$

by equation (2.2.4).

Using Theorem 2.2.3 in the above equation, we then have

$$2x_0^T(\bar{F}_N^+)^T\bar{F}_N^T\psi_N\bar{F}_N P_N Y = 2x_0^T S^T S\bar{F}_N P_N Y$$

$$= 2x_0^T\psi_N\bar{F}_N P_N Y \qquad (3.2.36)$$

The third to last term on the right side of equation (3.2.32) can be written as

$$2x_0^T\psi_N\bar{F}_N\bar{F}_N^+x_0 = 2x_0^T S^T S\bar{F}_N(S\bar{F}_N)^+Sx_0$$

$$= 2x_0^T(\bar{F}_N^+)^T Q_N\bar{F}_N^+x_0 \quad \text{by equation (3.2.34)}$$

$$= 2x_0^T\psi_N x_0 \quad \text{by equation (3.2.35)} \qquad (3.2.37)$$

Substituting equations (3.2.35)-(3.2.37) into (3.2.32) gives

$$L(Y) = Y^T P_N^T Q_N P_N Y - x_0^T\psi_N x_0 + e_N$$

Using equation (3.2.13), this becomes $L(Y) = Y^T P_N^T Q_N P_N Y - R^2$. Since $f(U) = 0$, it follows that $L(Y) = 0$, and thus

$$Y^T P_N^T Q_N P_N Y = R^2 \tag{3.2.38}$$

We can choose P_N such that $P_N^T P_N = I$ and $\Lambda_N = P_N^T Q_N P_N$ ia a diagonal matrix with the eigenvalues of Q_N along the diagonal.

The expression for U given in equation (3.2.31) can be put in a form not involving the pseudoinverse.

$$\begin{aligned} \overline{F}_N^+ &= (S\overline{F}_N)^+ S \\ &= \overline{F}_N^T (\overline{F}_N \overline{F}_N^T)^{-1} (S^T S)^{-1} S^T S \quad \text{by Theorem 2.2.4} \\ &= \overline{F}_N^T (\overline{F}_N \overline{F}_N^T)^{-1} \end{aligned} \tag{3.2.39}$$

Substituting equation (3.2.39) into (3.2.31) gives equation (3.2.26). Equation (3.2.38) has an infinite number of solutions since Λ_N is positive semidefinite and singular.

b) If $R = 0$, then by equations (3.2.28)-(3.2.29) it follows that $f(U) = h(U) = 0$. By Theorem 2.2.8, the value of U given by equation (3.2.30) is the value of U that minimizes $h(U)$ with the additional property that $\|U\|^2$ is minimum. Moreover, this value of U is unique because $\overline{F}_N^+$ is unique. By equations (3.2.30) and (3.2.39), we have that $U = \overline{F}_N^T (\overline{F}_N \overline{F}_N^T)^{-1} x_0$, and the second part of the theorem is proven. This results when $R = 0$ has been derived by a different method [CAD1], [CAD2].

QED

From Theorem 3.2.2b), a value of Y satisfying (3.2.19) and (3.2.17) always exists. This fact plus Theorem 3.2.4a) imply that the target can always be reached in $\bar{N}$ samples where $\bar{N}$ is the smallest integer greater than or equal to n/m. The following theorem shows that the boundary of the target can be reached in the same number or fewer number of samples than that required to reach the interior of the target set.

Theorem 3.2.5 Assume a sequence of controls exists such that $x^T(N_1T)x(N_1T) < R^2$ where $R > 0$. Then another sequence of controls exists such that $x^T(N_2T)x(N_2T) = R^2$ where $N_2 \leq N_1$.

Proof Assume the hypotheses of the theorem to be true. Furthermore, let $N \leq n/m$. Then by (3.2.8), $U^TQ_NU - 2d_N^TU + x_0^T\psi_Nx_0 < R^2$, or by using equation (3.2.13), $U^TQ_NU - 2d_N^TU + e_N < 0$. If we let $U = U_1 + Q_N^{-1}d_N$, then this inequality becomes $U_1^TQ_NU_1 - d_N^TQ_N^{-1}d_N + e_N < 0$ which implies that

$$e_N < d_N^TQ_N^{-1}d_N$$

since Q_N is positive definite. By theorem 3.2.3 this inequality is just the condition for the existence of a sequence of time-optimal controls to reach the boundary of the target set. Thus the boundary can always be reached in the same number of samples as that required to reach the interior of the target. On the other hand, by choosing an initial state close to the boundary of the target it is possible to construct examples that require fewer samples to reach the boundary than a point in the interior of the target.

If we suppose that the interior of the target can be reached in $N > n/m$, then by Theorem 3.2.4 we know that the boundary of the target can also be reached in N samples. Hence, for all N it follows that the boundary of the target can be reached in the same or fewer number of samples than that required to reach the interior of the target.

QED

The theorems presented above give a straightforward method of determining the minimum number of samples $\bar{N}$ required to reach the target and a sequence of time-optimal controls. It is noted that the value of $\bar{N}$ can be determined without diagonalizing the $Q_{\bar{N}}$ matrix, that is, $P_{\bar{N}}$ and $\Lambda_{\bar{N}}$ need not be computed. The procedure for determining $\bar{N}$ is illustrated in Figure 3.2.1.

Except for the special cases when a unique solution exists or R=0, it is necessary to use equation (3.2.15) if $\bar{N} \leq n/m$ or equation (3.2.27) if $\bar{N} > n/m$ to determine a sequence of time-optimal controls. Similarly, it is necessary to determine the diagonalizing matrix $P_{\bar{N}}$. For purposes of hand computation the eigenvalues of $Q_{\bar{N}}$ which form the diagonal elements of $\Lambda_{\bar{N}}$ can be found from the characteristic equation for $Q_{\bar{N}}$. For the case when $Q_{\bar{N}}$ has distinct eigenvalues, the columns of $P_{\bar{N}}$ can be chosen

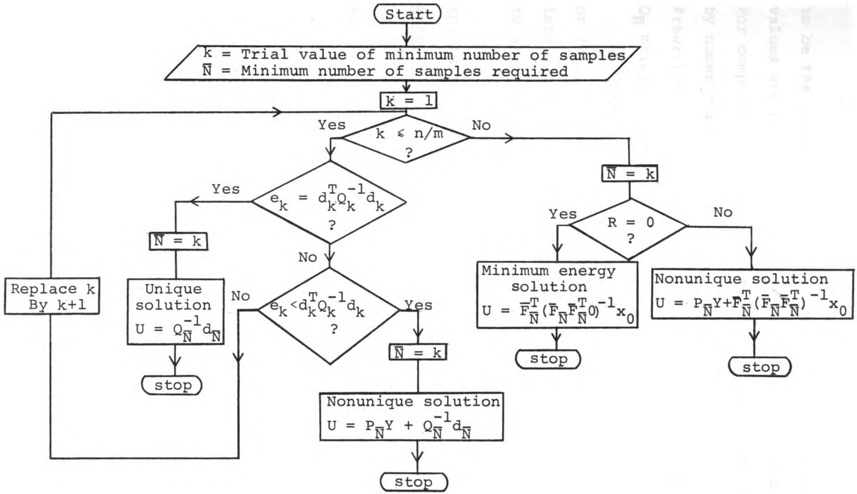


Fig. 3.2.1 Flowchart for determining minimum number of samples for open-loop system.

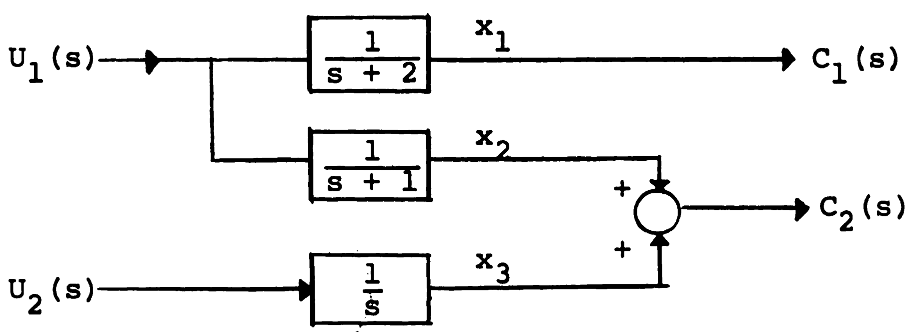
to be the normalized eigenvectors of $Q_{\bar{N}}$. If the eigenvalues are not distinct, other methods can be used [DER1]. For computer computation the $P_{\bar{N}}$ and $\Lambda_{\bar{N}}$ matrix can be found by numerical methods based on the theory of Jacobi [RAL1]. Prewritten FORTRAN programs exist for diagonalizing the $Q_{\bar{N}}$ matrix, that is, for determining $P_{\bar{N}}$ and $\Lambda_{\bar{N}}$ [KUO1], [IBM1].

Once $\Lambda_{\bar{N}}$ is known, a solution of equation (3.2.15) or (3.2.27) is easily found. The result is substituted into equation (3.2.17) if $\bar{N} \leq n/m$ or (3.2.26) if $\bar{N} > n/m$ to give a sequence of time optimal controls.

The theory presented thus far is illustrated by the following example.

<u>Example 3.2.1</u>. Consider the two input systems described by the following transfer function matrix and corresponding block diagram.

$$\begin{bmatrix} C_1(s) \\ C_2(s) \end{bmatrix} \quad \begin{bmatrix} \frac{1}{s+2} & 0 \\ \frac{1}{s+1} & \frac{1}{s} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix}$$



It is assumed that the control signals $u_1(t)$ and $u_2(t)$ are the output of zero-hold devices. The sampling period is T=1 second. The differential equations corresponding to (3.2.40) are

$$\frac{dc_1}{dt} + 2c_1(t) = u_1(t) \qquad (3.2.41)$$

$$\frac{d^2c_2}{dt^2} + \frac{dc_2}{dt} = \frac{du_1(t)}{dt} + \frac{du_2(t)}{dt} + u_2(t) \qquad (3.2.42)$$

By defining

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad (3.2.43)$$

and by the above figure

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \qquad (3.2.44)$$

We want to find a sequence of controls which drive the system from the initial state $x(0) = (10 \quad -10 \quad 10)^T$ to the target $x^T(\bar{N}T)x(\bar{N}T) \leq 1$ in the fewest number of samples, $\bar{N}$.

<u>Solution</u> The discrete-time system corresponding to 3.2.44) is

$$x(k+1) = Cx(k) + Du(k) \qquad (3.2.45)$$

where

$$C = \begin{bmatrix} e^{-2} & 0 & 0 \\ 0 & e^{-1} & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad (3.2.46)$$

$$D = \begin{bmatrix} (1-e^{-2})/2 & 0 \\ 1-e^{-1} & 0 \\ 0 & 1 \end{bmatrix} \tag{3.2.47}$$

From (3.2.4), and (3.2.36)-(3.2.37)

$$F_0 = -C^{-1}D = \begin{bmatrix} -3.19453 & 0 \\ -1.71828 & 0 \\ 0 & -1 \end{bmatrix} \tag{3.2.48}$$

$$F_1 = -C^{-2}D = \begin{bmatrix} -23.6045 & 0 \\ -4.6708 & 0 \\ 0 & -1 \end{bmatrix} \tag{3.2.49}$$

Setting N=1, we find that

$$\psi_1 = C^TC = \begin{bmatrix} 0.018315 & 0 & 0 \\ 0 & 0.135335 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2.50}$$

$$Q_1 = \bar{F}_1^T\psi_1\bar{F}_1 = F_0^T\psi_1F_0 = \begin{bmatrix} 0.58649 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.2.51}$$

$$d_1^T = x_0^T\psi_1\bar{F}_1 = x_0^T\psi_1F_0 = (1.74034 \quad -10.0000) \tag{3.2.52}$$

$$e_1 = x_0^T\psi_1x_0 - R^2 = 115.365 - 1^2 = 114.365 \tag{3.2.53}$$

By (3.2.51)-(3.2.53)

$$d_1^TQ_1^{-1}d_1 - e_1 = -9.2008$$

Thus

$$e_1 > d_1^TQ_1^{-1}d_1 \tag{3.2.54}$$

Also, n/m = 3/2 > N = 1. Therefore, we can use Theorem 3.2.3 which says that no solution exists for N=1. Setting N=2, we obtain

$$\psi_2 = (C^2)^T C^2 = \begin{bmatrix} 0.000335 & 0 & 0 \\ 0 & 0.018315 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3.2.55)$$

$$Q_2 = \overline{F}_2^T \psi_2 \overline{F}_2 = [F_0, F_1]^T \psi_2 [F_0, F_1] = \begin{bmatrix} 0.05750 & 0 & 0.17229 & 0 \\ 0 & 1 & 0 & 1 \\ 0.17229 & 0 & 0.58649 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

(3.2.56)

Since N=2 > n/m = 3/2, we know that the target can be reached in two sampling periods. The transformation given by (3.2.26) is then used where

$$P_2 = \begin{bmatrix} 0.28469 & 0 & -0.95862 & 0 \\ 0 & 0.70711 & 0 & -0.70711 \\ 0.95862 & 0 & 0.28469 & 0 \\ 0 & 0.70711 & 0 & 0.70711 \end{bmatrix} \quad (3.2.57)$$

The eigenvalues of Q_2 are

$$\gamma_1 = 0.63765 \quad \gamma_2 = 2.0000 \quad \gamma_3 = 0.006333 \quad \gamma_4 = 0.0000$$

(3.2.53)

Then by (3.2.27) we have that

$$0.63765y_0^2 + 2y_1^2 + 0.00633y_2^2 + 0y_3^2 = 1 \qquad (3.2.59)$$

One possible solution to (3.2.59) is

$$y_0 = 0 \quad y_1 = 0.7071 \quad y_2 = 0 \quad y_3 = 0 \qquad (3.2.60)$$

In this case the transformation given by (3.2.26) yields

$$U = P_2Y + \bar{F}_2^T(\bar{F}_2\bar{F}_2^T)^{-1}x_0 \tag{3.2.61}$$

$$= \begin{bmatrix} 0.28469 & 0 & -0.95862 & 0 \\ 0 & 0.70711 & 0 & -0.70711 \\ 0.95862 & 0 & 0.28469 & 0 \\ 0 & 0.70711 & 0 & 0.70711 \end{bmatrix} \begin{bmatrix} 0 \\ 0.7071 \\ 0 \\ 0 \end{bmatrix}$$

$$+ \begin{bmatrix} 0.18218 & -.92067 & 0 \\ 0 & 0 & -0.50000 \\ -0.06702 & 0.12460 & 0 \\ 0 & 0 & -0.50000 \end{bmatrix} \begin{bmatrix} 10 \\ -10 \\ 10 \end{bmatrix} = \begin{bmatrix} 11.029 \\ -4.500 \\ -1.916 \\ -4.500 \end{bmatrix} \tag{3.2.62}$$

In the above we have used the fact that $\bar{F}_N = [F_0, F_1]$.

From (3.2.62) we have

$$u(0) = \begin{bmatrix} 11.029 \\ -4.500 \end{bmatrix} \qquad u(1) = \begin{bmatrix} -1.916 \\ -4.500 \end{bmatrix} \tag{3.2.63}$$

Substituting (3.2.63) into the state equation (3.2.45) gives the following sequence of time-optimal states.

$$x(0) = (10 \quad -10 \quad 10)^T$$
$$x(1) = (6.1206 \quad 3.2926 \quad 5.500)^T$$
$$x(2) = (0 \quad 0 \quad 1)^T$$

and

$$x^T(2)x(2) = 0^2 + 0^2 + 1^2 = 1$$

Thus the target is reached in N=2 samples.

Another sequence of time-optimal controls can be found from (3.2.59) by choosing

$$y_0 = 1 \quad y_1 = 0 \quad y_2 = 7.5659 \quad y_3 = 10 \tag{3.2.64}$$

Substituting (3.2.64) into (3.2.61) gives

$$u(0) = \begin{bmatrix} 4.0605 \\ -12.0711 \end{bmatrix} \qquad u(1) = \begin{bmatrix} 1.1964 \\ 2.0711 \end{bmatrix}$$

This sequence of controls gives the following sequence of states

$$x(0) = (10 \quad -10 \quad 10)^T$$

$$x(1) = (3.1088 \quad -1.1121 \quad -2.0711)^T$$

$$x(2) = (0.93795 \quad 0.34712 \quad 0 \quad)^T$$

and $x^T(2)x(2) = 1.000$ as required.

3.3 Minimization of Energy When Time-Optimal Sequence is Not Unique

In the preceeding section it was shown that in general the sequence of time-optimal controls that drives the initial state to the target was not unique. This allows for the possibility of optimizing the system according to another criterion. The criterion chosen here is to minimize the energy required to drive the initial state to the target. For computational purposes we would like a method of determining this sequence of controls that is adaptable to computer solution. The problem is now stated more formally.

Problem Statement Given the discrete-time system described by equation (3.2.1). Let N be any integer greater than or equal to the minimum number of samples required to reach the target described by (3.1.3). Find a sequence of controls which minimize the total energy required to drive the system to the target set; that is, we wish to minimize J where

$$J = \sum_{i=0}^{N-1} u^T(iT)u(iT)$$

<u>Solution</u> Let $U = (u(0), u(T), \ldots, u[(N-1)T])^T$. Then J can be written as

$$J = U^T U \tag{3.3.1}$$

From equation (3.2.8) we have

$$x^T(NT)x(NT) = U^T Q_N U - 2d_N^T U + x_0^T \psi_N x_0 \tag{3.3.2}$$

Adjoining equation (3.3.2) to (3.3.1) by means of a Lagrange multiplier $(-1/\gamma)$, we obtain for the Lagrangian L

$$\begin{aligned} L &= U^T U - (1/\gamma)(x^T(NT)x(NT) - R^2) \\ &= U^T U - (1/\gamma)(U^T Q_N U - 2d_N^T U + x_0^T \psi_N x_0 - R^2) \end{aligned}$$

Setting $\frac{\partial L}{\partial U} = 0$, we arrive at the following equation.

$$U = -(\gamma I - Q_N)^{-1} d_N \tag{3.3.3}$$

Substituting equation (3.3.3) into (3.3.2) and setting $x^T(NT)x(NT) - R^2 = 0$, the following equation is obtained

$$d_N^T(\gamma I - Q_N)^{-1} Q_N (\gamma I - Q_N)^{-1} d_N + 2d_N^T(\gamma I - Q_N)^{-1} d_N + e_N = 0 \tag{3.3.4}$$

where e_N is given by equation (3.2.13). Equation (3.3.4) can be written as

$$d_N^T \mathrm{adj}(\gamma I - Q_N) Q_N \mathrm{adj}(\gamma I - Q_N) d_N + 2d_N^T \mathrm{adj}(\gamma I - Q_N) d_N c_N(\gamma) + e_N c_N^2(\gamma) = 0 \tag{3.3.5}$$

where

$$(\gamma I - Q_N)^{-1}c_N(\gamma) = \text{adj}(\gamma I - Q_N) = G_0\gamma^{Nm-1}$$

$$+ G_1\gamma^{Nm-2}+\ldots+G_{Nm-2}\gamma + G_{Nm-1} \quad (3.3.6)$$

and

$$c_N(\gamma) = |\gamma I - Q_N| = s_0\gamma^{Nm} + s_1\gamma^{Nm-1} +\ldots+s_{Nm-1}\gamma + s_{Nm} \quad (3.3.7)$$

The s_i and G_i can be found from Theorem 2.1.10. Equation (3.3.5) is a polynomial of order 2Nm in γ. The problem is to reduce (3.3.5) to the form

$$\rho_0\gamma^{2Nm} + \rho_1\gamma^{2Nm-1} +\ldots+\rho_{2Nm-1}\gamma + \rho_{2Nm} = 0 \quad (3.3.8)$$

where the ρ_i are constants to be determined. If $N > 2$, the reduction of (3.3.5) to (3.3.8) by hand computation becomes very tedious. The bulk of this section is devoted to finding the ρ_i in a fashion that is adaptable to computer solution and thus to cases where N is large. It will also be shown that the polynomial given in Equation (3.3.8) is actually of lower order than shown if $N > n/m$. We begin by proving the following theorem.

Theorem 3.3.1 Let $N \geq n/m$. Then the expansion of the conjoint matrix given in Theorem 2.1.10 for the matrix Q_N has the property that $G_{n+1} = G_{n+2} = \ldots = G_{Nm} = 0$.

Proof By equation (3.2.10), $Q_N = \bar{F}_N^T\psi_N\bar{F}_N$. The matrix $\bar{F}_N$ is of order nxNm and of rank n by Assumption 2 of Section 3.1. By Theorem 2.1.1 $\text{rank}(Q_N) \leq \min[\text{rank}(F_N), \text{rank}(\psi_N)]=$ n. Since ψ_N is nonsingular, $\psi_N\bar{F}_N$ is of rank n because multiplication of a matrix by a nonsingular matrix does not change the rank. By Theorem 2.1.2,

$$\begin{aligned} \text{rank}(Q_N) &\geq \text{rank}(\bar{F}_N^T) + \text{rank}({}_N\bar{F}_N) - n \\ &\geq n + n - n = n \end{aligned}$$

Therefore rank$(Q_N) = n$. Since Q_N is symmetric, Nm-n of the eigenvalues of Q_N are zero zy Theorem 2.1.7. The characteristic equation (3.3.7) becomes $c_N(\gamma) = \gamma^{Nm-n}(\gamma^n + s_1\gamma^{n-1} + \ldots + s_n)$. That is,

$$s_{n+1} = s_{n+2} = \ldots = s_{Nm} = 0 \qquad (3.3.9)$$

and the nonzero eigenvalues γ_i of Q_N satisfy the equation

$$\gamma_i^n + s_i\gamma_i^{n-1} + \ldots + s_n = 0 \quad i\text{-}1,2,\ldots,n \qquad (3.3.10)$$

By Theorem 2.1.10, the quantity Q_NG_n can be written as

$$\begin{aligned} Q_NG_n &= Q_N^2G_{n-1} + s_nQ_N \\ &= Q_N^2(Q_NG_{n-2} + s_{n-1}I) + s_nQ_N \\ &= Q_N^3G_{n-2} + s_{n-1}Q_N^2 + s_nQ_N \\ &\vdots \\ &= Q_N^{n+1} + s_1Q_N^n + s_2Q_N^{n-1} + \ldots + s_nQ_N \end{aligned}$$

From Theorem 2.1.12 and the fact that the sum of symmetric matrices is symmetric, we have that Q_NG_n is symmetric. By Theorem 2.1.11 the eigenvalues θ_i of Q_NG_n are

$$\theta_i = \gamma_i(\gamma_i^n + s_1\gamma_i^{n-1} + \ldots + s_n) \qquad i = 1,2,\ldots,Nm$$

where γ_i are the eigenvalues of Q_N. Thus for those eigenvalues of Q_N which are zero we have $\theta_i = 0$, and by equation (3.3.10) we have that $\theta_i = 0$ also for the nonzero eigenvalues of Q_N. Thus $Q_N G_n$ is a symmetric matrix with all zero eigenvalues. By Theorem 2.1.7 this implies that

$$Q_N G_n = 0 \tag{3.3.11}$$

From Theorem 2.1.10 we have that

$$G_i = Q_N G_{i-1} + s_i I \qquad i=n+1, n+2, \ldots, Nm$$

By equations (3.3.9) and (3.3.11) this implies that $G_i = 0$ for $i \geq n+1$.

QED

Theorem 3.3.2 a) The polynomial in equation (3.3.4) can be reduced to the form given by equation (3.3.8) where the ρ_i are as follows:

$$\rho_i = \begin{cases} e_N & \text{for } i=0 \\ 2(d_N^T d_N + e_N s_1) & \text{for } i=1 \\ 2d_N^T G_{i-1} d_N + \sum_{j=0}^{i-2} d_N^T (G_j Q_N G_{i-j-2} + 2G_j s_{i-j-1}) d_N \\ \quad + e_N \sum_{j=0}^{i} s_j s_{i-j} & \text{for } 2 \leq i \leq Nm \\ \sum_{j=i-Nm-1}^{Nm-1} [d_N^T (G_j Q_N G_{i-j-2} + 2G_j s_{i-j-1}) d_N + e_N s_{j+1} s_{i-j-1}] \\ & \text{for } Nm+1 \leq i \leq 2Nm \end{cases}$$

where the G_i and s_i are generated recursively by the equations

$$\begin{aligned}
G_0 &= I & s_0 &= 1 \\
G_1 &= Q_N G_0 + s_1 I & s_1 &= -\operatorname{tr}(Q_N) \\
G_2 &= Q_N G_1 + s_2 I & s_2 &= -\tfrac{1}{2}\operatorname{tr}(Q_N G_1) \\
&\vdots & &\vdots \\
G_{Nm-1} &= Q_N G_{Nm-2} + s_{Nm-1} I & s_{Nm} &= -(1/Nm)\operatorname{tr}(Q_N G_{Nm-1}) \\
G_{Nm} &= Q_N G_{Nm-1} + s_{Nm} I = 0
\end{aligned}$$

b) The quantities in the summations given above have the following property.

$$G_k Q_N G_j = G_j Q_N G_k \qquad j,k=0,1,\ldots,Nm-1$$

c) If $N \geqslant n/m$, the polynomial given by equation (3.3.8) reduces to a polynomial of order 2n, that is, equation (3.3.4) becomes

$$\bar{\rho}_0 \gamma^{2n} + \bar{\rho}_1 \gamma^{2n-1} + \ldots + \bar{\rho}_{2n-1}\,\gamma + \bar{\rho}_{2n} = 0$$

where the $\bar{\rho}_i$ are given by

$$\bar{\rho}_i = \begin{cases}
e_N & \text{for } i=0 \\
2(d_N^T d_N + e_N s_1) & \text{for } i=1 \\
2d_N^T G_{i-1} d_N + \sum_{j=0}^{i-2} d_N^T (G_j Q_N G_{i-j-2} + 2G_j s_{i-j-1}) d_N & \\
\qquad + e_N \sum_{j=0}^{i} s_j s_{i-j} & \text{for } 2 \leqslant i \leqslant n \\
\sum_{j=i-n-1}^{n-1} [d_N^T (G_j Q_N G_{i-j-2} + 2G_j s_{i-j-1}) d_N + e_N s_{j+1} s_{i-j-1}] & \\
& \text{for } n+1 \leqslant i \leqslant 2n
\end{cases}$$

<u>Proof</u> Let us expand the first term in equation (3.3.5) by using (3.3.6).

$$\mathrm{adj}(\gamma I - Q_N)Q_N\mathrm{adj}(\gamma I - Q_N)$$

$$= (I\gamma^{Nm-1}+G_1\gamma^{Nm-2}+G_2\gamma^{Nm-3}+\ldots+G_{Nm-2}\gamma+G_{Nm-1})Q_N(I\gamma^{Nm-1}$$

$$+G_1\gamma^{Nm-2}+G_2\gamma^{Nm-3}+\ldots+G_{Nm-2}\gamma+G_{Nm-1})$$

$$=Q_N\gamma^{2Nm-2}+(Q_NG_1 + G_1Q_N)\gamma^{2Nm-3}+ (Q_NG_2 + G_1Q_NG_1 + G_2Q_N)\gamma^{2Nm-4}$$

$$+ (Q_NG_3 + G_1Q_NG_2 + G_2Q_NG_1 + G_3Q_N)\gamma^{2Nm-5}$$

$$+ \ldots + (Q_NG_{Nm-2} + G_1Q_NG_{Nm-3} + \ldots + G_{Nm-3}Q_NG_1 + G_{Nm-2}Q_N)\gamma^{Nm}$$

$$+ (Q_NG_{Nm-1} + G_1Q_NG_{Nm-2} + G_2Q_NG_{Nm-3} + \ldots + G_{Nm-2}Q_NG_1$$

$$+ G_{Nm-1}Q_N)\gamma^{Nm-1}$$

$$+ (G_1Q_NG_{Nm-1} + G_2Q_NG_{Nm-2} + G_3Q_NG_{Nm-3} + \ldots + G_{Nm-2}Q_NG_2$$

$$+ G_{Nm-1}Q_NG_1)\gamma^{Nm-2} + \ldots$$

$$+ (G_{Nm-3}Q_NG_{Nm-1} + G_{Nm-2}Q_NG_{Nm-2} + G_{Nm-1}Q_NG_{Nm-3})\gamma^2$$

$$+ (G_{Nm-2}Q_NG_{Nm-1} + G_{Nm-1}Q_NG_{Nm-2})\gamma + G_{Nm-1}Q_NG_{Nm-1} \qquad (3.3.12)$$

Since we wish to find $d_N^T\mathrm{adj}(\gamma I - Q_N)Q_N\mathrm{adj}(\gamma I - Q_N)d_N$, we can use equation (3.3.12) to write

$$d_N^T\mathrm{adj}(\gamma I - Q_N)Q_N\mathrm{adj}(\gamma I - Q_N)d_N$$

$$= \theta_0\gamma^{2Nm} + \theta_1\gamma^{2Nm-1} + \theta_2\gamma^{2Nm-2} +$$

$$\ldots + \theta_{2Nm-1}\gamma + \theta_{2Nm} \qquad (3.3.13)$$

where

$$\theta_i = \begin{cases} 0 & \text{for } i=0,1 \\ \sum_{j=0}^{i-2} d_N^T G_j Q_N G_{i-2-j} d_N & \text{for } 2 \leq i \leq Nm \\ \sum_{j=i-Nm-1}^{Nm-1} d_N^T G_j Q_N G_{i-2-j} d_N & \text{for } Nm+1 \leq i \leq 2Nm \end{cases} \tag{3.3.14}$$

Let us turn to the second term in equation (3.3.5), that is, $\operatorname{adj}(\gamma I - Q_N)\ c_N(\gamma)$.

$$\begin{aligned} &\operatorname{adj}(\gamma I - Q_N)\ c_N(\gamma) \\ &= (G_0\gamma^{Nm-1}+G_1\gamma^{Nm-2}+G_2\gamma^{Nm-3}+\ldots+G_{Nm-2}\gamma+G_{Nm-1})(s_0\gamma^{Nm} \\ &\qquad +s_1\gamma^{Nm-1}+s_2\gamma^{Nm-2}+\ldots+s_{Nm-1}\gamma+s_{Nm}) \\ &= G_0\gamma^{2Nm-1} + (G_0s_1+G_1s_0)\gamma^{2Nm-2}+(G_0s_2+G_1s_1+G_2s_0)\gamma^{2Nm-3}+ \ldots \\ &\quad +(G_0s_{Nm-1}+G_1s_{Nm-2}+G_2s_{Nm-3}+ \ldots + G_{Nm-2}s_1 + G_{Nm-1}s_0)\gamma^{Nm} \\ &\quad +(G_0s_{Nm} + G_1s_{Nm-1} + G_2s_{Nm-2} + \ldots + G_{Nm-3}s_3 \\ &\qquad + G_{Nm-2}s_2 + G_{Nm-1}s_1)\gamma^{Nm-1} \\ &\quad +(G_1s_{Nm} + G_2s_{Nm-1} + \ldots + G_{Nm-3}s_4 + G_{Nm-2}s_3 + G_{Nm-1}s_2)\gamma^{Nm-2} \\ &\quad + (G_2s_{Nm} + G_3s_{Nm-1} + \ldots + G_{Nm-2}s_4 + G_{Nm-1}s_3)\gamma^{Nm-3}+ \ldots \\ &\quad + (G_{Nm-3}s_{Nm}+G_{Nm-2}s_{Nm-1}+G_{Nm-1}s_{Nm-2})\gamma^2 \\ &\quad + (G_{Nm-2}s_{Nm}+G_{Nm-1}s_{Nm-1})\gamma + G_{Nm-1}s_{Nm} \end{aligned} \tag{3.3.15}$$

The expression for $d_N^T\operatorname{adj}(\gamma I - Q_N)\ d_Nc_N(\gamma)$ can then be written as $2d_N^T\operatorname{adj}(\gamma I - Q_N)\ d_Nc_N(\gamma)$

$$= \zeta_0\gamma^{2Nm} + \zeta_1\gamma^{2Nm-1}=\zeta_2\gamma^{2Nm-2}+\ldots+\zeta_{2Nm-1}\gamma + \zeta_{2Nm} \tag{3.3.16}$$

where

$$\zeta_i = \begin{cases} 0 & \text{for } i=0 \\ 2\sum_{j=0}^{i-1} d_N^T G_j s_{i-j-1} d_N & \text{for } 1 \leq i \leq Nm \\ 2\sum_{j=i-1-Nm}^{Nm-1} d_N^T G_j s_{i-j-1} d_N & \text{for } Nm+1 \leq i \leq 2Nm \end{cases} \tag{3.3.17}$$

Consider now the last term in equation (3.3.5), that is, $c_N^2(\gamma)$.

$$\begin{aligned} c_N^2(\gamma) &= (s_0\gamma^{Nm} + s_1\gamma^{Nm-1} + s_2\gamma^{Nm-2} + \ldots + s_{Nm-1}\gamma + s_{Nm})(s_0\gamma^{Nm} \\ &\quad + s_1\gamma^{Nm-1} + s_2\gamma^{Nm-2} + \ldots + s_{Nm-1}\gamma + s_{Nm}) \\ &= s_0^2\gamma^{2Nm} + (s_0s_1 + s_1s_0)\gamma^{2Nm-1} + (s_0s_2 + s_1s_1 + s_2s_0)\gamma^{2Nm-2} \\ &\quad + (s_0s_3 + s_1s_2 + s_2s_1 + s_3s_0)\gamma^{2Nm-3} \\ &\quad + \ldots + (s_0s_{Nm} + s_1s_{Nm-1} + s_2s_{Nm-2} + \ldots + s_{Nm-2}s_2 + s_{Nm-1}s_1 \\ &\quad + s_{Nm}s_0)\gamma^{Nm} \\ &\quad + (s_1s_{Nm} + s_2s_{Nm-1} + \ldots + s_{Nm-1}s_2 + s_{Nm}s_1)\gamma^{Nm-1} \\ &\quad + (s_2s_{Nm} + s_3s_{Nm-1} + \ldots + s_{Nm-1}s_3 + s_{Nm}s_2)\gamma^{Nm-2} \\ &\quad + \ldots + (s_{Nm-1}s_{Nm} + s_{Nm}s_{Nm-1})\gamma + s_{Nm}^2 \end{aligned} \tag{3.3.18}$$

Therefore,

$$e_N c_N^2(\gamma) = \phi_0\gamma^{2Nm} + \phi_1\gamma^{2Nm-1} + \phi_2\gamma^{2Nm-2} + \ldots + \phi_{2Nm-1}\gamma + \phi_{2Nm} \tag{3.3.19}$$

where

$$\phi_i = \begin{cases} e_N & \text{for } i = 0 \\ e_N \sum_{j=0}^{i} s_j s_{i-j} & \text{for } 1 \leq i \leq Nm \\ e_N \sum_{j=i-Nm}^{Nm} s_j s_{i-j} & \text{for } Nm+1 \leq i \leq 2Nm \end{cases} \quad (3.3.20)$$

Let

$$\rho_i = \theta_i + \zeta_i + \phi_i \qquad i = 0,1,\ldots,2Nm \quad (3.3.21)$$

We then have by equations (3.3.5), (3.3.13), (3.3.16), (3.3.19) and (3.3.21)

$$\rho_0 \gamma^{2Nm} + \rho_1 \gamma^{2Nm-1} + \ldots + \rho_{2Nm-1}\gamma + \rho_{2Nm} = 0 \quad (3.3.22)$$

where by equations (3.3.14), (3.3.17), (3.3.20) and (3.3.21)

$$\rho_i = \begin{cases} e_N & \text{for } i = 0 \\ 2d_N^T d_N + e_N(s_0 s_1 + s_1 s_0) & \text{for } i=1 \\ \sum_{j=0}^{i-2} d_N^T G_j Q_N G_{i-2-j} d_N + 2\sum_{j=0}^{i-1} d_N^T G_j s_{i-j-1} d_N \\ \qquad + e_N \sum_{j=0}^{i} s_j s_{i-j} & \text{for } 2 \leq i \leq Nm \\ \sum_{j=i-Nm-1}^{Nm-1} d_N^T G_j Q_N G_{i-2-j} d_N + 2 \sum_{j=i-Nm-1}^{Nm-1} d_N^T G_j s_{i-j-1} d_N \\ \qquad + e_N \sum_{j=i-Nm}^{Nm} s_j s_{i-j} & \text{for } Nm+1 \leq i \leq 2Nm \end{cases}$$

By reordering the subscripts, the above equations can be written as

$$
\rho_i = \begin{cases}
e_N & \text{for } i=0 \\
2(d_N^T d_N + e_N s_1) & \text{for } i=1 \\
2d_N^T G_{i-1} d_N + \sum_{j=0}^{i-2} d_N^T (G_j Q_N G_{i-j-2} + 2G_j s_{i-j-1}) d_N & \\
\quad + e_N \sum_{j=0}^{i} s_j s_{i-j} & \text{for } 2 \leqslant i \leqslant Nm \\
\sum_{j=i-Nm-1}^{Nm-1} \{ d_N^T (G_j Q_N G_{i-j-2} + 2G_j s_{i-j-1}) d_N & \\
\quad + e_N s_{j+1} s_{i-j-1} \} & \text{for } Nm+1 \leqslant i \leqslant 2Nm
\end{cases} \tag{3.3.23}
$$

Thus the first part of the theorem has been proven.

b) From Theorem 2.1.10, G_k can be written as

$$
\begin{aligned}
G_k &= Q_N G_{k-1} + s_k I \\
&= Q_N (Q_N G_{k-2} + s_{k-1} I) + s_k I \\
&= Q_N^2 G_{k-2} + s_{k-1} Q_N + s_k I \\
&\vdots \\
&= Q_N^k + s_1 Q_N^{k-1} + s_2 Q_N^{k-2} + \dots + s_k I
\end{aligned}
$$

From this last equation we see that

$$Q_N G_k = G_k Q_N \tag{3.3.24}$$

$$G_j G_k = G_k G_j \tag{3.3.25}$$

By equation (3.3.24)

$$G_kQ_NG_j = G_kG_jQ_N \tag{3.3.26}$$

Similarly, $G_jQ_NG_k = G_jG_kQ_N$. However, by equation (3.3.25), this last equation becomes

$$G_jQ_NG_k = G_kG_jQ_N \tag{3.3.27}$$

By equating the right sides of equations (3.3.26) and (3.3.27), we have $G_kQ_NG_j = G_jQ_NG_k$, and the second part of the theorem is proven.

c) Let $N \geq n/m$. From Theorems 2.1.10 and 3.3.1 we have

$$\mathrm{adj}(\gamma I - Q_N) = \gamma^{Nm-n-1}(G_0\gamma^n+G_1\gamma^{n-1} +\ldots+G_{n-1}\gamma+G_n) \tag{3.3.28}$$

$$c_N(\gamma) = \gamma^{Nm-n}(s_0\gamma^n+s_1\gamma^{n-1} +\ldots+s_{n-1}\gamma+s_n) \tag{3.3.29}$$

Let us again expand the first term in equation (3.3.5) while using (3.3.28).

$$\mathrm{adj}(\gamma I - Q_N)Q_N\mathrm{adj}(\gamma I - Q_N)\gamma^{-2(Nm-n)}$$

$$= G_0Q_NG_0\gamma^{2n-2}+(G_0Q_NG_1+G_1Q_NG_0)\gamma^{2n-3}$$
$$+ (G_0Q_NG_2+G_1Q_NG_1+G_2Q_NG_0)\gamma^{2n-4}+\ldots+(G_0Q_NG_{n-3}+G_1Q_NG_{n-4}+\ldots$$
$$+G_{n-4}Q_NG_1+G_{n-3}Q_NG_0)\gamma^{n+1} +(G_0Q_NG_{n-2}+G_1Q_NG_{n-3}+\ldots$$
$$+G_{n-3}Q_NG_1+G_{n-2}Q_NG_0)\gamma^n + (G_0Q_NG_{n-1}+G_1Q_NG_{n-2}+\ldots$$
$$+G_{n-2}Q_NG_1+G_{n-1}Q_NG_0)\gamma^{n-1} +(G_0Q_NG_n+G_1Q_NG_{n-1}+G_2Q_NG_{n-2}+\ldots$$
$$+G_{n-2}Q_NG_2+G_{n-1}Q_NG_1+G_nQ_NG_0)\gamma^{n-2} +\ldots$$
$$+(G_{n-3}Q_NG_{n-1}+G_{n-2}Q_NG_{n-2}+G_{n-1}Q_NG_{n-3})\gamma^2$$

$$+(G_{n-2}Q_NG_{n-1}+G_{n-1}Q_NG_{n-2})\gamma+G_{n-1}Q_NG_{n-1} \qquad (3.3.30)$$

From equation (3.3.11) and the second part of this theorem, we have

$$Q_NG_n = G_nQ_N = 0 \qquad (3.3.31)$$

If we substitute equation (3.3.31) into (3.3.30), we have

$$d_N^T\text{adj}(\gamma I - Q_N)Q_N\text{adj}(\gamma I - Q_N)d_N\gamma^{-2(Nm-n)}$$

$$= \bar{\theta}_0\gamma^{2n}+\bar{\theta}_1\gamma^{2n-1}+\bar{\theta}_2\gamma^{2n-2} + \ldots + \bar{\theta}_{2n-1}\gamma+\bar{\theta}_{2n} \qquad (3.3.32)$$

where

$$\bar{\theta}_i = \begin{cases} 0 & \text{for } i=0,1 \\ \sum_{j=0}^{i-2} d_N^TG_jQ_NG_{i-2-j}d_N & \text{for } 2 \leq i \leq n \\ \sum_{j=i-n-1}^{n-1} d_N^TG_jQ_NG_{i-2-j}d_N & \text{for } n+1 \leq i \leq 2n \end{cases} \qquad (3.3.33)$$

Expanding the second term in equation (3.3.5) while using equations (3.3.28) and (3.3.29), we get

$$\text{adj}(\gamma I - Q_N)c_N(\gamma)\gamma^{-2(Nm-n)+1}$$

$$= (G_0\gamma^n+G_1\gamma^{n-1}+G_2\gamma^{n-2}+\ldots+G_{n-1}\gamma+G_n)(s_0\gamma^n+s_1\gamma^{n-1}+s_2\gamma^{n-2}+\ldots$$

$$+s_{n-1}\gamma+s_n) = G_0s_0\gamma^{2n}+(G_0s_1+G_1s_0)\gamma^{2n-1}+(G_0s_2+G_1s_1+G_2s_0)\gamma^{2n-2}$$

$$+\ldots+(G_0s_{n-2}+G_1s_{n-3}+\ldots+G_{n-2}s_0)\gamma^{n+2}+(G_0s_{n-1}+G_1s_{n-2}+\ldots$$

$$+G_{n-2}s_1+G_{n-1}s_0)\gamma^{n+1}+(G_0s_n+G_1s_{n-1}+\ldots+G_{n-1}s_1+G_ns_0)\gamma^n$$

$$+(G_1s_n+G_2s_{n-1}+\ldots+G_{n-1}s_2+G_ns_1)\gamma^{n-1} +\ldots+(G_{n-3}s_n+G_{n-2}s_{n-1}$$

$$+G_{n-1}s_{n-2}+G_ns_{n-3})\gamma^3+(G_{n-2}s_n+G_{n-1}s_{n-1}+G_ns_{n-2})\gamma^2$$

$$+(G_{n-1}s_n+G_ns_{n-1})\gamma+G_ns_n \qquad (3.3.34)$$

Since we are trying to determine $d_N^T\text{adj}(\gamma I - Q_N)d_Nc_N(\gamma)$, we show that this quantity can be simplified by showing that $d_N^TG_ns_id_N = 0$ for i=0,1,...,n. To prove this, we have by Theorem 2.1.10

$$G_i = Q_NG_{i-1} + s_iI \quad \text{for } i=0,1,2,\ldots,n \text{ with } G_{-1} \triangleq 0$$

Then by equation (3.3.31)

$$G_nG_i = G_nQ_NG_{i-1} + s_iG_n = s_iG_n$$

Therefore, by equation (3.2.12)

$$d_N^Ts_iG_nd_N = d_N^TG_nG_id_N = x_0^T\psi_N\bar{F}_NG_nG_i\bar{F}_N^T\psi_Nx_0$$

However, by equations (3.3.31) and (3.2.10)

$$0 = Q_NG_n = \bar{F}_N^T\psi_N\bar{F}_NG_N$$

Therefore,

$$y^T\bar{F}_N^T\psi_N\bar{F}_NG_nG_i\bar{F}_N^T\psi_N\bar{F}_N\, y = 0 \quad \text{for all } y$$

Let $x = \bar{F}_Ny$. Then $x^T\psi_N\bar{F}_NG_nG_i\bar{F}_N^T\psi_Nx = 0$. $\bar{F}_N$ is of order nxNm and be assumption N > n/m. This plus Assumption 2 of Section 3.1 imply that $\bar{F}_N$ is of full rank and $x^T\psi_N\bar{F}_NG_nG_i\bar{F}_N^T\psi_Nx = 0$ for all x. Thus

$$d_N^Ts_iG_nd_N = 0 \quad \text{for } i=0,1,\ldots,n \qquad (3.3.35)$$

Therefore, by equations (3.3.34) and (3.3.35)

$$d_N^T \mathrm{adj}(\gamma I - Q_N) d_N c_N(\gamma)\gamma^{-2(Nm-n)}$$

$$= d_N^T[G_0 s_0 \gamma^{2n-1} + (G_0 s_1 + G_1 s_0)\gamma^{2n-2} + (G_0 s_2 + G_1 s_1 + G_2 s_0)\gamma^{2n-3} + \ldots$$

$$+ (G_0 s_{n-2} + G_1 s_{n-3} + \ldots + G_{n-2} s_0)\gamma^{n+1} + (G_0 s_{n-1} + G_1 s_{n-2} + \ldots$$

$$+ G_{n-2} s_1 + G_{n-1} s_0)\gamma^n + (G_0 s_n + G_1 s_{n-1} + G_2 s_{n-2} + \ldots$$

$$+ G_{n-2} s_2 + G_{n-1} s_1)\gamma^{n-1} + (G_1 s_n + G_2 s_{n-1} + \ldots + G_{n-2} s_3 + G_{n-1} s_2)\gamma^{n-2} + \ldots$$

$$+ (G_{n-3} s_n + G_{n-2} s_{n-1} + G_{n-1} s_{n-2})\gamma^2 + (G_{n-2} s_n + G_{n-1} s_{n-1})\gamma$$

$$+ G_{n-1} s_n] d_N$$

or $2d_N^T \mathrm{adj}(\gamma I - Q_N) d_N c_N(\gamma)\gamma^{-2(Nm-n)}$

$$= \bar{\zeta}_0 \gamma^{2n} = \bar{\zeta}_1 \gamma^{2n-1} + \ldots + \bar{\zeta}_{2n-1}\gamma + \bar{\zeta}_{2n} \qquad (3.3.36)$$

where

$$\bar{\zeta}_i = \begin{cases} 0 & \text{for } i=0 \\ 2\sum_{j=0}^{i-1} d_N^T G_j s_{i-j-1} d_N & \text{for } 1 \leq i \leq n \\ 2\sum_{j=i-1-n}^{n-1} d_N^T G_j s_{i-j-1} d_N & \text{for } n+1 \leq i \leq 2n \end{cases} \qquad (3.3.37)$$

Let us now expand the third term in equation (3.3.5) using equation (3.3.29).

$$c_N^2(\gamma)\gamma^{-2(Nm-n)}$$

$$= (s_0\gamma^n + s_1\gamma^{n-1} + s_2\gamma^{n-2} + \ldots + s_{n-1}\gamma + s_n)(s_0\gamma^n + s_1\gamma^{n-1} + s_2\gamma^{n-2} + \ldots$$

$$+ s_{n-1}\gamma + s_n) = s_0^2\gamma^{2n} + (s_0 s_1 + s_1 s_0)\gamma^{2n-1} + (s_0 s_2 + s_1 s_1 + s_2 s_0)\gamma^{2n-2}$$

$$+\ldots+(s_0s_{n-1}+s_1s_{n-2}+\ldots s_{n-1}s_0)\gamma^{n+1}+(s_0s_n+s_1s_{n-1}+s_2s_{n-2}+\ldots$$

$$+s_{n-1}s_1+s_ns_0)\gamma^n+(s_1s_n+s_2s_{n-1}+\ldots+s_{n-1}s_2+s_ns_1)\gamma^{n-1}$$

$$+(s_{n-1}s_n+s_ns_{n-1})\gamma+s_n^2$$

Therefore,

$$e_Nc_N^2(\gamma)\gamma^{-2(Nm-n)} = \bar{\phi}_0\gamma^{2n}+\bar{\phi}_1\gamma^{2n-1}+\ldots+\bar{\phi}_{2n-1}\gamma+\bar{\phi}_{2n} \tag{3.3.38}$$

where

$$\bar{\phi}_i = \begin{cases} e_N & \text{for } i=0 \\ e_N \sum_{j=0}^{i} s_js_{i-j} & \text{for } 1 \leqslant i \leqslant n \\ e_N \sum_{j=i-n}^{n} s_js_{i-j} & \text{for } n+1 \leqslant i \leqslant 2n \end{cases} \tag{3.3.39}$$

Let

$$\bar{\rho}_i = \bar{\theta}_i + \bar{\zeta}_i + \bar{\phi}_i \qquad i=0,1,\ldots,2n$$

If we substitute equations (3.3.32), (3.3.36) and (3.3.38) into (3.3.5) and multiply both sides by the common factor $\gamma^{-2(Nm-n)}$, we obtain the equation

$$\bar{\rho}_0\gamma^{2n} + \bar{\rho}_1\gamma^{2n-1} + \ldots + \bar{\rho}_{2n-1}\gamma + \bar{\rho}_{2n} = 0$$

where the $\bar{\rho}_i$ are given above. By reordering the terms in the summation we obtain the result given in the third part of the theorem.

QED

Theorem 3.3.2 states that in order to find the sequence of time-energy optimal controls it is necessary to find the roots of a polynomial of order less than or equal to 2n where n is the order of the system. Once the roots of equation (3.3.8) have been found, the complex roots can be discarded and the real roots γ_i examined. The value of γ_i is substituted into equation (3.3.3) to obtain a candidate for an energy optimal control sequence. The quantity $U^T U$ corresponding to each value of γ_i is computed. The root that has the smallest value of $U^T U$ is the one requiring minimum energy to reach the boundary of the target. Since the expression for U was obtained by setting the first derivative of the Lagrangian equal to zero, this condition is also a necessary condition for determining the maximum energy required to reach the target in N samples. Thus one of the roots of the polynomial (3.3.8) corresponds to the sequence of controls requiring maximum energy to reach the target. For computational purposes the roots of the polynomial (up to 36 th order) can be computed by a pre-written FORTRAN program [IBM1].

It has been shown that the target can always be reached for some N such that $Nm \geq n$. If we choose N such that $Nm > n$ the minimum energy solution may require considerably less energy (all illustrated in Example 3.3.1). In this case the NmxNm matrices G_i appearing in Theorem 3.3.2 may be of large dimension requiring long computational time. An alternate formulation is now given which mitigates this problem. Some preliminary definitions are given first. Let

$$b_N = C^N x(0) \qquad \text{(nx1 vector)}$$

$$W_N = C^N \bar{F}_N \qquad \text{(nxNm matrix)}$$

$$K_N = W_N W_N^T \qquad \text{(nxn matrix)}$$

Then by equations (3.2.10), (3.2.12), and (3.2.13)

$$Q_N = W_N^T W_N$$

$$d_N = W_N^T b_N$$

$$e_N = b_N^T b_N - R^2$$

The expression for U given in equation (3.3.3) then becomes

$$U = -(\gamma I - Q_N)^{-1} d_N = -(\gamma I - W_N^T W_N)^{-1} W_N^T b_N$$

Applying Theorem 2.1.21 to this last term with A replaced by γI, B replaced by W_N^T and C replaced by I, we get an alternate expression for U.

$$U = - W_N^T (\gamma I - K_N)^{-1} b_N$$

From equations (3.2.5), (3.2.11) and the above definitions, we have that

$$x_N \triangleq x(NT) = b_N - W_N U$$

Therefore,

$$x_N = b_N + W_N (\gamma I - W_N^T W_N)^{-1} W_N^T b_N$$

$$= [I + W_N(\gamma I - W_N^T W_N)^{-1} W_N^T] b_N$$

We now apply Theorem 2.1.21 to the quantity in brackets with A replaced by I, B replaced by W_N and C replaced by I/γ. This gives $x_N = (I - K_N\gamma^{-1})^{-1} b_N$. Therefore, $x_N^T x_N$ = $b_N^T(I-K_N\gamma^{-1})^{-2} b_N$, and if the boundary of the target is to be reached in N samples, we require that

$$b_N^T(I-K_N\gamma^{-1})^{-2} b_N - R^2 = 0 \tag{3.3.40}$$

This can be rewritten as

$$\gamma^2 b_N^T(\gamma I-K_N)^{-2} b_N - R^2 = 0$$

or

$$\gamma^2 b_N^T \operatorname{adj}(\gamma I-K_N) \operatorname{adj}(\gamma I-K_N) b_N - c_N^2(\gamma) R^2 = 0$$

where

$$c_N(\gamma) = |\gamma I - K_N| = \bar{s}_0\gamma^n + \bar{s}_1\gamma^{n-1} + \ldots + \bar{s}_{n-1}\gamma + \bar{s}_n$$

$$\operatorname{adj}(\gamma I-K_N) = \bar{G}_0\gamma^{n-1} + \bar{G}_1\gamma^{n-2} + \ldots + \bar{G}_{n-2}\gamma + \bar{G}_{n-1}$$

Making these substitutions and collecting terms, the above polynomial reduces to the polynomial given in the following theorem.

<u>Theorem 3.3.3</u> a) If the sequence of time-optimal controls is not unique, the sequence of energy-optimal controls for this value of N is given by

$$U = -W_N^T(\gamma I-K_N)^{-1} b_N \tag{3.3.41}$$

where γ is a root of the polynomial

$$\alpha_0\gamma^{2n} + \alpha_1\gamma^{2n-1} + \ldots + {}_{2n-1}\gamma + \alpha_{2n} = 0$$

The α_i are given by

$$\alpha_i = \begin{cases} \sum_{j=0}^{i} b_N^T(\overline{G}_j\overline{G}_{i-j})b_N - R^2\overline{s}_j\overline{s}_{i-j} & \text{for } 0 \leq i \leq Nm \\ \sum_{j=i-Nm}^{Nm} b_N^T(\overline{G}_j\overline{G}_{i-j})b_N - R^2\overline{s}_j\overline{s}_{i-j} & \text{for } Nm+1 \leq i \leq 2Nm \end{cases}$$

where $\overline{s}_i$ and $\overline{G}_i$ are given by

$$\overline{G}_0 = I \qquad \overline{s}_0 = 1$$

$$\overline{G}_1 = K_N\overline{G}_0 + \overline{s}_1 I \qquad \overline{s}_1 = -\operatorname{tr}(K_N\overline{G}_0)$$

$$\vdots \qquad \vdots$$

$$\overline{G}_{n-1} = K_N\overline{G}_{n-2} + \overline{s}_{n-1}I \qquad \overline{s}_n = -(1/n)\operatorname{tr}(K_N\overline{G}_{n-1})$$

$$\overline{G}_n = K_N\overline{G}_{n-1} + \overline{s}_n I = 0$$

b) If $N \geq n/m$, the α_i become

$$\alpha_i = \begin{cases} \sum_{j=0}^{i} \left[b_N^T(\overline{G}_j\overline{G}_{i-j})b_N - R^2\overline{s}_j\overline{s}_{i-j}\right] & \text{for } 0 \leq i \leq n-1 \\ \sum_{j=i-n+1}^{n-1} \left[b_N^T(\overline{G}_j\overline{G}_{i-j})b_N - R^2\overline{s}_j\overline{s}_{i-j}\right] - 2R^2 s_n s_{i-n} & \text{for } n \leq i \leq 2n-2 \\ -2^2\overline{s}_{n-1}\overline{s}_n & \text{for } i = 2n-1 \\ -R^2\overline{s}_n^2 & \text{for } i = 2n \end{cases}$$

The proof of this theorem is similar to that for Theorem 3.3.2 so it is not repeated. The advantage of this expansion is that the $\overline{G}_i$ are of order nxn instead of NmxNm as in Theorem 3.3.2. Thus when N is large the matrix multiplication is less likely to become prohibitive.

The results presented thus far have been for determining time-energy optimal controls to the boundary of the target. It has already been shown that the boundary of the target can be reached in less than or an equal number of samples as to reach the interior of the target. Theorem 3.3.2 gives a procedure for determining a sequence of minimum energy controls to the boundary of the target. The question might be asked if the minimum energy solution to the interior of the target could be less than that to the boundary of the target. The following theorem shows that this can only happen for the trivial case when $u(0) = u(T) = \ldots = u[(N-1)T] = 0$. Heuristically, we see this case could arise if the system had negative eigenvalues. By letting $u(0)=u(T)=\ldots=u[(N-1)T] = 0$, the system "coasts" towards the origin and thus enters the target. If the sampling period is such that $x(NT)$ is contained in G, then this is the minimum energy sequence.

<u>Theorem 3.3.4</u> Let the initial state lie outside the target set, and let $N \geq N_m$ where N_m is the minimum number of samples required to reach the target. If $e_N \leq 0$, $U = 0$ is the sequence of minimum energy controls that drive the initial state to the target. If $e_N > 0$, the sequence of minimum energy controls drives the initial state to the boundary of the target.

<u>Proof</u> By equations (3.1.3), (3.2.8) and (3.2.13) we have that $\bar{x} \varepsilon G=\{x(NT):x^T(NT)x(NT) \leq R^2\}$ if and only if

$\overline{U}\epsilon H \equiv \{U: U^T Q_N U - 2d_N^T U + e_N \leq 0\}$ where $\overline{U}$ and $\overline{x}$ are related by $\overline{x}[(k+1)T] = C\overline{x}(kT) + Du(kT)$, $\overline{U} = [(u(0), u(T), \ldots, u[(N-1)T]]^T$.

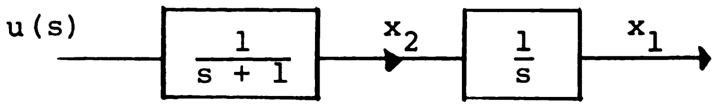
The sequence U=0 is contained in H if and only if $e_N \leq 0$. This sequence is, of course, the minimum energy solution. Suppose $e_N > 0$. Then the sequence U=0 does not belong to H. Furthermore, the sequence with minimum $\|U\|^2$ must be a boundary point of H. To show this, assume that an interior point p of H is such that $\|p\|^2$ = minimum of $\|U\|^2$. Since p is an interior point, there exists a neighborhood S about p that is contained in H. Consider the line segment from the origin (U=0) to p. Let q be the point of intersection of the boundary of S and the line segment. Then by construction q must lie between p and the origin. Therefore, $\|q\|^2 < \|p\|^2$ since q is closer to the origin. Hence by contradiction, the sequence of energy-optimal controls are boundary points of H. The boundary points of H are those sequences of controls that drive the initial state to the boundary of G, and thus the result follows.

QED

The following examples illustrate the theory presented above.

<u>Example 3.3.1</u> Consider the single input system described by the following transfer function and corresponding block diagram.

$$G(s) = \frac{1}{s(s+1)}$$



It is assumed that the control signal u(t) is the output of a zero-hold device. The sampling period is T = 1 second. The differential equations corresponding to the above system are:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) \qquad (3.3.42)$$

We want to find a sequence of controls which drive the system from the initial state $x(0) = (10,-12)^T$ to the target $x^T(\bar{N}T)x(\bar{N}T) < 1$ in the fewest number of samples $\bar{N}$. If the sequence is not unique, we want to determine a sequence that minimizes the total energy to reach the target. Compare the minimum energy for $\bar{N}$ samples with that for $\bar{N} + 1$ samples.

<u>Solution</u> The discrete-time system corresponding to equation (3.3.42) is:

$$x(k + 1) = Cx(k) + Du(k) \qquad (3.3.43)$$

where

$$C = \begin{bmatrix} 1 & 1-e^{-1} \\ 0 & e^{-1} \end{bmatrix} \qquad (3.3.44)$$

$$D = \begin{bmatrix} e^{-1} \\ 1-e^{-1} \end{bmatrix} \qquad (3.3.45)$$

Using the technique described in Section 3.2, we find that the minimum number of samples required to reach the target is $\bar{N} = 2$. Moreover, the solution is not unique. By routine calculations, we find that

$$Q_2 = \begin{bmatrix} 0.6431 & 0.4293 \\ 0.4293 & 0.5349 \end{bmatrix} \quad (3.3.46)$$

$$d_2 = (0.6662 \quad 1.1649)^T \quad (3.3.47)$$

$$e_2 = 1.7788 \quad (3.3.48)$$

$$P_2 = \begin{bmatrix} 0.7500 & -0.6615 \\ 0.6615 & 0.7500 \end{bmatrix} \quad (3.3.49)$$

$$\Lambda_2 \quad \begin{bmatrix} 1.0217 & 0 \\ 0 & 0.15627 \end{bmatrix} \quad (3.3.50)$$

From Theorem 3.2.2, we have

$$U = P_2 Y + Q_2^{-1} d_2 \quad (3.3.51)$$

where

$$Y^T \Lambda_2 Y = R^2 \quad (3.3.52)$$

If we substitute equation (3.3.50) into (3.3.52) and recall that R = 1, we have

$$\frac{Y_0^2}{0.9788} + \frac{Y_1^2}{6.3992} = 1 \quad (3.3.53)$$

Equation (3.3.53) is that of an ellipse; the square of the semiaxes are 0.9788 and 6.3992. The ellipse is shown in Figure 3.3.1a. The interpretation of this ellipse is that every point on the ellipse corresponds to a sequence of time-optimal controls. To determine the actual

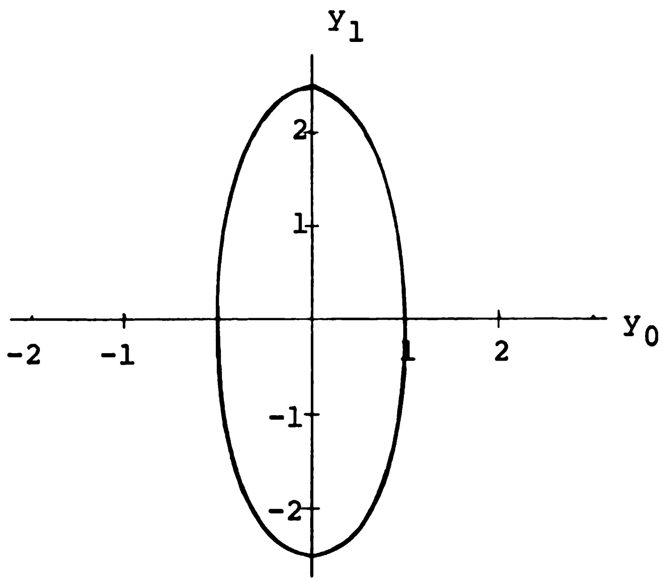


Fig. 3.3.1a. Locus of time-optimal controls in y_0-y_1 plane for Example 3.3.1.

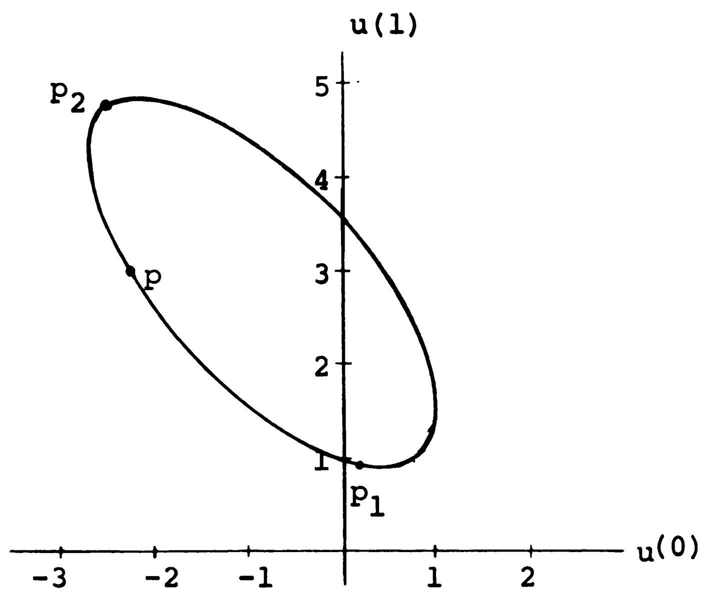


Fig. 3.3.1b. Locus of time-optimal controls in u(0)-u(1) plane for Example 3.3.1.

sequence of controls U, we use the transformation given by equation (3.3.51). This transformation rotates and shifts the coordinates, but distances and angles are preserved. Thus after the transformation, the locus of time-optimal controls is again an ellipse. This ellipse is shown in Figure 3.3.1b. Any point on this ellipse represents a sequence of time-optimal controls to the boundary of the target set. Points inside the ellipse map into points inside the target. For example, if we take the point "p" in Figure 3.3.1b, this gives the sequence of time-optimal controls

$$p = U = \begin{bmatrix} u(0) \\ u(1) \end{bmatrix} = \begin{bmatrix} -2.243 \\ 3.049 \end{bmatrix}$$

If we substitute this sequence of controls into the state equation (3.3.43), we find that

$x(0) = (10 \quad -12)^T \quad x(1) = (1.59 \quad -5.83)^T$

$x(2) = (-0.976 \quad -0.218)^T$ and $x^T(2)\, x(2) = 1$ as desired.

Returning to the problem of determining the minimum energy time-optimal controls, we see from Figure 3.3.1b, that this sequence is the one corresponding to the point on the ellipse that is closest to the origin. To determine this sequence we use the first part of Theorem 3.3.2.

$$G_0 = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \begin{aligned} S_0 &= 1 \\ S_1 &= -\mathrm{tr}(Q_2) = -1.1780 \end{aligned}$$

$$G_1 = Q_2 + s_1 I = \begin{bmatrix} -0.5349 & 0.4293 \\ 0.4293 & -0.6431 \end{bmatrix} \qquad (3.3.54)$$

$$s_2 = 0.15966$$

The polynomial discussed in Theorem 3.3.2 then becomes

$$\rho_0 \gamma^4 + \rho_1 \gamma^3 + \rho_2 \gamma^2 + \rho_3 \gamma^2 + \rho_4 = 0 \qquad (3.3.55)$$

where

$$\rho_0 = e_2 = 1.7788 \qquad (3.3.56)$$

$$\rho_1 = 2(d_2^T d_2 + e_2 s_1) = -0.58916 \qquad (3.3.57)$$

Similarly by Theorem 3.3.2

$$\rho_2 = -0.41598 \quad \rho_3 = 0.37616 \quad \rho_4 = -0.02549 \qquad (3.3.58)$$

Substituting equations (3.3.56)-(3.3.58) into (3.3.55) and solving for the roots of the latter, we obtain

$$\gamma_1 = -0.6300 \qquad (3.3.59)$$

$$\gamma_2 = 0.07439 \qquad (3.3.60)$$

These values of γ_i are substituted into equation (3.3.3) to obtain a sequence of controls.

For $\gamma = \gamma_1$, $u(0) = 0.2125$ $u(1) = 0.9216$ Energy=0.895 (3.3.61)

and for $\gamma = \gamma_2$, $u(0) = -2.492$ $u(1) = 4.853$ Energy=29.76 (3.3.62)

Thus the sequence corresponding to γ_1 is the minimum energy solution while the sequence corresponding to γ_2 is the maximum energy solution. After substituting the controls into the state equation (3.3.43), we obtain the following sequence of states:

$x(0) = (10 \quad -12)^T \quad x(1) = (2.493 \quad -4.280)^T$

$x(2) = (0.126 \quad -0.992)^T$ and $x^T(2)x(2) = 1$

$x(0) = (10 \quad -12)^T \quad x(1) = (1.498 \quad -5.990)^T$

$x(2) = (-0.503 \quad 0.864)^T$ and $x^T(2)x(2) = 1$

The sequence of controls given by equations (3.3.61) and (3.3.62) are shown as points P_1 and P_2 respectively in Figure 3.3.1b. As mentioned previously, the minimum energy solution (P_1) is closest to the origin while the maximum energy solution (P_2) lies on the ellipse at the farthest point from the origin.

We now turn to the second part of the problem, that is, determining the minimum energy solution for N = 3 samples. In this case we find

$$Q_3 = \begin{bmatrix} 0.84354 & 0.72170 & 0.39048 \\ 0.72170 & 0.64306 & 0.42933 \\ 0.39048 & 0.42933 & 0.53491 \end{bmatrix}$$

$$d_3 = (1.3337 \quad 1.2153 \quad 0.89361)^T$$

$$e_3 = 1.3241$$

Since N > n, we can use the third part of Theorem 3.3.2. We find

$$G_0 = I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{aligned} S_0 &= 1 \\ S_1 &= -\mathrm{tr}(Q_3) = -2.02152 \end{aligned}$$

$$G_1 = Q_3G_0+s_1I = \begin{bmatrix} -1.17798 & 0.72170 & 0.39048 \\ 0.72170 & -1.37845 & 0.42933 \\ 0.39048 & 0.42933 & -1.48661 \end{bmatrix}$$

$$S2 = -\frac{1}{2}tr(Q_3G_1) = 0.4800$$

$$G_2 = Q_3G_1+s_2I = \begin{bmatrix} 0.15966 & -0.21840 & 0.05874 \\ -0.21840 & 0.29874 & -0.08034 \\ 0.05874 & -0.08034 & 0.02161 \end{bmatrix}$$

$$S_32 = -\frac{1}{3}tr(Q_3G_2) = 0$$

From the third part of Theorem 3.3.2,

$$\bar{\rho}_0\gamma^4 + \bar{\rho}_1\gamma^3 + \bar{\rho}_2\gamma^2 + \bar{\rho}_3\gamma + \bar{\rho}_4 = 0 \qquad (3.3.63)$$

where

$\bar{\rho}_0 = 1.3241$, $\bar{\rho}_1 = 2.7551$, $\bar{\rho}_2 = -4.8603$,

$\bar{\rho}_3 = 1.9406$, $\bar{\rho}_4 = -0.23040$

The real roots of equation (3.3.63) are

$$\gamma_1 = 0.25884,\ \gamma_2 = 0.29313,\ \gamma_3 = 0.69018,\ \gamma_4 = -3.3229 \qquad (3.3.64)$$

If we substitute the γ_i into equation (3.3.3) to obtain the sequence of controls, we find

For $\gamma = \gamma_1$, u(0)=-0.008189 u(1)=0.61190 u(2)=2.29686
and Energy = 5.650 (3.3.65)

For $\gamma = \gamma_2$, u(0)=1.72465 u(1)=1.01916 u(2)=-0.89911
and Energy = 4.8215 (3.3.66)

For $\gamma = \gamma_3$, u(0)=1.3112 u(1)=1.1614 u(2)=-0.7539
and Energy = 3.6368 (3.3.67)

For $\gamma = \gamma_4$, $u(0) = 0.2619$ $u(1)=0.2395$ $u(2)=0.1785$
and Energy = 0.15778 (3.3.68)

By comparing (3.3.65) - (3.3.68), we see the sequence of controls given by (3.3.68) is the minimum energy solution. The corresponding sequence of states using the optimal control is

$$
\begin{aligned}
x(0) &= (10 \quad -12)^T \\
x(1) &= (2.5109 \quad -4.2490)^T \\
x(2) &= (-0.08689 \quad -1.41176)^T \\
x(3) &= (-0.91363 \quad -0.40654)^T \\
\text{and} \quad x^T(3)x(3) &= 1
\end{aligned}
$$

The minimum energy for N = 3 is 17.7% of that for N = 2.

<u>Example 3.3.2</u> In this example we wish to determine the sequence of minimum energy controls for N=2 in Example 3.3.1 by using Theorem 3.3.3 instead of Theorem 3.3.2.

<u>Solution</u> Using the numerical values for C^2, $x(0)$, and D in Example 3.3.1, we find that

$$b_2 = C^2 x(0) = (-0.37597 \quad -1.62403)^T$$

$$W_2 = C^2 \overline{F}_2 = \begin{bmatrix} -0.76746 & -0.36788 \\ -0.23254 & -0.63212 \end{bmatrix}$$

$$K_2 = W_2 W_2^T = \begin{bmatrix} 0.72432 & 0.41101 \\ 0.41101 & 0.45366 \end{bmatrix}$$

Then

$$\overline{G}_0 = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \begin{aligned} \overline{s}_0 &= 1 \\ \overline{s}_1 &= -\operatorname{tr}(K_2) = -1.17798 \end{aligned}$$

$$\overline{G}_1 = \overline{K}_2 \overline{G}_0 + \overline{s}_1 I = \begin{bmatrix} -0.45366 & 0.41101 \\ 0.41101 & -0.72432 \end{bmatrix} \qquad \overline{s}_2 = 0.15966$$

$$G_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

If we substitute these values of $\bar{s}_i$ and $\bar{G}_i$ into the second part of Theorem 3.3.3, we find that

$$\alpha_0\gamma^4 + \alpha_1\gamma^3 + \alpha_2\gamma^2 + \alpha_3\gamma + \alpha_4 = 0 \qquad (3.3.69)$$

where

$\alpha_0 = 1.788,\ \alpha_1 = -0.58922,\ \alpha_2 = -0.41596,\ \alpha_3 = 0.37615,$

$\alpha_4 = -0.02549$ (3.3.70)

Upon comparing the polynomial given by (3.3.69)-(3.3.70) with the corresponding polynomial (3.3.55)-(3.3.58) in Example 3.3.1, we see that they are the same except for roundoff error. Thus the roots are the same and in particular, the root γ corresponding to the minimum energy solution is the same. If we substitute this value of γ into the expression for U given in Theorem 3.3.3, we obtain the same sequence of controls as that given in Example 3.3.1.

In Example 3.3.1 it is noted that the sequence of energy optimal controls corresponds to the negative root of a polynomial. It will be shown that this is always the case. Furthermore, the negative root is unique implying that the sequence of energy optimal controls is unique. We begin by proving the following theorem.

Theorem 3.3.5 Assume there exists a sequence of controls that drives the initial state to the target set. Then a sequence of energy optimal controls is given by

$$U = -W_N^T(\gamma I - K_N)^{-1}b_N \qquad (3.3.71)$$

where $b_N = C^N x(0)$, $W_N = C^N \overline{F}_N$ and $K_N = W_N W_N^T$, and γ is a root of the equation

$$\sum \beta_i \left(\frac{\gamma}{\gamma - \lambda_j}\right)^2 = R^2. \qquad (3.3.72)$$

Here the summation is over all the distinct eigenvalues λ_j of K_N. Moreover,

$$\beta_j = \sum [(P^T b_N)_i]^2 \qquad (3.3.73)$$

where the summation is over all entries i such that $\lambda_i = \lambda_j$. (If all eigenvalues are distinct, β_j is the square of the j th element of $P^T b_N$.) P is any orthogonal matrix which diagonalizes the K_N matrix. That is, $\Lambda = P^T K_N P$, $P^T P = I$ and Λ is a diagonal matrix with diagonal elements equal to the eigenvalues of K_N.

The minimum energy is

$$J = \sum \beta_i \frac{\lambda_i}{(\gamma - \lambda_i)^2}$$

where the summation is over all distinct eigenvalues λ_i of K_N.

<u>Proof</u> From equations (3.3.40) and (3.3.41), we have that the sequence of energy optimal controls is given by

$$U = -W_N^T(\gamma I - K_N)^{-1}b_N \qquad (3.3.74)$$

where γ is a root of the equation

$$b_N^T(I - K_N\gamma^{-1})^{-2}b_N - R^2 = 0 \qquad (3.3.75)$$

Since K_N is a symmetric matrix, there exists a matrix P such that $K_N = P \Lambda P^T$ where $P^TP = I$ and Λ is a diagonal matrix with diagonal elements equal to the eigenvalues of K_N. The quantity $(I - K_N\gamma^{-1})^{-2}$ in equation (3.3.75) can then be written in a different way.

$$
\begin{aligned}
(I - \gamma^{-1}K_N)^{-2} &= [I - \gamma^{-1}P \Lambda P^T]^{-2} \\
&= [PP^T - \gamma^{-1}P \Lambda P^T]^{-2} \\
&= [P(I - \gamma^{-1} \Lambda)P^T]^{-1}[P(I - \gamma^{-1} \Lambda)P^T]^{-1} \\
&= (P^T)^{-1}(I - \gamma^{-1} \Lambda)^{-1}P^{-1}(P^T)^{-1}(I - \gamma^{-1} \Lambda)P^{-1} \\
&= P(I - \gamma^{-1}\Lambda)^{-2}P^T \qquad (3.3.76)
\end{aligned}
$$

From equations (3.3.75)-(3.3.76) we have

$$(P^Tb_N)^T(I - \gamma^{-1} \Lambda)^{-2}(P^Tb_N) = R^2 \qquad (3.3.77)$$

The quantity $(I - \gamma^{-1} \Lambda)^{-2}$ is a diagonal matrix. Let $\phi_i = (\frac{\gamma}{\gamma - \lambda_i})^2$, i=1,2,...,r where r is the number of distinct eigenvalues λ_i of K_N. Then if $(P^Tb_N)_i$, i=1,2,...,n are the elements of the vector P^Tb_N, we have

$$
\left[(P^Tb_N)_1, (P^Tb_N)_1, \ldots, (P^Tb_N)_n\right]
\begin{bmatrix}
\phi_1 & & & & & & 0 & 0 \\
 & \ddots & & & & & & \\
0 & & \phi_1 & & & & 0 & 0 \\
\vdots & & & \phi_2 & & & \vdots & \vdots \\
\vdots & & & & \ddots & & \vdots & \vdots \\
\vdots & & & & & \phi_2 & \vdots & \vdots \\
 & & & & & & \ddots & 0 \\
0 & & & & & 0 & \phi_r & 0 \\
 & & & & & & & \ddots \\
0 & \cdots & & & & 0 & & \phi_r
\end{bmatrix}
\begin{bmatrix}
(P^Tb_N)_1 \\
(P^Tb_N)_2 \\
\vdots \\
(P^Tb_N)_n
\end{bmatrix}
= R^2
$$

This can be written as

$$\left(\frac{\gamma}{\gamma - \lambda_1}\right)^2 \sum_{\substack{\text{all } i \ni \\ \lambda_i=\lambda_1}} (P^T b_N)_i + \left(\frac{\gamma}{\gamma - \lambda_2}\right)^2 \sum_{\substack{\text{all } i \ni \\ \lambda_i=\lambda_2}} (P^T b_N)_i + \ldots$$

$$+ \left(\frac{\gamma}{\gamma - \lambda_r}\right)^2 \sum_{\substack{\text{all } i \ni \\ \lambda_i=\lambda_r}} (P^T b_N)_i = R^2$$

By defining β_j as in equation (3.3.73), we have that

$$\sum_{\substack{\text{distinct} \\ \lambda_j}} \beta_j \left(\frac{\gamma}{\gamma - \lambda_j}\right)^2 = R^2$$

and the first part of the theorem is proven.

To determine the expression for the minimum energy, we use the expression for U given by equation (3.3.71).

$$\begin{aligned} U^T U &= b_N^T(\gamma I - K_N)^{-1} W_N W_N^T (\gamma I - K_N)^{-1} b_N \\ &= b_N^T(\gamma I - K_N)^{-1} K_N (\gamma I - K_N)^{-1} b_N \\ &= \gamma^{-2} b_N^T (I - \gamma^{-1} P \Lambda P^T)^{-1} P \Lambda P^T (I - \gamma^{-1} P \Lambda P^T)^{-1} b_N \\ &= \gamma^{-2} b_N^T [P(I - \gamma^{-1} \Lambda) P^T]^{-1} P \Lambda P^T [P(I - \gamma^{-1} \Lambda) P^T]^{-1} b_N \\ &= \gamma^{-2} b_N P (I - \gamma^{-1} \Lambda)^{-1} \Lambda (I - \gamma^{-1} \Lambda)^{-1} P^T b_N \\ &= \sum \frac{\beta_i \lambda_j}{(\gamma - \lambda_j)^2} \end{aligned}$$

where the summation is over all the distinct eigenvalues of K_N.

QED

In general, the P matrix is not unique. However, the results given above do not depend on this nonuniqueness. A column p_j of P corresponding to a distinct eigenvalue is normalized so that $\|p_j\|^2 = \|p_i\|^2 = 1$. It follows that p_j can assume only two directions; the one opposite the other. $(P^T b_N)_j^2$ is then the same for either choice of p_j. If an eigenvalue is repeated q times, then there are still q independent eigenvectors associated with this eigenvalue since the matrix to be diagonalized, K_N, is symmetric. It then follows that $\sum_{j=1}^{q} (P^T b_N)_j^2$ is the same regardless of the choice of P.

Theorem 3.3.5 can then be used to determine additional properties of the energy optimal sequence of controls.

<u>Theorem 3.3.6</u> Let $e_N > 0$. Then equation (3.3.72) has a unique negative root γ_1. If $\mu = -\gamma_1$, then

$$\frac{\lambda min}{\alpha} \leqslant \mu \leqslant \frac{\lambda max}{\alpha}$$

where $\alpha = [(e_N + R^2)/R^2]^{\frac{1}{2}} - 1$. An approximation to the negative root γ_1 is

$$\gamma_1 \doteq -\frac{tr(K_N)}{\alpha n}$$

<u>Proof</u> Let $\beta = \beta_1 + \beta_2 + \ldots + \beta_n = (P^T b_N)^T (P^T b_N)$

$$= b_N^T P \, P^T b_N = b_N^T b_N$$

Then by Theorem 3.3.5 and equation (3.2.13),

$$\sum \frac{\beta_i}{\beta}\left(\frac{\gamma}{\gamma - \lambda_i}\right)^2 = \frac{R^2}{b_N^T b_N} = \frac{R^2}{x_0^T (C^N)^T C^N x_0} = \frac{R^2}{e_N + R^2} = \frac{1}{(1+\alpha)^2}$$

where $\alpha = [(e_N + R^2)/R^2]^{\frac{1}{2}} - 1$. If $e_N > 0$, then $\frac{1}{(1+\alpha)^2} < 1$.

Let $\mu = -\gamma > 0$. Then

$$\sum \frac{\beta_i}{\beta}\left(\frac{\mu}{\mu + \lambda_i}\right)^2 = \frac{1}{(1+\alpha)^2} < 1 \qquad (3.3.78)$$

The quantity $\beta_i/\beta \leq 1$ and $\Sigma\beta_i/\beta = 1$. Therefore, since $\left(\frac{\mu}{\mu + \lambda_i}\right)^2 < 1$, it follows that there always exists a value of μ which satisfies (3.3.78). The value of μ is unique sincd $\left(\frac{\mu}{\mu + \lambda_i}\right)^2$ is monotone increasing for $\mu > 0$, and the sum of monotone increasing functions in monotone increasing. Thus for a given α there is only one value of μ such that (3.3.78) is satisfied. Let $\lambda_i = \lambda_{max}$. Then

$$\left(\frac{\mu}{\mu + \lambda_{min}}\right)^2 = \sum \frac{\beta_i}{\beta}\left(\frac{\mu}{\mu + \lambda_{min}}\right)^2 \leq \frac{1}{(1+\alpha)^2}$$

Let $\lambda_i = \lambda_{min}$. Then

$$\left(\frac{\mu}{\mu + \lambda_{min}}\right)^2 = \sum \frac{\beta_i}{\beta}\left(\frac{\mu}{\mu + \lambda_{min}}\right)^2 \geq \frac{1}{(1+\alpha)^2}$$

Therefore,

$$\left(\frac{\mu}{\mu + \lambda_{max}}\right)^2 \leq \frac{1}{(1 + \alpha)^2} \leq \left(\frac{\mu}{\mu + \lambda_{min}}\right)^2$$

or

$$\frac{\mu}{\mu + \lambda_{max}} \leq \frac{1}{1 + \alpha} \leq \frac{\mu}{\mu + \lambda_{min}}$$

which implies that

$$\frac{\mu + \lambda_{max}}{\mu} \geq 1 + \alpha \geq \frac{\mu + \lambda_{min}}{\mu}$$

or

$$\lambda_{min} \leq \mu\alpha \leq \lambda_{max} \tag{3.3.79}$$

and the second part of the theorem is proven. From (3.3.79) we see that $\mu\alpha$ must lie between the smallest and largest eigenvalue of K_N. As an approximation, we could choose the average value of the eigenvalues. Thus

$$\mu\alpha \doteq \sum_{i=1}^{n} \frac{\lambda_i}{n}$$

By Theorem 2.1.19, the sum of the eigenvalues of K_N is just the trace of K_N. Thus

$$\gamma \doteq - \frac{tr(K_N)}{\alpha_n}$$

QED

<u>Theorem 3.3.7</u> Let $\varepsilon_N > 0$. The unique sequence of minimum energy controls is given by equation (3.3.71) where γ is the negative root of (3.3.72).

<u>Proof</u> Let γ_a be a root of (3.3.72). It is first shown that $\gamma_a < \lambda_{max}$. Suppose the opposite is true, that is, $\gamma_a \geq \lambda_{max} > 0$. Then $0 < \frac{\lambda_{max}}{\gamma_a} \leq 1$ and $1 - \frac{\lambda_{max}}{\gamma_a} \leq 1$.

Therefore, $\left(1 - \frac{\lambda_{max}}{\gamma_a}\right)^{-2} \geq 1$. This implies that $\sum \frac{\beta_i}{\beta} \left(\frac{\gamma_a}{\gamma_a - \lambda_{max}}\right)^2 \geq 1$ since $\sum \frac{\beta_i}{\beta} = 1$. This is a contradiction since by assumption γ_a is such that

$$\sum \frac{\beta_i}{\beta}\left(\frac{\gamma_a}{\gamma_a - \lambda_i}\right)^2 = \frac{R^2}{e_N + R^2} < 1$$

Therefore, $\gamma_a < \lambda_{max}$.

Let us return to the expression for the Lagragian L given in the second equation after equation (3.3.2). Taking the second derivative, we find that

$$\frac{d^2L}{dU^2} = 2(I - \frac{Q_N}{\gamma})$$

A necessary condition for a minimum energy solution is that $I - Q_N/\gamma$ be positive semidefinite. The sufficient condition is that $I - Q_N/\gamma$ be positive definite. If $\gamma < 0$, then $I - Q_N/\gamma$ is positive definite since Q_N is positive semidefinite. Thus the negative root of (3.3.72) corresponds to a minimum energy solution. Suppose that $\gamma > 0$. By the above result we have that $\lambda_{max} > \gamma > 0$. Then $-Q_N/\gamma$ has a negative eigenvalue greater than 1 and $I - Q_N/\gamma$ has a negative eigenvalue. Thus $I - Q_N/\gamma$ is not positive semidefinite and therefore no value of $\gamma > 0$ can correspond to a minimum energy solution. Hence, the minimum energy solution corresponds to the unique negative root of (3.3.72)

QED

We can now compare the results given in Theorem 3.3.7 with those of Theorems 3.3.2 and 3.3.3. In the latter, the sequence of optimum controls was found by solving for all the roots of a polynomial. The coefficients ρ_i of this polynomial are obtained from a sequence of matrix multiplications and additions. Due to roundoff, the values of the ρ_i may be slightly in error. Sometimes this can result in a large change in the value of the roots. The formulation given in Theorem 3.3.7 does not require the ρ_i to be found. However, it is necessary to determine the diagonalizing matrix P and the eigenvalues of K_N. If we define

$$h(\gamma) = \sum \beta_j \left(\frac{\gamma}{\gamma - \lambda_j} \right)^2 - R^2 \qquad (3.3.80)$$

in equation (3.3.72) then by the above results we know that $h(\gamma)$ is monotone on $(-\infty, 0)$. Furthermore, $\gamma = \gamma_1 = - \operatorname{tr}(K_N)/\alpha n$ is an approximate root of $h(\gamma)$. This allows us to find the negative root of $h(\gamma)$ by a numerical technique such as Newton's Method [JAM1] using γ_1 as a starting value.

<u>Example 3.3.3</u> We wish to determine the sequence of minimum energy controls for Example 3.3.1 (N=2) by using Theorems 3.3.5-3.3.7.

<u>Solution</u> A matrix that diagonalizes the K_2 matrix in Example 3.3.2 is

$$P = \begin{bmatrix} 0.81017 & -0.58620 \\ 0.58620 & 0.81017 \end{bmatrix}$$

The eigenvalues of K_2 are

$$\lambda_1 = 1.02171 \text{ and } \lambda_2 = 0.15627 \qquad (3.3.81)$$

From Example 3.3.2, $b_2 = (-0.37597 \ \ -1.62403)^T$. Thus

$$P^T b_2 = \begin{bmatrix} 1.25661 \\ -1.09535 \end{bmatrix}$$

$$\beta_1 = (P^T b_2)_1 = (1.25661)^2 = 1.5791 \qquad (3.3.82)$$

and

$$\beta_2 = (P^T b_2)_2 = 1.1998 \qquad (3.3.83)$$

Substituting (3.3.81)-(3.3.83) into (3.3.80) gives

$$h(\gamma) = 1.5791\left(\frac{\gamma}{\gamma-1.02171}\right)^2 + 1.1998\left(\frac{\gamma}{\gamma-0.15627}\right)^2 - 1 \qquad (3.3.84)$$

From Theorem 3.3.6, an approximate root of (3.3.84) is

$$\gamma \doteq \gamma_1 = -\frac{tr(K_2)}{2\alpha} \qquad (3.3.84)$$

where $\alpha = [(e_2 + R^2)/R^2]^{\frac{1}{2}} - 1$.

From Example 3.3.1, $e_2 = 1.7788$ and $R = 1$. From Example 3.3.2, $tr(K_2) = 1.1780$. Thus $\gamma_i \doteq -0.88$. Using γ_1 as an initial guess of the root of $h(\gamma)$, we find by Newton's Method that the negative root of $h(\gamma)$ is

$$\gamma = -0.62999 \qquad (3.3.85)$$

By Theorem 3.3.7 this value of γ corresponds to the minimum energy sequence of controls. If we compare (3.3.85) with the negative root given by (3.3.59) which was obtained by a different method, we see that they are the same except for roundoff error. Thus the sequence of energy optimal controls are the same as those obtained in Examples 3.3.1 and 3.3.2.

3.4 Time-Optimal Control Using Feedback Control

The previous sections of this chapter were concerned with determining the minimum number of samples and a corresponding sequence of controls which drive the system from a given initial state to a hyperspherical target. For each different initial state it is necessary to perform a new set of calculations to determine a sequence of time-optimal controls. To avoid these calculations, it would be desirable to synthesize the controller as a sample function of the state of the system. It was shown in Section 3.1 that under the assumption that $N = n/m =$ integer and $|D, CD, C^2D, \ldots, C^{N-1}D| \neq 0$ that the target could always be reached in N samples. On the other hand, if $N < n/m$ the target can not always be reached for an arbitrary initial state. Since we wish to synthesize the controller as a function of an arbitrary state, the discussion will be limited to the case where $N = n/m$.

The problem of time-optimal control using a feedback controller has been studied previously [TOU3], [KUO2] for the case when the target is the origin. It is shown here that for the case of a hyperspherical target the optimal control law can be put in a form that differs from that of the case when the target is the origin only by a constant. The problem is now stated more formally.

Problem Statement Given the linear discrete-time system described by

$$x\,[(k+1)T] = Cx(kT) + Du(kT) \qquad (3.4.1)$$
$$x(0) = x_0$$

We wish to determine a sequence of controls $u(kT)$, $k = 0,1,\ldots,N-1$ which drives the system from an arbitrary initial state x_0 to the target described by

$$\{x(NT) : x^T(NT)\ x(NT) \leq R^2\}$$

Where R is a real number. We require that the control be a function of the state of the system. It is assumed that $N = n/m$ is an integer.

Solution The problem is solved under the same assumptions as those given in the open-loop case. That is, it is assumed that the system is controllable and that $|D, CD, \ldots, C^{N-1}D| \neq 0$. As discussed previously, the second assumption is no assumption at all if the system has a single input.

By using a technique similar to that used to derive equation (3.2.3), we can write the solution to (3.4.1) as

$$x[(k + N)T] = C^N x(kT) + C^N \sum_{i=0}^{N-1} C^{-(i + 1)} D\, u[(k + i)T]$$

$$= C^N x(kT) - C^N \overline{F}_N U_k$$

where $\overline{F}_N = [F_0, F_1, \ldots, F_{N-1}]$ and the F_i are given by equation (3.2.4).

$$U_k = [u(kT), u[(k + 1)T], \ldots, u[(k + N-1)T]]^T$$

If we set $x[(k + N)T] = 0$, then

$$\begin{bmatrix} u(kT) \\ u[(k + 1)T] \\ \vdots \\ u[(k + N-1)T] \end{bmatrix} = \overline{F}_N^{-1} \; x(kT) \qquad (3.4.2)$$

$\overline{F}_N^{-1}$ exists by the assumption that $|D, CD, \ldots, C^N D^{-1}| \neq 0$. Premultiplying both sides by the matrix $[I_m 0]$ where I_m is an mxm identity matrix gives

$$u(kT) = [I_m 0]\overline{F}_N^{-1} x(kT) \qquad k = 0, 1, \ldots, N-1 \qquad (3.4.3)$$

This expression for u(kT) represents a sequence of time-optimal controls that drive the system to the origin. For the case of a hyperspherical target we assume that u(kT) has the following form

$$u(kT) = [I_m 0]\overline{F}_N^{-1} x(kT) + vR$$

where v is to be determined. The following theorem provides a solution to the time-optimal control problem.

<u>Theorem 3.4.1</u> Given the system described by (3.4.1) and the assumptions that $|D, CD, \ldots, C^{N-1}D| \neq 0$ and $N = n/m =$

integer. Then a sequence of feedback time-optimal controls to the target is given by $u(kT) = [I_m 0]\overline{F}_N^{-1} x(kT) + vR$

$k = 0,1,\ldots,N-1$ (3.4.4)

where v is any vector that satisfies the inequality

$$v^T G^T G v \leq 1 \qquad (3.4.5)$$

where

$$G = \sum_{i=0}^{N-1} (C + D[I_m 0]\overline{F}_N^{-1})^i D \qquad (3.4.6)$$

If we choose v such that

$$v^T G^T G v = 1$$

the sequence of controls given by (3.4.4) drives the system to the boundary of the hypersphere. If we choose $v = 0$, the controller drives the system to the origin.

<u>Proof</u> From (3.4.1) and (3.4.4)

$$x(T) = [C + D[I_m 0]\overline{F}_N^{-1}]x_0 + DvR$$

$$x(2T) = Cx(T) + Du(T)$$

$$= [C + D[I_m 0]\overline{F}_N^{-1}]\, x(T) + DvR$$

$$= [C + D[I_m 0]\overline{F}_N^{-1}]^2 x_0 + [C + D[I_m 0]\overline{F}_N^{-1}]\ DvR + DvR$$

or in general

$$x(NT) = [C + D[I_m 0]\overline{F}_N^{-1}]^N x_0 + \sum_{i=0}^{N-1} [C + D[I_m 0]\overline{F}_N^{-1}]^i\ DvR \qquad (3.4.7)$$

If we set $R = 0$

$$x(NT) = [C + D[I_m 0]\overline{F}_N^{-1}]^N x_0$$

It has been shown that with R = 0, $u(kT) = \overline{F}_N^{-1}x(kT)$, k = 0,1,...,N-1 drives the system to the origin in N samples. Thus

$$[C + D[I_m 0]\overline{F}_N^{-1}]^N x_0 = 0$$

Since x_0 is arbitrary, it follows that

$$\phi_C = [C + D[I_m 0]\overline{F}_N^{-1}]^N = 0 \qquad (3.4.8)$$

Thus ϕ_C is a nilpotent matrix of index N. The properties of ϕ_C have been studied by several authors [CAD3], [FAR1]. (For a correction to the latter reference, see [TUE1].)

Substituting (3.4.8) into (3.4.7), we get

$$x(NT) = \sum_{i=0}^{N-1} [C + D[I_m 0]\overline{F}_N^{-1}]^i DvR$$

Since we require that the target be reached in N samples, we have

$$R^2 \geq x^T(NT)x(NT) = R^2 \; v^T G^T Gv \qquad (3.4.9)$$

where G is given by (3.4.6). Thus we require that

$$v^T G^T Gv \leq 1$$

From (3.4.9) we see that if we choose $v^T G^T Gv = 1$, then x(NT) will lie at the boundary of the target. If v = 0, it has already been shown that x(NT) = 0.

It is worth noting that neither $\overline{F}_N$ nor v depend on x(kT) or R and that an expression for v can always be found to satisfy (3.4.5). The technique is illustrated by the following example.

<u>Example 3.4.1</u> Given the system described in Example 3.3.1 with $X(0) = (10 \quad -12)^T$, and $R = 1$. We wish to determine a sequence of time-optimal control which drive the initial state to the target using feedback control.

<u>Solution</u> From Example 3.3.1, we have for $N = 2$

$$\bar{F}_2 = [F_0, F_1] = \begin{bmatrix} 0.71828 & 3.67077 \\ -1.71828 & -4.67077 \end{bmatrix}$$

$$D = \begin{bmatrix} 0.36788 \\ 0.63212 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0.63212 \\ 0 & 0.36788 \end{bmatrix}$$

Thus

$$v^T G^T G v = v^T D^T [I_2 + C + D[I_2 0]\bar{F}_2^{-1}]^T [I_2 + C + D[I_2 0]F_2^{-1}]Dv$$

Making the appropriate substitutions while noting that v is a scalar in this case we have

$$v^T G^T G v = 0.39955\ v^2$$

By Theorem 3.4.1 we require that

$$0.39955 v^2 \leqslant 1$$

or

$$-1.582 \leqslant v \leqslant 1.582 \qquad (3.4.10)$$

If we choose $v = 1$, the feedback controller is given by

$$u(kT) = [1 \ 0]\bar{F}_2^{-1} x(k) + 1$$

$$= [-1.582 \ -1.243] x(k) + 1 \qquad (3.4.11)$$

Thus with $k = 0$

$$u(0) = [-1.582 \ -1.243] \begin{bmatrix} 10 \\ -12 \end{bmatrix} + 1 = 0.0998$$

From the state equation, we have

$$x(1) = \begin{bmatrix} 1 & .63212 \\ 0 & .36788 \end{bmatrix} \begin{bmatrix} 10 \\ -12 \end{bmatrix} + \begin{bmatrix} 0.36788 \\ 0.63212 \end{bmatrix} (0.0998) = \begin{bmatrix} 2.4513 \\ -4.3515 \end{bmatrix}$$

From (3.4.11) the control at the next step is

$$u(1) = [-1.582 \quad -1.243] \begin{bmatrix} 2.4513 \\ -4.3515 \end{bmatrix} + 1 = 2.5323$$

and

$$x(2) = \begin{bmatrix} 1 & .63212 \\ 0 & .36788 \end{bmatrix} \begin{bmatrix} 2.4513 \\ -4.3515 \end{bmatrix} + \begin{bmatrix} 0.36788 \\ 0.63212 \end{bmatrix} (2.5323) = \begin{bmatrix} 0.6322 \\ 0 \end{bmatrix}$$

Therefore, $x^T(2)x(2) = 0.3997 < 1$ as desired. The energy required is $u^2(0) + u^2(1) = 6.423$.

From (3.4.10) and Theorem 3.4.1 a sequence of control to the boundary of the target can be found by choosing $v = \pm 1.5820$. If we repeat the above steps, we get the following

For $v = +1.5820$	For $v = -1.5820$
$u(0) = 0.6818$	$u(0) = -2.4822$
$u(1) = 2.3182$	$u(1) = 3.4823$
$x(0) = (10 \; -12)^T$	$x(0) = (10 \; -12)^T$
$x(1) = (2.665, -3.984)^T$	$x(1) = (1.5014, -5.9836)^T$
$x(2) = (1.000, 0.000)^T$	$x(2) = (-1.000, 0)^T$
$x^T(2)x(2) = 1.000$	$x^T(2)x(2) = 1.000$
Energy = 5.839	Energy = 18.288

Thus the minimum energy feedback controller requires an energy of 5.839. From Example 3.3.1, the minimum energy open-loop controller required an energy of 0.895. Thus the optimum closed-loop controller requires 6.52 times as much energy as the open-loop controller. As an additional comparison, if we set R = 0 to find the sequence of controls which drive the system to the origin, we get

$$u(0) = -0.9002 \qquad x(0) = (10\ -12)^T$$
$$u(1) = 2.9002 \qquad x(1) = (2.083,\ -4.984)^T$$
$$x(2) = (0.000,\ 0.000)^T$$
$$\text{Energy} = 9.2215$$

Using feedback control, the energy required to drive the system to the origin is 1.579 times as much as that required to drive the system to the boundary using v = 1.5820.

The feedback controller has the advantage of a simple form independent of the initial state. However, there are two disadvantages. First, a certain period of time is required to perform the matrix multiplication and addition required to find u(kT). If the sampling period is too short, this cannot be done. Second, it may turn out in practice that not all the components of x(kT) are available for a measurement. In this case an estimate can be made of the unmeasureable components of x(kT). If the system is <u>observable</u> and all past inputs and outputs are available, then x(kT) can be determined exactly. For example, let the state equation and output equation be given by

$$x[(k+1)T] = Cx(kT) + Du(kT)$$
$$y(kT) = Bx(kT) \qquad k = 0,1,\ldots,N-1$$

Then under the assumption that $[BC^{-1}, BC^{-2}, \ldots, BC^{-n}]^{-1}$ exists, it can be shown [KUO2] that

$$x(kT) = \begin{bmatrix} BC^{-1} \\ BC^{-2} \\ \vdots \\ BC^{-n} \end{bmatrix}^{-1} \begin{bmatrix} y[(k-1)T] \\ y[(k-2)T] \\ \vdots \\ y[(k-n)T] \end{bmatrix}$$

$$+ \begin{bmatrix} BC^{-1}D & 0 & \cdots & 0 \\ BC^{-2}D & BC^{-1}D & & \vdots \\ \vdots & \vdots & & \vdots \\ BC^{-n}D & BC^{-n+1}D & \cdots & BC^{-1}D \end{bmatrix} \begin{bmatrix} u[(k-1)T \\ \vdots \\ u[(k-n)T] \end{bmatrix}$$

Thus if n past measurements of the output $y(kT)$ and input $u(kT)$ are available, we can determine $x(kT)$ exactly. For example, if $k = 0$, we can determine $x(0)$ if we have the outputs $y(-T), \ldots, y(-nT)$ and inputs $u(-T), \ldots, u(-nT)$. Once $x(0)$ is determined, we can find $u(0)$ by using Theorem 3.4.1.

3.5 Statement and Solution of Time-Optimal Control Problem with Single Input Delay

This section is devoted to the time-optimal control problem when there is a time delay in the control signal. Such a delay could occur when the control signal is sent through long transmission lines. To simplify the results, it is assumed that the delay time is an integral multiple of the sampling period. Such an approximation would be applicable when the sampling period is small compared with the delay time. For the case when the target set is the origin, the problem has been studied by several authors [KUR1], [KOP1], [KUO2]. The problem is now stated.

Problem Statement Given a linear time-invariant system described by

$$\dot{x} = Ax + Bu(t - pT) \tag{3.5.1}$$
$$x(0) = x_0$$

where

$x(t)$ is a $n \times 1$ vector
$u(t)$ is a $m \times 1$ vector
A is a $n \times n$ matrix
B is a $n \times m$ matrix
p is a positive integer described below

It is assumed that $u(t)$ is of the sampled-data type so that

$$u(t - pT) = u(kt - pT) \text{ for } kT \leq t < (k + 1)T. \tag{3.5.2}$$

where T is the sampling period. Since it is assumed that the delay time T_d is an integral multiple of the sampling period, we have

$$T_d = pT$$

where p is a positive integer. We wish to find the minimum number of samples $\bar{N}$ and a corresponding sequence of controls $u(t)$, $u(2t)$,..., $u[(\bar{N} - p - 1)T]$ such that $x(\bar{N}T) \varepsilon G$ where G is given by $\{x(\bar{N}T) : x^T(\bar{N}T)x(\bar{N}T) \leq R^2\}$. It is assumed that $x(0)$ (assumed to lie outside the target) and the past controls $u(-T)$, $u(-2T)$,..., $u(-pT)$ are known.

<u>Solution</u> We make the same two assumptions as for the undelayed system. That is, it is assumed that (1) the system is completely controllable and (2)

$$\text{rank}[D, CD, \ldots, C^{N-1}D] = \text{maximum} \qquad (3.5.3)$$

for $N > 0$.

Proceeding in a fashion similar to that used in deriving the solution to the undelayed system, we write the solution to (3.5.1) as

$$x(t) = \phi(t - t_0)x(t_0) = \int_{t_0}^{t} \phi(t - \tau)Bu(\tau - pT)d\tau$$

Using equation (3.5.2), the above equation can be written as the following difference equation.

$$x[(k + 1)T] = Cx(kT) + Du(kT - pT) \qquad (3.5.4)$$

where

$$D = \int_0^T \phi(T - \tau)Bd\tau \qquad (3.5.5)$$

$$C = \phi(T)$$

The solution of (3.5.4) is

$$x(kT) = C^k x(0) = C^k \sum_{i=0}^{k-1} C^{-1-i}D\, u[(i - p)T]$$

As in equation (3.2.4), we define

$$F_j = -C^{-(j + i)}D \qquad j = 0,1,\ldots,N - 1 \qquad (3.5.7)$$

Then

$$x(kT) = C^k x_0 - C^k \sum_{i=0}^{k-1} F_i u[(i - p)T \qquad (3.5.8)$$

Assuming that the past initial sequence u(-T), u(-2T),..., u(-pT) is known, then by equation (3.5.8),

$$k - 1 - p \geq 0 \tag{3.5.9}$$

We can then write equation (3.5.8) as

$$x(kT) = C^k\theta_p - C^k \sum_{i=p}^{k-1} F_i u[(i - p)T] \tag{3.5.10}$$

where

$$\theta_p = x_0 - \sum_{i+0}^{p-1} F_i u[(i - p)T] \tag{3.5.11}$$

As in the undelayed system, it can be shown that the boundary of the target can be reached in less than or an equal number of samples as to reach the interior of the hypersphere. The study will then be restricted to that of the boundary of the hypersphere.

By using (3.5.10) we have that

$$x^T(NT)x(NT) =$$

$$\left[\theta_p - \sum_{i=p}^{N-1} F_i u[(i - p)T]\right]^T \psi_N \left[\theta_p - \sum_{i=p}^{N-1} F_i u[(i - p)T]\right] \tag{3.5.12}$$

where, as in (3.2.7)

$$\psi_N = (C^N)^T C^N \tag{3.5.13}$$

Expanding (3.5.12) and performing the same steps as in obtaining (3.2.8), we get

$$x^T(NT)x(NT) = U_p^T \bar{Q}_N U_p - 2\bar{d}_N^T U_p + \theta_p^T \psi_N \theta_p \tag{3.5.14}$$

where

$$U_p = (u(0),\ u(T), \ldots, u[(N-1-p)T])^T \qquad (3.5.15)$$

$$\bar{Q}_N = \bar{F}^T_{N,p}\psi_N \bar{F}_{N,p} \qquad (3.5.16)$$

$$\bar{F}_{N,p} = [F_p, F_{p+1}, \ldots, F_{N-1}] \text{ nx(N-p)m matrix} \qquad (3.5.17)$$

Thus

$$\bar{Q}_N = \begin{bmatrix} F^T_p\psi_N F_p & F^T_p\psi_N F_{p+1} \cdot & \cdot \cdot F_p\psi_N F_{N-1} \\ F^T_{p+1}\psi_N F_p & & \vdots \\ \vdots & & \\ F^T_{N-1}\psi_N F_p & \cdot \quad \cdot \quad \cdot & F^T_{N-1}\psi_N F_{N-1} \end{bmatrix} \text{(N-p)mx(N-p)m matrix} \qquad (3.5.18)$$

$$\bar{d}_N = \theta^T_p \psi_N \bar{F}_{N,p} \qquad (3.5.19)$$

If $p = 0$ (no delay), then

$$U_p = U$$
$$\bar{Q}_N = Q_N$$
$$\bar{d}_N = d_N$$

where U, Q_N and d_N are defined in Section 3.2.

<u>Theorem 3.5.1</u> $\bar{Q}_N$ is a symmetric matrix with the following properties:

a) If $N \leq n/m + p$, $\bar{Q}_N$ is positive definite

b) If $N > n/m + p$, $\bar{Q}_N$ is positive semidefinite and singular

<u>Proof</u> $\bar{Q}^T_N = (\bar{F}^T_{N,p}\ \psi_N \bar{F}_{N,p})^T = \bar{F}^T_{N,p}\ \psi^T_N \bar{F}_{N,p} = \bar{F}^T_{N,p}\ \psi_N \bar{F}_{N,p} = \bar{Q}_N$

Hence $\bar{Q}_N$ is symmetric since ψ_N is symmetric.

a) By assumption (2),

$$\text{rank}[D,CD,\ldots,C^{N'-1}D] = N'm \text{ if } N' \leq n/m$$

Since C is nonsingular, this implies that

$$\text{rank}[C^pD,C^{p+1}D,\ldots,C^{p+N'-1}D] = N'm \quad \text{if } N' \leq n/m$$

Thus $[C^pD, C^{p+1}D\ldots,C^{p+N'-1}D]$ is of maximum rank if $N' \leq n/m$.

Let $N = N' + p$. Then

$\text{rank }(\bar{F}_{N,p}) = \text{rank}[C^pD, C^{p+1}D,\ldots,C^{N-1}D] = \text{maximum}$ if $N \leq n/m + p$.

By Theorem 2.1.5, $\bar{F}^T_{N,p}\bar{Q}_N F_{N,p}$ is positive definite.

b) If $N > n/m + p$, then $\bar{Q}_N$ is positive semidefinite because

$$y^T\bar{Q}_N y = y^T\bar{F}^T_{N,p}\psi_N\bar{F}_{N,p}y = z^T\psi_N z \geq 0$$

where $z = \psi_{N,p}y$, and ψ_N is positive definite. $\bar{Q}_N$ is not positive definite because $\bar{Q}_N$ is not of rank $(N-p)m$. This follows from Theorem 2.1.1:

$$\text{rank }(\bar{Q}_N) \leq \min\ [\text{rank}(F_{N,p}),\ \text{rank}(\psi_N)]$$

ψ_N is of rank n. $\bar{F}_{N,p}$ is of order $nx(N-p)m$ where $(N-p)m > n$. Thus $\bar{Q}_N$ is an $(N-p)mx(N-p)m$ matrix of rank less than $(N-p)m$. Hence $\bar{Q}_N$ is singular and not positive definite.

Since we are looking for the value of N and the corresponding sequence of controls such that

$$x^T(NT)x(NT) = R^2$$

we have that

$$f(U) = U_p^T \bar{Q}_N U_p - \overline{2d}_N^T U_p + \bar{e}_N = 0 \qquad (3.5.20)$$

where

$$\bar{e}_N = \theta_p^T \psi_N \theta_p - R^2 \qquad (3.5.21)$$

Equation (3.5.20) is of the same form as equation (3.2.14) for the undelayed system. We can then apply the same arguments as in Section 3.2 for obtaining an optimal sequence of controls. This leads to the following theorems which correspond to Theorems 3.2.2 - 3.2.4. Since the proofs are the same except that $\bar{Q}_N$, $\bar{d}_N$, $\bar{e}_N$ replace Q_N, d_N, e_N, respectively, the proofs are not repeated.

<u>Theorem 3.5.2</u> a) Let $N \leq n/m + p$. Then the expression for f(U) given by equation (3.5.20) can be reduced to

$$Y_p^T \Lambda_N Y_p = \bar{g}_N = 0 \qquad (3.5.22)$$

where

$$\bar{g}_N = \bar{e}_N - \bar{d}_N^T \bar{Q}_N^{-1} \bar{d}_N \qquad (3.5.23)$$

by the transformation

$$U_p = P_N Y_p + \bar{Q}_N^{-1} \bar{d}_N \qquad (3.5.24)$$

where P_N is such that $P_N^T P_N = I$ and $\Lambda_N = P_N^T \bar{Q}_N P_N$ (3.5.25) is a diagonal matrix whose diagonal elements are the eigenvalues of $\bar{Q}_N$.

b) If $N = n/m + p$, equation (3.5.22) becomes

$$Y_p^T \Lambda_N Y_p - R^2 = 0 \qquad (3.5.26)$$

<u>Theorem 3.5.3</u> Let $N \leq n/m + p$. Then the following is true.

a) If $\bar{e}_N = \bar{d}_N^T \bar{Q}_N^{-1} \bar{d}_N$, the unique sequence of controls such that $x(NT) \varepsilon\ \partial G$ (the boundary of G) is given by

$$U_p = \bar{Q}_N^{-1} \bar{d}_N$$

b) If $\bar{e}_N < \bar{d}_N^T \bar{Q}_N^{-1} \bar{d}_N$, a nonunique sequence of controls exists such that $x(NT) \varepsilon\ G$.

c) If $\bar{e}_N > \bar{d}_N^T \bar{Q}_N^{-1} \bar{d}_N$, no sequence of controls exists such that $x(NT) \varepsilon\ G$.

<u>Theorem 3.5.4</u> a) If $N > n/m + p$, a nonunique sequence of controls exists such that $x(NT) \varepsilon\ \partial G$ (the boundary of G). A sequence of controls can be found from

$$U_p = P_N Y + \bar{F}_{N,p}^T (\bar{F}_{N,p} \bar{F}_{N,p}^T)^{-1} x_0 \qquad (3.5.27)$$

where

$$Y^T \Lambda_N Y = R^2 \qquad (3.5.28)$$

P_N is chosen such that $P_N^T P_N = I$ and $\Lambda_N = P_N^T \bar{Q}_N P_N$ is a diagonal matrix with diagonal elements equal to the eigenvalues of $\bar{Q}_N$.

b) If $R = 0$, $U_p = \bar{F}_{N,p}^T (\bar{F}_{N,p} \bar{F}_{N,p}^T)^{-1} x_0$ is the sequence of controls which require minimum energy to reach the origin in N samples.

The above theorems are combined together to give Figure 3.5.1 which describes the method of determining the value of $\bar{N}$ and a sequence of time-optimal controls. The procedure is illustrated in Example 3.5.1.

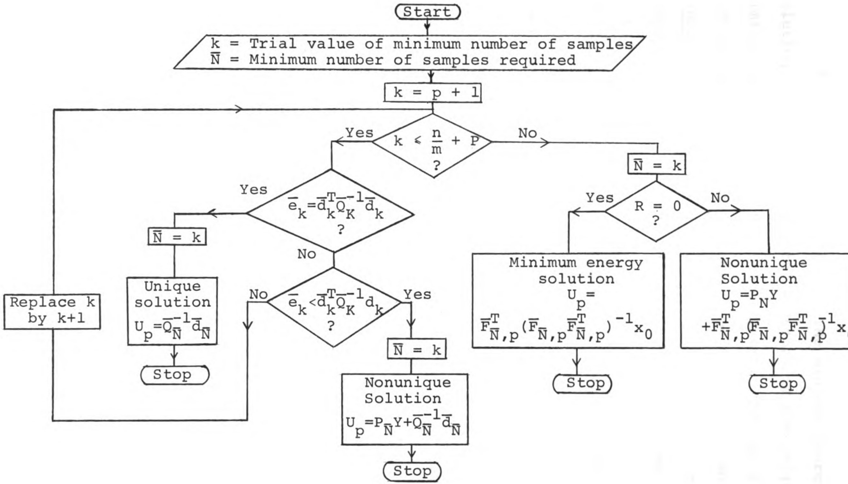


Fig. 3.5.1 Flowchart for determining minimum number of samples for system with input delay.

The method of determining the minimum energy solution when the time-optimal sequence is not unique is similar to that of the undelayed system. Theorem 3.3.1 for the undelayed system becomes the following theorem.

Theorem 3.5.5 Let $N \geq n/m + p$. The expansion of the conjoint matrix given in Theorem 2.1.10 for the matrix $\bar{Q}_N$ has the property that $G_{n+1} = G_{n+2} = \ldots = G_{(N-p)m} = 0$.

Similarly, Theorem 3.3.2 becomes the following theorem for the delayed system. The proofs of these two theorems are nearly the same as those for the undelayed system so they are omitted.

Theorem 3.5.6 a) If the sequence of time-optimal controls is not unique, the sequence of energy-optimal controls for the same value of N is given by

$$U_p = -(\gamma I - \bar{Q}_N)^{-1}\bar{d}_N$$

where γ is a root of the polynomial

$$\rho_o\gamma^{2(N-p)m} + \rho_1\gamma^{2(N-p)m-1} + \ldots + \rho_{2(N-p)m-1}\gamma$$

$$+ \rho_{2(N-p)m} = 0$$

The ρ_i are given by

$$\rho_i = \begin{cases} \bar{e}_N & \text{for } i=0 \\ 2(\bar{d}_N^T\bar{d}_N + \bar{e}_N s_1) & \text{for } i=1 \\ 2\bar{d}_N^T G_{i-1}\bar{d}_N + \sum_{j=0}^{i-2} \bar{d}_N^T (G_j\bar{Q}_N G_{i-j-2} + 2G_j s_{i-j-1})\bar{d}_N + \bar{e}_N \sum_{j=0}^{i} s_j s_{i-j} & \\ & \text{for } 2 \leq i \leq (N-p)m \\ \sum_{j=i-(N-p)m-1}^{(N-p)m-1} \bar{d}_N^T (G_j\bar{Q}_N G_{i-j-2} + 2G_j s_{i-j-1})\bar{d}_N + \bar{e}_N s_{j+1} s_{i-j-1} & \\ & \text{for } (N-p)m+1 \leq i \leq 2(N-p)m \end{cases}$$

where the G_i and s_i are generated recursively by the equations

$$G_0 = I \qquad s_0 = 1$$

$$G_1 = \bar{Q}_N G_0 + s_1 I \qquad s_1 = -tr(\bar{Q}_N)$$

$$\vdots \qquad \vdots$$

$$G_{(N-p)m-1} = \bar{Q}_N G_{(N-p)m-2} + s_{(N-p)m-1} I \qquad s_{(N-p)m} = -\frac{1}{(N-p)m} tr(\bar{Q}_N G_{(N-p)m-1})$$

$$G_{(N-p)m} = \bar{Q}_N G_{(N-p)m-1} + s_{(N-p)m} = 0$$

b) The quantities in the summations given above have the following property.

$$G_k \bar{Q}_N G_j = G_j \bar{Q}_N G_k \qquad j,k = 0,1,\ldots,(N-p)m-1$$

c) If $N \geq n/m + p$, the polynomial in part a) above reduces to a polynomial of order 2n, that is, the polynomial becomes

$$\bar{\rho}_0\gamma^{2n} + \bar{\rho}_1\gamma^{2n-1} + \ldots + \bar{\rho}_{2n-1}\gamma + \bar{\rho}_{2n} = 0$$

where the $\bar{\rho}_i$ are given by

$$\bar{\rho}_i = \begin{cases} \bar{e}_N & \text{for } i = 0 \\ 2(\bar{d}_N^T\bar{d}_N + \bar{e}_N s_1) & \text{for } i = 1 \\ 2\bar{d}_N^T G_{i-1}\bar{d}_N + \sum_{j=0}^{i-2} \bar{d}_N^T (G_j\bar{Q}_N G_{i-j-2} + 2G_j s_{i-j-1})\bar{d}_N & \\ \quad + \bar{e}_N \sum_{j=0}^{i} s_j s_{i-j} & \text{for } 2 \leq i \leq n \\ \sum_{j=i-n-1}^{n-1} \bar{d}_N^T (G_j\bar{Q}_N G_{i-j-2} + 2G_j s_{i-j-1})\bar{d}_N + \bar{e}_N s_{j+1} s_{i-j-1} & \\ & \text{for } n+1 \leq i \leq 2n \end{cases}$$

where the G_i and s_i are given in part a).

Examples 3.5.1 and 3.5.2 show how the above results can be used.

<u>Example 3.5.1</u> Given the system

$$x(k+1) = Cx(k) + Du(k-1)T \quad (3.5.29)$$

$$x(0) = (10 \quad 12)^T$$

$$u(-1) = 0$$

where C and D are given by (3.3.44) and (3.3.45), respectively.

We wish to determine the minimum number of samples $\bar{N}$ and a corresponding sequence of controls such that $x^T(\bar{N}T)x(\bar{N}T) \leq 1$.

<u>Solution</u> Using Figure 3.5.1, we have that $\bar{N} \geq p+1 = 2$. Let N=2. Then by (3.5.21) and (3.5.11),

$$\bar{e}_2 = \theta_1^T \psi_2 \theta_1 - R^2 = x^T(0)\psi_2 x(0) - R^2$$

Using equation (3.5.13), this becomes

$$\bar{e}_2 = (10 \quad 12) \begin{bmatrix} 1 & 0.8647 \\ 0.8647 & 0.7660 \end{bmatrix} \begin{bmatrix} 10 \\ -12 \end{bmatrix} - 1 = 416.82$$

From equation (3.5.19),

$$\bar{d}_2^T = x^T(0)\psi_2 \bar{F}_{2,1} = x^T(0)\psi_2 F_1$$

$$= (10 \quad 12) \begin{bmatrix} 1 & 0.8647 \\ 0.8647 & 0.7660 \end{bmatrix} \begin{bmatrix} 3.6708 \\ -4.6708 \end{bmatrix}$$

$$= -8.5225$$

$$\bar{Q}_2 = \bar{F}_{2,1}^T \psi_2 \bar{F}_{2,1} = F_1^T \psi_2 F_1 = 0.53491$$

Thus $\bar{d}_2^T \bar{Q}_2^{-1} \bar{d}_2 - \bar{e}_2 = -281.03$. Therefore $\bar{e}_2 > \bar{d}_2^T \bar{Q}_2^{-1} \bar{d}_2$, and we know by Theorem 3.5.3 that N is not large enough.

Setting N=3, we obtain

$$\bar{e}_3 = x^T(0)\psi_3 x(0) - R^2 = 457.426$$

$$\bar{d}_3^T = x^T(0)\psi_3 \bar{F}_{3,1} = x^T(0)\psi_3 [F_1, F_2] = (-16.5644 \ -8.2511) \tag{3.5.30}$$

$$\bar{Q}_3 = \bar{F}_{3,1}^T \psi_3 \bar{F}_{3,1} = \begin{bmatrix} 0.6431 & 0.4293 \\ 0.4293 & 0.5349 \end{bmatrix} \tag{3.5.31}$$

We then have $\bar{d}_3^T \bar{Q}_3^{-1} \bar{d}_3 - \bar{e}_3 = 1.000$. Therefore, $\bar{e}_3 < \bar{d}_3^T \bar{Q}_3^{-1} \bar{d}_3$ and by Theorem 3.5.3 we know that N=3 is optimal, and that the sequence of controls is not unique. To obtain the controls we use the transformation given by Theorem 3.5.2. For this example we have

$$\Lambda_3 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 1.0217 & 0 \\ 0 & 0.15627 \end{bmatrix}$$

where λ_1 and λ_2 are the eigenvalues of $\bar{Q}_3$. The P_3 matrix, composed of the normalized eigenvalues of $\bar{Q}_3$, is

$$P_3 = \begin{bmatrix} 0.7500 & -0.6614 \\ 0.6615 & 0.7500 \end{bmatrix} \quad (3.5.32)$$

From the second part of Theorem 3.5.2, we then have

$$1.0217y_0^2 + 0.15627y_1^2 = 1 \quad (3.5.33)$$

One solution of (3.5.33) is

$$Y_1 = (y_0, y_1)^T = (0.7500,\ 1.6497)^T \quad (3.5.34)$$

The controls are found from Theorem 3.5.2,

$$U_1 = P_3Y_1 + \bar{Q}_3^{-1}\bar{d}_3 \quad (3.5.35)$$

Substituting equations (3.5.30)-(3.5.32) and (3.5.34) into (3.3.35) gives

$$U_1 = \begin{bmatrix} u(0) \\ u(1) \end{bmatrix} = \begin{bmatrix} -33.835 \\ 13.042 \end{bmatrix} \quad (3.5.36)$$

$$\text{Energy} = u^2(-1) + u^2(0) + u^2(1) = 1145$$

When the controls given by (3.5.36) and the past control $u(-1) = 0$ are substituted into the state equation (3.5.29), we have the following sequence of states

$$x(0) = (10 \quad 12)^T$$
$$x(1) = (17.585 \quad 4.146)^T$$
$$x(2) = (7.9288 \quad -19.764)^T$$
$$x(3) = (0.2335 \quad 0.9732)^T$$

and $x^T(3)x(3) = 1.00$ as desired.

Example 3.5.2 Determine the sequences of controls for the system described in Example 3.5.1 which require minimum and maximum energy to reach the boundary of the target.

Solution From Theorem 3.5.6, the sequence of controls is given by

$$U_p = -(\gamma I - \bar{Q}_3)^{-1}\bar{d}_3 \tag{3.5.37}$$

where γ is a real root of the polynomial

$$\bar{\rho}_0\gamma_4 + \bar{\rho}_1\gamma^3 + \bar{\rho}_2\gamma^2 + \bar{\rho}_3\gamma + \bar{\rho}_4 = 0 \tag{3.5.38}$$

and we have used the fact that N=3,m=1, and n=2. The values of the $\bar{\rho}_i$ are found from the information given in Example 3.5.1 and Theorem 3.5.6.

$$s_0 = 1$$

$$s_1 = -\operatorname{tr}(\bar{Q}_3) = -\operatorname{tr}\begin{bmatrix} 0.6431 & 0.4293 \\ 0.4293 & 0.5349 \end{bmatrix} = -1.178$$

$$G_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$G_1 = \bar{Q}_3G_0 + s_1I = \bar{Q}_3 + s_1I = \begin{bmatrix} -0.5349 & 0.4293 \\ 0.4293 & -0.6431 \end{bmatrix}$$

$$s_2 = -\frac{1}{2}\operatorname{tr}(\bar{Q}_3G_1) = 0.1597$$

Thus by Theorem 3.5.6

$$\bar{\rho}_0 = \bar{e}_3 = 457.43 \tag{3.5.39}$$

$$\bar{\rho}_1 = 2(\bar{d}_3^T\bar{d}_3 + \bar{e}_3s_1) = -392.78 \tag{3.5.40}$$

$$\bar{\rho}_2 = 2\bar{d}_3^T G_1 \bar{d}_3 + \bar{d}_3^T (G_0 \bar{Q}_3 G_0 + 2G_0 s_1)\bar{d}_3 + \bar{e}_3 (s_0 s_2 + s_1 s_1 + s_2 s_0)$$

$$= 157.85 \qquad (3.5.41)$$

$\bar{\rho}_3$ and $\bar{\rho}_4$ are found from the equation

$$\bar{\rho}_i = \sum_{j=i-3}^{1} \bar{d}_3^T (G_j \bar{Q}_3 G_{i-j-2} + 2G_j s_{i-j-1})\bar{d}_3 + \bar{e}_3 s_{j+1} s_{i-j-1}$$

Thus

$$\bar{\rho}_3 = 0.3450 \text{ and } \bar{\rho}_4 = -0.0260 \qquad (3.5.42)$$

Substituting (3.5.39)-(3.5.42) into (3.5.38) gives

$$457.43\gamma^4 - 392.78\gamma^3 + 157.85\gamma^2 + 0.3450\gamma - 0.0260 = 0$$

The real roots of this equation are

$$\gamma_1 = -0.01351 \qquad (3.5.43)$$

$$\gamma_2 = 0.01159 \qquad (3.5.44)$$

If we substitute (3.5.43) into (3.5.37), we get

$$u(-1) = 0 \quad u(0) = -31.53 \quad u(1) = 9.637 \qquad (3.5.45)$$

The energy required is $u^2(-1) + u^2(0) + u^2(1) = 1087$.

The sequence of states resulting from the application of these controls is

$$\begin{aligned} x(0) &= (10 \quad 12)^T \\ x(1) &= (17.59 \quad 4.415)^T \\ x(2) &= (8.777 \quad -18.31)^T \\ x(3) &= (0.750 \quad -0.643)^T \\ x^T(3)x(3) &= 1.00 \end{aligned}$$

If we choose γ_2 given by (3.4.44), we get

$$u(-1) = 0 \quad u(0) = -35.07 \quad u(1) = 13.01 \qquad (3.5.46)$$

The energy required is 1399, and the corresponding sequence of states is

$$x(0) = (10 \quad 12)^T$$
$$x(1) = (17.59 \quad 4.415)^T$$
$$x(2) = (7.473 \quad -20.55)^T$$
$$x(3) = (-0.730 \quad 0.664)^T$$
$$x^T(3)x(3) = 1.00$$

From the above we see that the sequence of controls given by (3.5.45) requires minimum energy while that of (3.5.46) requires maximum energy.

3.6 Statement and Solution of Stochastic Time-Optimal Control Proglem

In the preceding sections it has been assumed that the state of the system could be determined exactly. In some cases due to noise this can not be done. This section is devoted to studying a stochastic version of the problem discussed in Section 3.1.

Problem Statement Given the system described by

$$x[(k+1)T] = Cx(kT) + Du(kT) + Ew(kT) \quad k=0,1,\ldots,N-1 \tag{3.6.1}$$

where

$x(kT)$ is a nx1 state vector
$u(kT)$ is a mx1 nonrandom vector to be determined
C is a nonsingular nxn matrix
D is a nxm matrix
E is a nxr matrix
$w(kT)$ is a sequence of rx1 zero mean independent random vectors.

The initial state is given by

$$x(0) = x_0 + v \qquad (3.6.2)$$

where x_0 is a known vector and v is a zero mean random vector assumed to be independent of w(kT) for k = 0,1,..., N-1. We wish to find the smallest value of N and a corresponding sequence of controls u(kT) such that

$$E(x^T(NT)x(NT)) \leq R^2$$

In the Appendix it is shown how equation (3.6.1) is obtained from the corresponding continuous-time system when the noise is "white."

Solution In preparation to solving this problem, the following definitions and assumptions are given.

Definition Let $W = (w(0), w(T), \ldots, w[(N-1)T])^T$. The matrix M_N is defined by

$$M_N = E(WW^T) \quad \text{(NrxNr matrix)} \qquad (3.6.3)$$

Definition The covariance matrix, V, is defined by

$$V = g(vv^T) \quad \text{(nxn matrix)} \qquad (3.6.4)$$

The matrix M_N and V describe the noise disturbance and are assumed to be known.

Assumption 1 It is assumed that the deterministic system corresponding to (3.6.1) and (3.6.2) is completely controllable. That is, if w(kT) = 0, k = 0, 1,..., N-1 and v = 0, then the resulting deterministic system is completely controllable. From Theorem 2.3.1 this means that

$$\text{rank}[D, CD, \ldots, C^{n-1}D] = n.$$

Assumption 2 It is assumed that

$$\text{rank}[D, CD, \ldots, C^{N-1}D] = \text{maximum for } N > 0.$$

These two assumptions are the same as those given previously in Section 3.1 and thus the comments concerning the second assumption apply here also.

The solution to equation (3.6.1) is

$$x(NT) = C^N x(0) + C^N \sum_{i=0}^{N-1} C^{i-N} Du[(N-1-i)T]$$

$$+ C^N \sum_{i=0}^{N-1} C^{i-N} Ew\,[(N-1-i)T] \qquad (3.6.5)$$

This result follows by the same argument as that used to obtain equation (3.2.5).

$$\text{Let } H_j = -C^{-(j+1)}E \quad \text{for } j = 0,1,\ldots,N-1 \qquad (3.6.6)$$

$$F_j = -C^{-(j+1)}D \quad \text{for } j = 0,1,\ldots,N-1 \qquad (3.6.7)$$

Then by (3.6.5)-(3.6.7)

$$x(NT) = C^N x(0) - C^N \sum_{i=0}^{N-1} F_{N-1-i}\, u[(N-1-i)T]$$

$$-C^N \sum_{i=0}^{N-1} H_{N-1-i}\, w[(N-1-i)T]$$

$$= C^N x(0) - C^N \sum_{i=0}^{N-1} F_i\, u(iT) - C^N \sum_{i=0}^{N-1} H_i\, w(iT)$$

Thus $x^T(NT)x(NT)$ is given by

$$x^T(NT)x(NT) = \left[x(0) - \sum_{i=0}^{N-1} F_i u(iT) - \sum_{i=0}^{N-1} H_i w(iT)\right] \psi_N \left[x(0) - \sum_{j=0}^{N-1} F_j u(jT) - \sum_{j=0}^{N-1} H_j w(jT)\right] \qquad (3.6.8)$$

where, as previously,

$$\psi_N = (C^N)^T C^N \tag{3.6.9}$$

Expanding (3.6.8), we obtain

$$x^T(NT)x(NT) = x^T(0)\psi_N x(0) - 2x^T(0)\psi_N \sum_{i=0}^{N-1} F_i u(iT)$$

$$- 2x^T(0)\psi_N \sum_{i=0}^{N-1} H_i w(iT)$$

$$+\left(\sum_{i=0}^{N-1} F_i u(iT)\right)^T \psi_N \left(\sum_{j=0}^{N-1} F_j u(jT)\right)$$

$$+ 2\left(\sum_{i=0}^{N-1} F_i u(iT)\right)^T \psi_N \left(\sum_{j=0}^{N-1} H_j w(jT)\right)$$

$$+\left(\sum_{i=0}^{N-1} H_i w(iT)\right)^T \psi_N \left(\sum_{j=0}^{N-1} H_j w(jT)\right) \tag{3.6.10}$$

Taking the expected value of (3.6.10) while noting the assumptions of independent and zero mean, we obtain

$$E[x^T(NT)x(NT)] = x_0^T\psi_N x_0 + E(v^T\psi_N v) - 2x_0^T\psi_N \sum_{i=0}^{N-1} F_i u(iT)$$

$$+\left(\sum_{i=0}^{N-1} F_i u(iT)\right)^T \psi_N\left(\sum_{j=0}^{N-1} F_j u(jT)\right)$$

$$+ E\left[\left(\sum_{i=0}^{N-1} H_i w(iT)\right)^T \psi_N \left(\sum_{j=0}^{N-1} H_j w(jT)\right)\right] = x_0^T\psi_N x_0$$

$$+ E(v^T\psi_N v) - 2d_N^T U + U^T Q_N U + E(W^T A_N W) \tag{3.6.11}$$

where

$$Q_N = \bar{F}_N^T \psi_N \bar{F}_N \quad \text{(NmxNm matrix)} \tag{3.6.12}$$

$$\bar{F}_N = [F_0, F_1, \ldots, F_{N-1}] \quad \text{(nxNm matrix)} \qquad (3.6.13)$$

$$A_N = H_N{}^T \psi_N H_N \quad \text{(NmxNm matrix)} \qquad (3.6.14)$$

$$H_N = [H_0, H_1, \ldots, H_{N-1}] \quad \text{(nxNm matrix)} \qquad (3.6.15)$$

$$d_N^T = x_0^T \psi_N \bar{F}_N \quad \text{(1 x Nm vector)} \qquad (3.6.16)$$

$$U = (u(0), u(T), \ldots, u[(N-1)T])^T \qquad (3.6.17)$$

It is noted that Q_N and $\bar{F}_N$ are the same as for the deterministic system of Section 3.1. The vector d_N given above is the same as that of Section 3.1 except that x_0 is interpreted here as the nominal initial state.

Since ψ_N is positive definite and symmetric, A_N is symmetric and at least positive semidefinite. The last term in equation (3.6.11) can then be written in an alternate fashion. Since A_N is positive semidefinite, it follows by Theorem 2.1.9 that there exists a matrix S such that $A_N = S^T S$.

Thus $E(W^T A_N W) = E\ (W^T S^T S\ W) = E(Z^T Z)$

where $Z = SW$

Therefore, $E(W^T A_N W) = tr[E(ZZ^T)]$

where $tr[E(ZZ^T)]$ denotes the trace of the $E(ZZ^T)$ matrix. Therefore,

$$E(W^T A_N W) = tr[E(SWW^T S^T)] = tr\ (SM_N S^T)$$

where M_N is defined by (3.6.3). Since tr (AB) = tr (BA) for conformal matrices A and B, it follows that

$$E(W^T A_N W) = \text{tr}\ (S^T S\ M_N) = \text{tr}\ (A_N M_N) \qquad (3.6.18)$$

By the same argument

$$E(v^T \psi_N v) = \text{tr}\ (\psi_N V) \qquad (3.6.19)$$

where V is defined by (3.6.4). Substituting (3.6.18) and (3.6.19) into (3.6.11) gives

$$E(x^T(NT)x(NT)) = U^T Q_N U - 2d_N^T U + x_0^T\ \psi_N x_0 + \text{tr}(A_N M_N)$$

$$+ \text{tr}(\psi_N V) \qquad (3.6.20)$$

Using the same argument as for the deterministic system, it can be shown that if there exists a number N_1 such that $E(x^T(N_1T)x(N_1T)) < R^2$, then there exists a number $N \leq N_1$ such that

$$E(x^T(NT)x(NT)) = R^2 \qquad (3.6.21)$$

If we let

$$\underline{e}_N = x_0^T \psi_N x_0 + \text{tr}(A_N M_N) + \text{tr}(\psi_N V) - R^2 \qquad (3.6.22)$$

then equations (3.6.20)-(3.6.21) give

$$f(U) \triangleq U^T Q_N U - 2d_N{}^T U + \underline{e}_N = 0 \qquad (3.6.23)$$

Thus, we are looking for the smallest value of N and a corresponding sequence of controls U such that (3.6.23) is satisfied. Because this equation is of the same form as equation (3.2.14) except that $\underline{e}_N$ replaces e_N, we have by the same argument the following theorems.

Theorem 3.6.1 a) Let $N \leq n/m$. The expression for f(U) given in equation (3.6.23) can be reduced to the form

$$Y^T \Lambda_N Y + \bar{g}_N = 0$$

where

$$\bar{g}_N = \underline{e}_N - d_N^T Q_N^{-1} d_N \tag{3.6.24}$$

by the transformation

$$U = P_N Y + Q_N^{-1} d_N \tag{3.6.25}$$

where P_N is such that $P_N^T P_N = I$ and $\Lambda_N = P_N^T Q_N P_N$ is a diagonal matrix with diagonal elements equal to the eigen-values of Q_N.

b) If $N = n/m$, equation (3.6.23) reduces

$$Y^T \Lambda_N Y + tr(A_N M_N) + tr\ (\psi_N V) - R^2 = 0$$

Theorem 3.6.2 Let $N \leq n/m$. Then the following is true.

a) If $\underline{e}_N = d_N^T Q_N^{-1} d_N$, the unique sequence of controls such that $E(x^T(NT)x(NT)) = R^2$ is given by

$$U = Q_N^{-1} d_N$$

b) If $\underline{e}_N < d_N^T Q_N^{-1} d_N$, a nonunique sequence of controls exists such that $E(x^T(NT)x(NT)) \leq R^2$.

c) If $\underline{e}_N > d_N^T Q_N^{-1} d_N$, no sequence of controls exists such that $E(x^T(NT)x(NT)) \leq R^2$.

Theorem 3.6.3 If $N > n/m$ and a solution to the optimal control problem exists, a sequence of controls is given by

$$U = P_N Y + \bar{F}_N^T (\bar{F}_N \bar{F}_N^T)^{-1} x_0$$

where

$$Y^T \Lambda_N Y = R^2 - tr(A_N M_N) - tr(\psi_N V)$$

and P_N is such that $P_N^T P_N = I$ and $\Lambda_N = P_N^T Q_N P_N$ is a diagonal matrix with diagonal elements equal to the eigenvalues of Q_N.

Theorem 3.6.4 a) Let $N > n/m$ and $tr(A_N M_N) + tr(\psi_N V) = R^2$. Then a sequence of controls such that $E(x^T(NT)x(NT)) = R^2$ is given by $U = \bar{F}_N^T(\bar{F}_N \bar{F}_N^T)^{-1} x_0$. This value of U has the property that $\|U\|^2$ is minimum.

b) If $tr(A_N M_N) + tr(\psi_N V) < R^2$, a nonunique sequence of controls exists such that $E(x^T(NT)x(NT)) \leq R^2$.

c) If $tr(A_N M_N) + tr(\psi_N V) > R^2$, then no sequence of controls exists such that $E(x^T(NT)x(NT)) \leq R^2$.

A significant difference between the solution to the stochastic and deterministic problems is that for the latter the target can always be reached in a number of samples equal to the smallest integer greater than or equal to n/m while the stochastic system has the additional requirement given by Theorem 3.6.4. It is possible that the noise and system parameters may be such that the third part of Theorem 3.6.4 would not hold for any N. Theorems 3.6.1-3.6.4 are combined into the flow chart given in Figure 3.6.1.

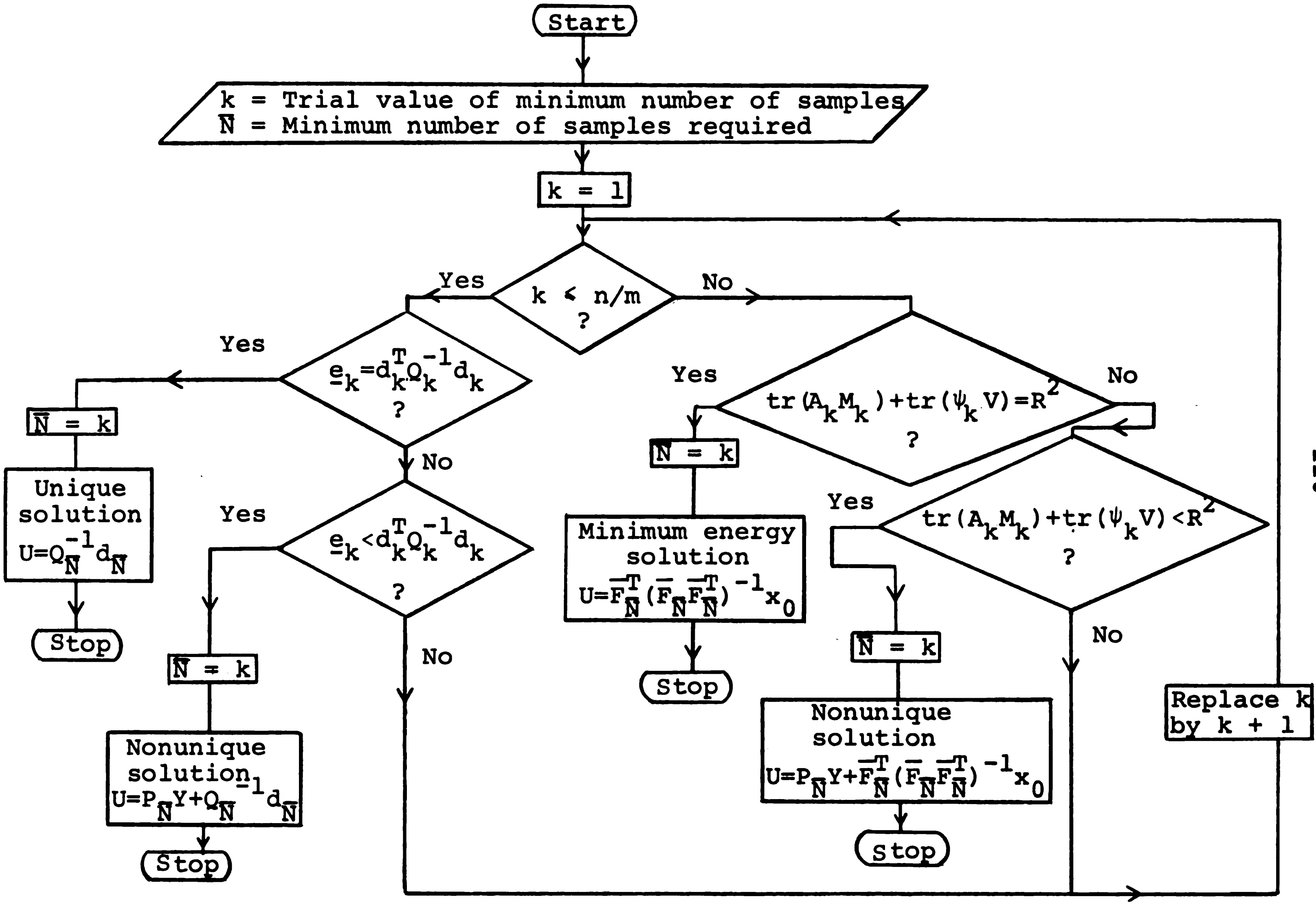


Fig. 3.6.1. Flowchart for determining minimum number of samples for stochastic system.

It should be noted that although we have found a sequence of controls such that $E(x^T(NT)x(NT)) = R^2$, this does not imply that $E(x(NT))$ lies on the boundary of the hypersphere (as in the deterministic case). Instead, we shall have $E(x^T(NT))E(x(NT)) \leq R^2$.

To show this, let $x(NT) = (x_1(NT), x_2(NT), \ldots, x_n(NT))^T$ and $\sigma_{x_i}^2$ equal the variance of $x_i(NT)$. Then

$$
\begin{aligned}
E(x^T(NT)x(NT)) &= E(x_1^2(NT)) + E(x_2^2(NT)) + \ldots + E(x_n^2(NT)) \\
&= E^2(x_1(NT)) + \sigma_{x_1}^2 + E^2(x_2(NT)) + \sigma_{x_2}^2 \\
&\quad + \ldots + E^2(x_n(NT) + \sigma_{x_n}^2 \\
&= E(x^T(NT))E(x(NT)) + \sum_{i=1}^{n} \sigma_{x_i}^2
\end{aligned}
$$

Therefore,

$$R^2 = E(x^T(NT))E(x(NT)) + \sum_{i=1}^{n} \sigma_{x_i}^2 \geq E(x^T(NT))E(x(NT))$$

$E(x(kT)$ can be found from equation (3.6.5). Taking the expected value of this equation, we obtain

$$E(x(kT)) = C^k x_0 - C^k \sum_{i=0}^{k-1} F_i u(iT) \qquad (3.6.26)$$

The method of determining the sequence of minimum energy controls once $\bar{N}$ is known is similar to that for the deterministic case. To determine the sequence of controls such that $J = \sum_{i=0}^{N-1} u(iT)u(iT)$ is minimum subject to the constraint $E(x^T(NT)x(NT)) = R^2$, we have only to replace e_N by $\underline{e}_N$ in Theorem 3.3.2. We then find the coefficients ρ_i and the roots γ_i of the polynomial (3.3.8). The γ_i are substituted into equation (3.3.3) to determine the sequence of energy-optimal controls.

An example is given to illustrate the above theory.

Example 3.6.1 Consider the system described by

$$x(k+1) = Cx(k) + Du(k) + Ew(k) \quad k=0,1,\ldots,N-1 \tag{3.6.27}$$

$$x(0) = x_0 \tag{3.6.28}$$

where

$$C = \begin{bmatrix} 1 & 1-e^{-1} \\ 0 & e^{-1} \end{bmatrix} \tag{3.6.29}$$

$$D = E = \begin{bmatrix} e^{-1} \\ 1-e^{-1} \end{bmatrix} \tag{3.6.30}$$

$$x_0 = (10, \quad -12)^T \tag{3.6.31}$$

The expressions for C, D and x_0 correspond to the values given in Example 3.3.1 for the deterministic system. From previous results we know that the deterministic system is controllable. We assume that v, w(k), k=0,1,..., N-1 are independent Gaussian random variables such that

$$E(v) = 0 \tag{3.6.32}$$

$$E(w(kT)]=0 \quad k=0,1,\ldots,N-1 \tag{3.6.33}$$

$$V = E(vv^T) = \begin{bmatrix} 4.8 & -4 \\ -4. & 4 \end{bmatrix} \tag{3.6.34}$$

$$M_N = E(WW^T) + \frac{1}{10 + (i+1)^2} I_N \quad i = 0,1,\ldots,N-1$$

$$= \begin{bmatrix} \frac{1}{11} & 0 & \cdots & 0 \\ 0 & \frac{1}{14} & 0 \cdots & 0 \\ \vdots & 0 & & \vdots \\ \vdots & \vdots & & 0 \\ 0 & 0 & \cdots & \frac{1}{10+N^2} \end{bmatrix} \qquad (3.6.35)$$

Equation (3.6.35) can be intrepreted as meaning that the noise variance decreases as time increases.

The sampling period is T=1. We wish to determine the smallest value of k and the corresponding sequence of controls such that $E(x^T(kT)x(kT)) \leq 1$.

Solution To determine the minimum number of samples N, we use the procedure outlined in Figure 3.6.1. Setting k=1, we obtain

$$Q_1 = F_0^T \psi_1 F_0$$

$$= (0.71828 \ -1.71828) \begin{bmatrix} 1. & 0.63212 \\ 0.63212 & 0.53491 \end{bmatrix} \begin{bmatrix} 0.71828 \\ -1.71828 \end{bmatrix} = 0.5349$$

(3.6.36)

$$d_1 = x_0^T \psi_1 F_0$$

$$= (10 \ -12) \begin{bmatrix} 1. & 0.63212 \\ 0.63212 & 0.53491 \end{bmatrix} \begin{bmatrix} 0.71828 \\ -1.71828 \end{bmatrix} = 1.90226$$

(3.6.37)

$\underline{e}_1 = x_0^T \psi_1 x_0 + \text{tr}\,(P_1 M_1) + \text{tr}\,(\psi_1 V) = 26.2497$

and $d_1^T Q_1^{-1} d_1 - \underline{e}_1 = -19.485$. Therefore,

$$\underline{e}_1 > d_1^T Q_1^{-1} d_1 \qquad (3.6.38)$$

Thus by Theorem 3.6.2, k=1 is not large enough. If we try k=2 we arrive at the same conclusion. Letting k=3, we obtain

$$Q_3 = \begin{bmatrix} F_0^T \psi_3 F_0 & F_0^T \psi_3 F_1 & F_0^T \psi_3 F_2 \\ F_1^T \psi_3 F_0 & F_1^T \psi_3 F_1 & F_1^T \psi_3 F_2 \\ F_2^T \psi_3 F_0 & F_2^T \psi_3 F_1 & F_2^T \psi_3 F_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.84354 & 0.72169 & 0.39048 \\ 0.72169 & 0.64306 & 0.42932 \\ 0.39048 & 0.42933 & 0.53490 \end{bmatrix} \qquad (3.6.39)$$

$$\psi_3 = \begin{bmatrix} 1. & 0.95021 \\ 0.95021 & 0.90538 \end{bmatrix} \qquad (3.6.40)$$

From (3.6.34), (3.6.35), (3.6.39) and (3.6.40)

$$\text{tr}(A_3 M_3) + \text{tr}(\psi_3 V) = \text{tr}(Q_3 M_3) + \text{tr}(\psi_3 V)$$

$$= 0.15077 + 0.81983 = 0.9706 \qquad (3.6.41)$$

Since R = 1, $\text{tr}(A_3 M_3) + \text{tr}(\psi_3 V) < R^2$. Thus by Theorem 3.6.4, we know that $\bar{N}=3$.

A sequence of optimal controls is found from Theorem 3.6.3.

$$\Lambda_3 = P_3^T Q_3 P_3 \tag{3.6.42}$$

The eigenvalues of Q_3 are

$$\lambda_1 = 0.2748 \quad \lambda_3 = 1.7467 \quad \lambda_3 = 0$$

The corresponding normalized eigenvectors form the columns of the P_3 matrix:

$$P_3 = \begin{bmatrix} -0.4678 & 0.6698 & 0.5767 \\ -0.1061 & 0.6053 & -0.7889 \\ 0.8774 & 0.4303 & 0.2122 \end{bmatrix}$$

and

$$\mathrm{tr}(A_3M_3) + \mathrm{tr}(\psi_3V) - R^2 = -0.0294$$

From Theorem 3.6.3 we have $Y^T\Lambda_3Y + \mathrm{tr}\,(A_3M_3) + \mathrm{tr}\,(\psi_3V) - R^2 = 0$

or

$$[y_0 \; y_1 \; y_2] \begin{bmatrix} 0.2748 & 0. & 0. \\ 0. & 1.7467 & 0. \\ 0. & 0. & 0. \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix} = 0.0294$$

or

$$0.2748\, y_0^2 + 1.7468\, y_1^2 = 0.0294$$

or

$$\frac{y_0^2}{0.10698} + \frac{y_1^2}{0.1683} = 1 \tag{3.6.42}$$

From (3.6.42) it is seen that as far as the time optimal sequence of controls is concerned y_2 is arbitrary. The choice of y_2 does affect the energy required, however. A possible solution of (3.6.42) is

$$y_0 = 0.3271 \quad y_1 = 0 \quad y_2 = 0$$

The sequence of controls is found from Theorem 3.6.3.

$$U = P_3Y + \bar{F}_3^{+}x_0 = P_3Y + \bar{F}_3^{T}(\bar{F}_3\bar{F}_3^{T})^{-1}x_0 \qquad (3.6.43)$$

Substituting the known quantities into the right side of (3.6.43), we obtain a sequence of time-optimal controls.

$$u(0) = 0.5657 \quad u(1) = 0.6509 \quad u(2) = 0.8826 \qquad (3.6.44)$$

To check the result a computer simulation of the system was made. Gaussian random variables were generated by means of the IBM subroutines RANDU and GAUSS [IBM1]. Although no claim was made about the sequence being independent, the sample auto-correlation indicated this to be the case. We then have the problem of determining the covariance matrix given by (3.6.34), i.e.,

$$E(vv^T) \quad \begin{bmatrix} 4.8 & -4. \\ -4. & 4. \end{bmatrix} \qquad (3.6.45)$$

Because of the independence of the random variables the covariance matrix will be diagonal unless a transformation is made. To overcome this problem let

$$v = Sz \qquad (3.6.46)$$

where S is to be determined. Let us also set

$$E(zz^T) = \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix} \quad (3.6.47)$$

and

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \quad (3.6.48)$$

We then have

$$E(vv^T) = SE(zz^T)S^T = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix} \begin{bmatrix} s_{11} & s_{21} \\ s_{12} & s_{22} \end{bmatrix}$$

(3.6.49)

Equating the right-hand sides of (3.6.45) and (3.6.49), we have

$$\alpha_{11}s_{11}^2 + \alpha_{22}s^2_{12} = 4.8$$

$$s_{11}s_{21}\alpha_{11} + s_{12}s_{22}\alpha_{22} = -4.$$

$$s^2_{21}\alpha_{11} + s^2_{22}\alpha_{22} = 4.$$

A solution to these three equations and six unknowns is

$\alpha_{11} = 2/3$ $s_{11} = 2$ $s_{12} = -1$ $s_{22} = 5/4$ $s_{21} = -1$ $\alpha_{22}=32/15$

Therefore,

$$E(zz^T) = \begin{bmatrix} 2/3 & 0 \\ 0 & 32/15 \end{bmatrix} \quad (3.6.50)$$

and

$$v = Sz = \begin{bmatrix} 2 & -1 \\ -1 & 5/4 \end{bmatrix} z \quad (3.6.51)$$

The independent random variables z are generated in the computer and multiplied by S to give the random variables with covariance matrix given by (3.6.45). Because of the assumption of independence, there is no problem in generating the w(iT) random variables.

By using the controls given in (3.6.44) and the state equations (3.6.27) and (3.6.28) a sequence of states can be found. In order to determine if the sequence of controls is correct, a monte carlo technique can be used. For this example, this consisted in simulating the system 500 times using different noise sequences. The following sample means were obtained.

$$E(x(0)) = \begin{bmatrix} 10.056 \\ -12.058 \end{bmatrix} \qquad E(x(1)) = \begin{bmatrix} 2.644 \\ -4.075 \end{bmatrix}$$

$$E(x(2)) = \begin{bmatrix} 0.3102 \\ -1.0832 \end{bmatrix} \qquad E(x(3)) = \begin{bmatrix} -0.04919 \\ 0.16058 \end{bmatrix}$$

$$E(x^T(3))\ E(x(3)) = 0.0282$$

$$E(x^T(3)\ x(3)) = 1.007$$

The results check since $E(x^T(3)x(3)) \approx 1.$ as desired.

CHAPTER IV

TIME-OPTIMAL CONTROL WITH HYPERRECTANGULAR TARGET SET

4.1 Statement of Control Problem

In this chapter the following problem is considered. Given the linear discrete-time system

$$x[(k+1)T] = Cx(kT) + Du(kT) \quad (4.1.1)$$

$$x(0) = x_0$$

where

x(kT) is a nx1 vector
u(kT) is a mx1 vector
D is a nxm matrix
C is a nxn nonsingular matrix

It is assumed that the initial state lies outside the target described by

$$H = \{x(NT) : -M_i \leq x_i(NT) \leq M_i \qquad i = 1,2,\ldots,n\} \quad (4.1.2)$$

where $x(NT) = (x_1(NT), x_2(NT), \ldots, x_n(NT))^T, M_i \geq 0$, and N is the minimum number of samples such that $x(NT) \varepsilon H$. We wish to determine N and a corresponding sequence of controls. If the sequence is not unique, we want to choose a sequence which minimizes the total fuel required to reach the target, that is, we want to minimize

$$J = \sum_{k=0}^{N-1} \sum_{j=1}^{m} |u_j(kT)| \tag{4.1.3}$$

where $u(kT) = (u_1(kT), u_2(kT), \ldots, u_m(kT))^T$. It is assumed that the magnitude of the controls is constrained, that is,

$$|u_i(kT)| \leq G_{i,k} \tag{4.1.4}$$

where $G_{i,k} > 0$. (The case where inequality (4.1.4) need not hold will also be discussed.)

It is assumed that a solution exists for some value of N. No other assumptions will be made.

4.2 Formulation of Time-Optimal Control Problem as a Solution to a Set of Linear Inequalities

It has been previously shown (equation (3.2.5)) that the solution of equation (4.1.1) is

$$x(kT) = C^k x_0 - C^k \sum_{j=0}^{k-1} F_j u(jT).$$

Let

$$c_k^k = \text{i-th row of } C^k \tag{4.2.1}$$

Then $x_i(kT) = c_i^k x_0 - c_i^k \sum_{j=0}^{k-1} F_j u(jT) \quad i = 1,2,\ldots,n$ and

$$x_i^2(kT) = \left[x_0 - \sum_{j=0}^{k-1} F_j u(jT)\right]^T \psi_{i,k} \left[x_0 - \sum_{j=0}^{k-1} F_j u(jT)\right]$$

where

$$\psi_{i,k} = (c_i^k)^T c_i^k \text{ (outer product of } c_i^k \text{ with itself)} \tag{4.2.2}$$

Letting k = N where N corresponds to the time at which the target is reached, we have

$$x_i^2(NT) = U^T\bar{F}_N^T\psi_{i,N}\bar{F}_N U - 2x_0^T\psi_{i,N}\bar{F}_N U + x_0^T\psi_{i,N}x_0 \qquad (4.2.3)$$

where

$$\bar{F}_N = (F_0, F_1, \ldots, F_{N-1}) \qquad (4.2.4)$$

$$U = (u(0), u(T), \ldots, u[(N-1)T])^T \text{ (Nmx1 vector)} \qquad (4.2.5)$$

and the F_i are given by equation (3.2.4).

From equation (4.1.2) we are looking for N and u(iT) such that

$$x_i^2(NT) \leqslant M_i^2 \qquad (4.2.6)$$

From equations (4.2.3) and (4.2.6) we then have

$$U^T Q_{i,N} U - 2d_{i,N}^T U + e_{i,N} \leqslant 0 \qquad i=1,2,\ldots,n \quad (4.2.7)$$

where

$$Q_{i,N} = \bar{F}_N^T\psi_{i,N}\bar{F}_N \qquad (4.2.8)$$

$$d_{i,N}^T = x_0^T\psi_{i,N}\bar{F}_N \qquad (4.2.8)$$

$$e_{i,N} = x_0^T\psi_{i,N}x_0 - M_i^2 \qquad (4.2.10)$$

Equation (4.2.7) represents a constraint on the controls which must be satisfied in order that the target be reached in N samples.

<u>Theorem 4.2.1</u> Let $Q_{i,N} \neq 0$. Then $Q_{i,N}$ is positive semi-definite, symmetric and of rank 1.

<u>Proof</u> From equation (4.2.2) we see that $\psi_{i,N}$ is formed from the rows of the matrix C^N. Since this latter matrix is nonsingular, it follows that all the rows are nonzero. Then by Theorem 2.1.13, it follows that $\psi_{i,N}$ is symmetric and of rank 1. Since $Q_{i,N} = \bar{F}_N^T \psi_{i,N} \bar{F}_N$, we have that $Q_{i,N}$ is symmetric and by Theorem 2.1.1 rank $(Q_{i,N}) \leq$ $\min\{\text{rank}(\bar{F}_N), \text{rank}(\psi_{i,N})\} \leq 1$. By assumption, $Q_{i,N} \neq 0$. Therefore, $\text{rank}(Q_{i,N}) \geq 1$. Consequently, by the above, $\text{rank}(Q_{i,N}) = 1$. We also have for an arbitrary vector y, $y^T \bar{F}_N \psi_{i,N} \bar{F}_N\, y = y^T \bar{F}_N^T (c_i^N)^T\; c_i^N \bar{F}_N y = z^T z \geq 0$ where $z = c_i^N \bar{F}_N y$. Thus $Q_{i,N} = \bar{F}_N^T \psi_{i,N} \bar{F}_N$ is positive semidefinite.

QED

It is because $\text{rank}(Q_{i,N}) = 1$ that much of the succeeding results depend. This fact is used to show that (except in a degenerate case) inequality (4.2.7) represents the equation of two hyperplanes and the region between the hyperplanes. (In the degenerate case the two hyperplanes come together to form a single hyperplane. This case occurs only when $M_i = 0$ in equation (4.1.2).) It shall be assumed at present that $Q_{i,N} \neq 0$. The case when $Q_{i,N} = 0$ will be considered later.

Let us set

$$f(U) = U^T Q_{i,N} U - 2d_{i,N}^T U + e_{i,N} \qquad (4.2.11)$$

We would like to perform a linear transformation on $f(U)$ to convert it into a simpler form [FRA2] as was done in

Chapter III. In this case, advantage will be taken of the fact that rank$(Q_{i,N}) = 1$. Let

$$U = P_{i,N}Y_i + C_{i,N} \tag{4.2.12}$$

Then f(U) becomes

$$g(Y_i) = (P_{i,N}Y_i + C_{i,N})^T Q_{i,N}(P_{i,N}Y_i + C_{i,N})$$
$$- 2d_{i,N}^T(P_{i,N}Y_i + C_{i,N}) + e_{i,N} = Y_i^T P_{i,N}^T Q_{i,N} Y_i$$
$$+ 2(C_{i,N}^T Q_{i,N} - d_{i,N}^T) P_{i,N} Y_i + C_{i,N}^T Q_{i,N} C_{i,N}$$
$$- 2d_{i,N}^T C_{i,N} + e_{i,N} \tag{4.2.13}$$

By letting

$$L_{i,N} = P_{i,N}^T(Q_{i,N}C_{i,N} - d_{i,N}) \tag{4.2.14}$$

$$g_{i,N} = C_{i,N}^T Q_{i,N} C_{i,N} - 2d_{i,N}^T C_{i,N} + e_{i,N} \tag{4.2.15}$$

equation (4.2.13) becomes

$$g(Y_i) = Y_i^T P_{i,N}^T Q_{i,N} P_{i,N} Y_i + 2L_{i,N}^T Y_i + g_{i,N} \tag{4.2.16}$$

We would like to choose $C_{i,N}$ such that $L_{i,N}$ in equations (4.2.14) and (4.2.16) is zero. If $Q_{i,N}$ had an inverse, we could choose $C_{i,N} = Q_{i,N}^{-1}d_{i,N}$ but, except when $N = 1$, we know by Theorem 4.1.1 that $Q_{i,N}^{-1}$ does not exist. From Theorem 2.2.7 we know that a vector of the form

$$C_{i,N} = Q_{i,N}^+ d_{i,N} + \alpha A_{i,N} d_{i,N} \quad \alpha \text{ a scalar} \tag{4.2.17}$$

minimizes the length of $Q_{i,N}C_{i,N} - d_{i,N}$ where $Q^{+}_{i,N}$ is the pseudoinverse of $Q_{i,N}$ and $A_{i,N}$ is any matrix such that

$$Q_{i,N}A_{i,N}d_{i,N} = 0 \qquad (4.2.18)$$

An expression for $A_{i,N}$ will be found such that $L_{i,N} = 0$ in equation (4.2.16).

<u>Theorem 4.2.2</u> The positive eigenvalue γ_i of $Q_{i,N}$ is

$$\gamma_i = tr(Q_{i,N}) \qquad (4.2.19)$$

Furthermore, in the expansion of $adj(\lambda I - Q_{i,N})$ given in Theorem 2.1.10 we have

$$G_i = 0 \quad \text{for } i \geqslant 2 \qquad (4.2.20)$$

<u>Proof</u> Since $Q_{i,N}$ is positive semidefinite, symmetric and of rank 1 by Theorem 4.2.1, we have

$$|\lambda I - Q_{i,N}| = \lambda^{Nm} - \gamma_i\lambda^{Nm-1} \text{ where } \gamma_i > 0 \qquad (4.2.21)$$

Comparing equation (4.2.21) with the similar expression in Theorem 2.1.10, we see that

$$s_0 = 1 \qquad (4.2.22)$$

$$s_1 = -\gamma_i \qquad (4.2.23)$$

$$s_j = 0 \quad \text{for } j > 1 \qquad (4.2.24)$$

A comparison of equation (4.2.23) with that for s_1 given in Theorem 2.1.10 shows that $\gamma_i = tr(Q_{i,N})$ as desired. Also, from Theorem 2.1.10 and equation (4.2.24) we see that

$$0 = s_2 = -\frac{1}{2}tr(Q_{i,N}G_1) \qquad (4.2.25)$$

Then by Theorem 2.1.19 we know that

$$tr(Q_{i,N}G_1) = z_1 + z_2 + \ldots + z_{Nm} = 0 \qquad (4.2.26)$$

where the z_i are the eigenvalues of $Q_{i,N}G_1$. By Theorem 2.1.1 $\text{rank}(Q_{i,N}G_1) \leq \min[\text{rank}(Q_{i,N}), \text{rank}(G_1)] \leq 1$. Thus we know that $Q_{i,N}G_1$ has at most one nonzero eigenvalue z_1. However, by equation (4.2.26) we have that $z_1 = 0$ also, since $z_2 = z_3 = \ldots = z_{Nm} = 0$. By Theorem 2.1.10 we also have that $Q_{i,N}G_1 = Q_{i,N}Q_{i,N} + s_1Q_{i,N}$. Therefore, $Q_{i,N}G_1$ is a symmetric matrix with all eigenvalues equal to zero. By Theorem 2.1.7 it follows that $\text{rank}(Q_{i,N}G_1) = 0$ and

$$Q_{i,N}G_1 = 0 \qquad (4.2.27)$$

If we substitute equations (4.2.27) and (4.2.24) into Theorem 2.1.10 we have $G_2 = 0$. This, plus equation (4.2.24) imply that $G_i = 0$ for $i \geq 2$.

QED

<u>Theorem 4.2.3</u> The constituent matrices Z_{11} and Z_{21} for $Q_{i,N}$ are

$$Z_{11} = \frac{\gamma_i I - Q_{i,N}}{\gamma_i} \qquad (4.2.28)$$

$$Z_{21} = \frac{Q_{i,N}}{\gamma_i} \qquad (4.2.29)$$

where

$$\gamma_i = tr(Q_{i,N}) \qquad (4.2.30)$$

is the nonzero eigenvalue of $Q_{i,N}$.

<u>Proof</u> By Theorem 2.1.15 the minimal polynomial $m(\lambda)$ is

$$m(\lambda) = \frac{D(\lambda)}{D_{Nm-1}(\lambda)} \qquad (4.2.31)$$

where $D(\lambda) = |\lambda I - Q_{i,N}|$ and $D_{Nm-1}(\lambda)$ is the greatest common divisor of all the minors of order Nm-1 of $\lambda I - Q_{i,N}$. Thus $D_{Nm-1}(\lambda)$ is the greatest common factor of $\operatorname{adj}(\lambda I - Q_{i,N})$. By Theorem 2.1.10, $\operatorname{adj}(\lambda I - Q_{i,N}) = I\lambda^{Nm-1} + G_1\lambda^{Nm-2} + G_2\lambda^{Nm-3}+\ldots+G_{Nm-1}$. Then by Theorem 4.2.2 this reduces to

$$\operatorname{adj}(\lambda I - Q_{i,N}) = \lambda^{Nm-2}(\lambda I + G_1)$$

From Theorems 2.1.10 and 4.2.2 we have $G_1 = Q_{i,N} - \gamma_i I$. Therefore,

$$\operatorname{adj}(\lambda I - Q_{i,N}) = \lambda^{Nm-2}[(\lambda - \gamma_i)I + Q_{i,N}]$$

The greatest common divisor, $D_{Nm-1}(\lambda)$, of $\operatorname{adj}(\lambda I - Q_{i,N})$ is then

$$D_{Nm-1}(\lambda) = \lambda^{Nm-2} \qquad (4.2.32)$$

Substituting equations (4.2.21) and (4.2.32) into equation (4.2.31), we obtain

$$m(\lambda) + \frac{(\lambda - \gamma_i)\lambda^{Nm-1}}{\lambda^{Nm-2}} = \lambda(\lambda - \gamma_i) \qquad (4.2.33)$$

Then if $h(\lambda)$ is a polynomial

$$\frac{h(\lambda)}{m(\lambda)} = \frac{k_1}{\lambda} + \frac{k_2}{\lambda - \gamma_i}$$

where

$$k_1 = \frac{h(\lambda)}{(\lambda - \gamma_i)}\Bigg|_{\lambda = 0} = -\frac{h(0)}{\gamma_i}$$

$$k_2 = \frac{h(\lambda)}{\lambda}\Big|\lambda = \gamma_i = \frac{h(\gamma_i)}{\gamma_i}$$

Therefore,

$$\frac{h(\lambda)}{m(\lambda)} = -\frac{h(0)}{\lambda\gamma_i} + \frac{h(\gamma_i)}{\gamma_i(\lambda - \gamma_i)}$$ or by equation (4.2.33),

$$h(\lambda) = -\frac{(\lambda - \gamma_i)}{\gamma_i} h(0) + \frac{\lambda}{\gamma_i} h(\gamma_i)$$

Using the properties of $h(\lambda)$ [KOE1], we can substitute $Q_{i,N}$ for λ and obtain

$$h(Q_{i,N}) = -\frac{(Q_{i,N} - \gamma_i I)}{\gamma_i} h(0) + \frac{Q_{i,N}}{\gamma_i} h(\gamma_i) \qquad (4.2.34)$$

From Theorems 2.1.16 and 2.1.17,

$$h(Q_{i,N}) = Z_{11}h(0) + Z_{21}h(\gamma_i) \qquad (4.2.35)$$

Comparing equations (4.2.34) and (4.2.35), we see that

$$Z_{11} = \frac{\gamma_i I - Q_{i,N}}{\gamma_i} \quad \text{and} \quad Z_{21} = \frac{Q_{i,N}}{\gamma_i}$$

QED

<u>Theorem 4.2.4</u> We may choose $A_{i,N}$ in equation (4.2.18) to be

$$A_{i,N} = \frac{\gamma_i I - Q_{i,N}}{\gamma_i} \qquad (4.2.36)$$

where $\gamma_i = tr(Q_{i,N})$.

<u>Proof</u> It suffices to show that $Q_{i,N}A_{i,N} = 0$. By equations (4.2.28) and (4.2.36) we have $A_{i,N} = Z_{11}$. From Theorem 2.1.17, $Z_{21}Z_{11} = 0$. Therefore, by equations (4.2.28) and (4.2.29)

$$\frac{Q_{i,N}}{\gamma_i}\left[\frac{\gamma_i I - Q_{i,N}}{\gamma_i}\right] = 0 \qquad (4.2.37)$$

which implies that

$$Q_{i,N}A_{i,N} = 0 \qquad (4.2.38)$$

QED

Theorem 4.2.5 The pseudoinverse $Q^+_{i,N}$ of $Q_{i,N}$ is given by

$$Q^+_{i,N} = \frac{Q_{i,N}}{\gamma_i^2} \qquad (4.2.39)$$

where $\gamma_i = tr(Q_{i,N})$.

Proof We need to show that the four identities in Theorem 2.2.3 are satisfied. From equation (4.2.37) we have

$$\gamma_i Q_{i,N} = Q^2_{i,N} \qquad (4.2.40)$$

which implies that

$$Q^3_{i,N} = \gamma_i^2 Q_{i,N} \qquad (4.2.41)$$

Then by equations (4.2.39) and (4.2.41),

$$Q_{i,N}Q^+_{i,N}Q_{i,N} = \frac{Q^3_{i,N}}{\gamma_i^2} = Q_{i,N}$$

and the first identity is satisfied. Similarly,

$$Q^+_{i,N}Q_{i,N}Q^+_{i,N} = \frac{Q^3_{i,N}}{\gamma_i^4} = \frac{Q_{i,N}}{\gamma_i^2} = Q^+_{i,N}$$

and the second identity is satisfied. Since $Q_{i,N}$ is symmetric, $Q^+_{i,N}$ will also by symmetric by the assumption on the form of $Q^+_{i,N}$. Therefore,

$$(Q_{i,N}Q^{+}_{i,N})^{T} = (Q_{i,N}\frac{Q_{i,N}}{\gamma_i^2})^{T} = Q^{T}_{i,N}\frac{Q^{T}_{i,N}}{\gamma_i^2} = Q_{i,N}Q^{+}_{i,N}$$

and the third identity is satisfied. Also,

$$(Q^{+}_{i,N}Q_{i,N})^{T} = (\frac{Q_{i,N}}{\gamma_i^2}Q_{i,N})^{T} = \frac{Q^{T}_{i,N}}{\gamma_i^2}Q^{T}_{i,N} = Q^{+}_{i,N}Q_{i,N}$$

and the fourth identity is verified.

QED

Using the expression for $C_{i,N}$ given by equation (4.2.17), we obtain

$$Q_{i,N}C_{i,N} - d_{i,N} = (Q_{i,N}Q^{+}_{i,N} - I)d_{i,N} + \alpha Q_{i,N}A_{i,N}d_{i,N}$$

By equation (4.2.38) this reduces to

$$Q_{i,N}C_{i,N} - d_{i,N} = (Q_{i,N}Q^{+}_{i,N} - I)d_{i,N} \qquad (4.2.42)$$

From equation (4.2.39)

$$Q_{i,N}Q^{+}_{i,N} = \frac{Q^{2}_{i,N}}{\gamma_i^2}$$

$$= \frac{Q_{i,N}}{\gamma_i} \text{ by equation (4.2.40)}$$

$$= -A_{i,N} + I \text{ by Theorem 4.2.4} \qquad (4.2.43)$$

Let us substitute equation (4.2.43) into equation (4.2.42).

$$Q_{i,N}C_{i,N} - d_{i,N} = - A_{i,N}d_{i,N} \qquad (4.2.44)$$

Let

$$h^2 = (-A_{i,N}d_{i,N})^T(-A_{i,N}d_{i,N})$$

$$= d_{i,N}^T A_{i,N} A_{i,N} d_{i,N}$$

Recalling from equations (4.2.28) and (4.2.36) that $Z_{11} = A_{i,N}$ and using Theorem 2.1.17, we have

$$h^2 = d_{i,N}^T A_{i,N} d_{i,N} \qquad (4.2.45)$$

<u>Theorem 4.2.6</u> h=0 in equation (4.2.45), and equation (4.2.16) reduces to

$$g(Y_i) = \gamma_i y_{0,i}^2 - M_i^2 \qquad (4.2.46)$$

where

$$Y_i = (y_{0,i}, y_{1,i}, \ldots, y_{Nm-1,i})^T; \; \gamma_i = tr(Q_{i,N}) \qquad (4.2.47)$$

<u>Proof</u> From equations (4.2.8) and (4.2.2), $Q_{i,N} = \bar{F}_N^T \psi_{i,N} \bar{F}_N = \bar{F}_N^T (c_i^N)^T c_i^N \bar{F}_N = BE$ where $B = \bar{F}_N^T (c_i^N)^T$ and $E = c_i^N \bar{F}_N$. $Q_{i,N}$ is of rank 1. Thus $B = E^T \neq 0$. Then B is of order Nmx1 and of rank 1 while E is 1xNm and of rank 1. Thus BE is a rank factorization of $Q_{i,N}$. From Theorem 2.2.1 we have $EQ_{i,N}^+ B = I$ or

$$c_i^N \bar{F}_N Q_{i,N}^+ \bar{F}_N^T (c_i^N)^T = I$$

Premultiplying this equation by $(c_i^N)^T$ and postmultiplying it by c_i^N, we get

$$\psi_{i,N} \bar{F}_N Q_{i,N}^+ \bar{F}_N^T \psi_{i,N} = \psi_{i,N} \qquad (4.2.48)$$

From equations (4.2.45) and (4.2.9)

$$h^2 = d_{i,N}^T A_{i,N} d_{i,N}$$

$$= x_0^T \psi_{i,N} \bar{F}_N A_{i,N} \bar{F}_N^T \psi_{i,N} x_0 \qquad (4.2.49)$$

Substituting equation (4.2.48) into (4.2.49) yields

$$h^2 = x_0^T \psi_{i,N} \bar{F}_N Q_{i,N}^+ \bar{F}_N^T \psi_{i,N} \bar{F}_N A_{i,N} \bar{F}_N^T \psi_{i,N} x_0$$

$$= x_0^T \psi_{i,N} \bar{F}_N Q_{i,N}^+ Q_{i,N} A_{i,N} \bar{F}_N^T \psi_{i,N} x_0 \text{ by equation (4.2.8)}$$

From equation (4.2.38), $Q_{i,N} A_{i,N} = 0$ in the above equation. Thus $h = 0$, and the first part of the theorem is proven.

Since $h=0$, then by equation (4.2.45)

$$A_{i,N} d_{i,N} = 0 \qquad (4.2.50)$$

Substituting equation (4.2.50) into (4.2.44) gives $Q_{i,N} C_{i,N} - d_{i,N} = 0$ which implies that $L_{i,N} = 0$ by equation (4.2.14). This in turn implies (by equation (4.2.16)) that

$$g(Y_i) = Y_i^T P_{i,N}^T Q_{i,N} P_{i,N} Y_i + g_{i,N}$$

$$= Y_i^T \Lambda_{i,N} Y_i + g_{i,N} \qquad (4.2.51)$$

where $P_{i,N}$ is chosen such that $P_{i,N}^T P_{i,N} = I$ and $\Lambda_{i,N} = P_{i,N}^T Q_{i,N} P_{i,N}$ is a diagonal matrix with the eigenvalues of $Q_{i,N}$ along the diagonal. From equation (4.2.15), $g_{i,N} = C_{i,N}^T Q_{i,N} C_{i,N} - 2d_{i,N}^T C_{i,N} + e_{i,N}$. By equations (4.2.17) and (4.2.50) this becomes

$$g_{i,N} = (Q^{+}_{i,N}d_{i,N})^{T}Q_{i,N}Q^{+}_{i,N}d_{i,N} - 2d^{T}_{i,N}Q^{+}_{i,N}d_{i,N} + e_{i,N}$$

$$= d^{T}_{i,N}Q^{+}_{i,N}Q_{i,N}Q^{+}_{i,N}d_{i,N} - 2d^{T}_{i,N}Q^{+}_{i,N}d_{i,N} + e_{i,N}$$

$$= d^{T}_{i,N}Q^{+}_{i,N}d_{i,N} - 2d^{T}_{i,N}Q^{+}_{i,N}d_{i,N} + e_{i,N}$$ by Theorem 2.2.3

$$= -d^{T}_{i,N}Q^{+}_{i,N}d_{i,N} + e_{i,N}$$

$$= -x^{T}_{0}\psi_{i,N}\bar{F}_{N}Q^{+}_{i,N}\bar{F}^{T}_{N}\psi_{i,N}x_{0} + e_{i,N}$$ by equation (4.2.9)

$$= -x^{T}_{0}\psi_{i,N}x_{0} + e_{i,N}$$ by equation (4.2.48)

$$= -x^{T}_{0}\psi_{i,N}x_{0} + x^{T}_{0}\psi_{i,N}x_{0} - M^{2}_{i}$$ by equation (4.2.10)

$$= - M^{2}_{i} \qquad (4.2.52)$$

Let us substitute equation (4.2.52) into (4.2.51).

$$g(Y_i) = Y^{T}_{i}\ \Lambda_{i,N}Y_i - M^{2}_{i} \qquad (4.2.53)$$

Since $Q_{i,N}$ is symmetric and of rank 1 by Theorem 4.2.1, $\Lambda_{i,N}$ will have only one nonzero element which is the nonzero eigenvalue of $Q_{i,N}$. The columns of $P_{i,N}$ can be ordered such that this nonzero element appears as the first element on the diagonal of $\Lambda_{i,N}$. Thus equation (4.2.53) reduces to $g(Y_i) = \gamma_i y^{2}_{0,i} - M^{2}_{i}$ where $\gamma_i = tr(Q_{i,N})$ and $Y_i = (y_{0,i}, y_{1,i}, \ldots, y_{Nm-1,i})^{T}$.

QED

From equations (4.2.50), (4.2.17) and (4.2.12) we see that

$$U = P_{i,N}Y_i + Q^{+}_{i,N}d_{i,N}$$

$$= P_{i,N}Y_i + \frac{Q_{i,N}d_{i,N}}{\gamma_i^2} \text{ by Theorem 4.2.5} \quad (4.2.54)$$

Equation (4.2.54) represents a rotation and shift between the Y_i and U coordinates. All angles and distances are preserved by this transformation. Let $g(Y_i) = 0$. Then

$$y_{0,i}^2 = \frac{M_i^2}{\gamma_i} \quad (4.2.55)$$

For the case when N=2 and m=1, equation (4.2.55) represents two parallel lines as shown in Figure 4.2.1a. In the higher dimensional case the lines are replaced by hyperplanes. If $g(Y_i) \leq 0$, equation (4.2.46) gives

$$\frac{-M_i}{\sqrt{\gamma_i}} \leq y_{0,i} \leq \frac{M_i}{\sqrt{\gamma_i}} \quad (4.2.56)$$

When N=2 and m=1, the inequality in (4.2.56) represents the region between the two lines $y_{0,i} = \pm M_i/\sqrt{\gamma_i}$. The transformation given by equation (4.2.54) will rotate and shift the hyperplanes. Thus the expression for U (the sequence of optimal controls) corresponding to inequality (4.2.56) will lie in the shaded region shown in Figure 4.2.1b. The problem to be considered next is the general expression for the hyperplanes in the U coordinates given the hyperplanes in the Y_i system.

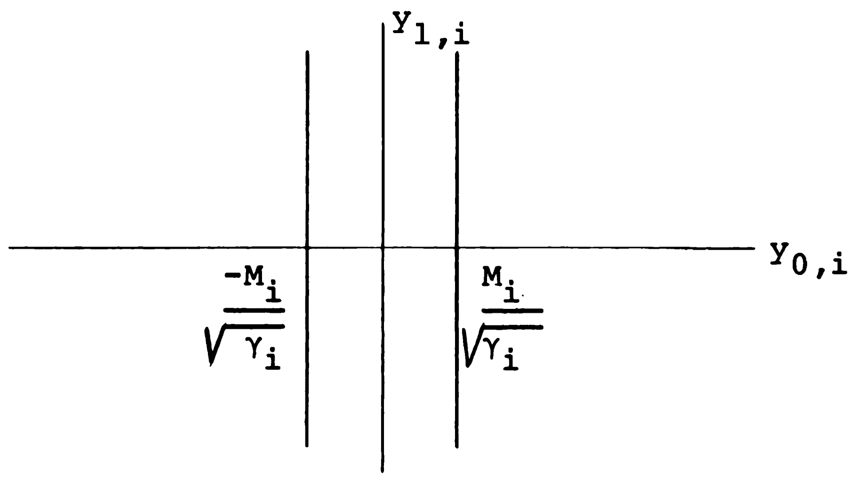


Fig. 4.2.1a. Constraints on the control sequence in the $y_{0,i}$-$y_{1,i}$ plane

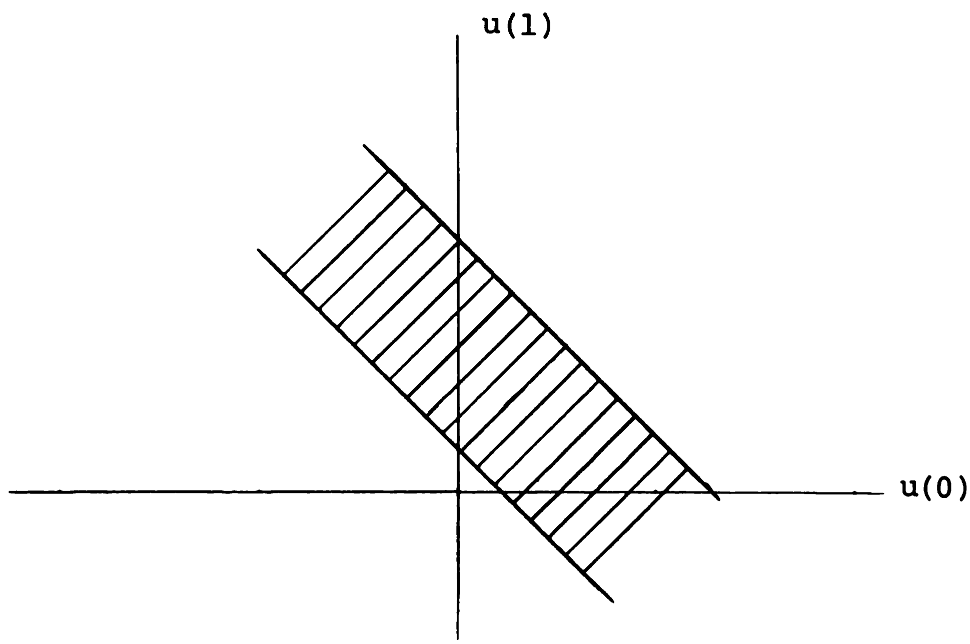


Fig. 4.2.1b. Constraints on the control sequence in the u(0) - u(1) plane

From equation (4.2.54) we see that U is related to Y_i by the $P_{i,N}$ matrix. We have chosen $P_{i,N}$ such that

$$\Lambda_{i,N} = P_{i,N}^T Q_{i,N} P_{i,N} \qquad (4.2.57)$$

$$P_{i,N}^T P_{i,N} = I \qquad (4.2.58)$$

Because $Q_{i,N}$ is of rank 1, $P_{i,N}$ can be found easily as shown by the following theorem.

__Theorem 4.2.7__ Let $P_{i,N} = (p_1^i, p_2^i, \ldots, p_{Nm}^i)$ where the p_j^i are the columns of the $P_{i,N}$ matrix. Then we can choose the p_j^i as follows.

$$p_1^i = \frac{\pm q^i}{\|q^i\|^{1/2}} \qquad p_j^i = \frac{\pm z_j^i}{\|z_j^i\|^{1/2}} \quad j=2,3,\ldots,Nm$$

q^i is any nonzero column of $Q_{i,N}$. z_j^i, $j=2,3,\ldots,Nm$ are any Nm-1 nonzero columns of $Q_{i,N} - \gamma_i I$ and $\gamma_i = tr(Q_{i,N})$ is the positive eigenvalue of $Q_{i,N}$. (The norm is given by $\|w\|^2 = w^T w$.)

__Proof__ From Theorem 2.1.18 we know that Nm-1 columns of the $P_{i,N}$ matrix can be found from the nonzero columns of

$$\frac{d^{Nm-2}}{d\lambda^{Nm-2}} \{adj(\lambda I - Q_{i,N})\}\Big|_{\lambda=0} \qquad (4.2.59)$$

From Theorem 2.1.10 we have

$$adj(\lambda I - Q_{i,N}) = I\lambda^{Nm-1} + G_1\lambda^{Nm-2} + \ldots + G_{Nm-2}\lambda + G_{Nm-1}$$

By Theorem 4.2.2 this becomes

$$adj(\lambda I - Q_{i,N}) = I\lambda^{Nm-1} + G_1\lambda^{Nm-2} \qquad (4.2.60)$$

Taking the derivative as indicated in (4.2.59), we have

$$\frac{d^{Nm-2}}{d\lambda^{Nm-2}}\{\mathrm{adj}(\lambda I - Q_{i,N})\}\Big|_{\lambda=0} = (Nm-1)!\lambda I + (Nm-2)!G_1\Big|_{\lambda=0}$$
$$= (Nm-2)!G_1 \qquad (4.2.61)$$

By Theorems 2.1.10 and 4.2.2

$$G_1 = Q_{i,N} + s_1 I$$
$$= Q_{i,N} - \gamma_i I \qquad (4.2.62)$$

Let z_j^i, $j=2,3,\ldots,Nm$ be any Nm-1 nonzero columns of $Q_{i,N} - \gamma_i I$. Then by the above, these columns are independent and can be used to determine Nm-1 columns of the $P_{i,N}$ matrix. By equation (4.2.58) we require the columns to be normalized. Thus we can set

$$p_j^i = \frac{\pm z_j^i}{\|z_j^i\|^{1/2}} \qquad j = 2,3,\ldots,Nm$$

and we have Nm-1 columns of $P_{i,N}$ that satisfy equation (4.2.58). The remaining column of $P_{i,N}$ can be found from any nonzero column of $\mathrm{adj}(\lambda I - Q_{i,N})\big|_{\lambda=\gamma_i}$ where γ_i is the nonzero eigenvalue of $Q_{i,N}$. By Theorem 4.2.2 we know this to be $\gamma_i = \mathrm{tr}(Q_{i,N})$. From equation (4.2.60) we then have

$$\mathrm{adj}(\lambda I - Q_{i,N})\Big|_{\lambda=\gamma_i} = I\gamma_i^{Nm-1} + G_1\gamma_i^{Nm-2}$$
$$= \gamma_i^{Nm-2}(\gamma_i I + G_1)$$
$$= \gamma_i^{Nm-2}(\gamma_i I + Q_{i,N} - \gamma_i I) \text{ by equation}$$

(4.2.62)

$$=\gamma_i^{Nm-2} Q_{i,N}$$

An eigenvector is then proportional to any nonzero column of $Q_{i,N}$. Let q^i be one of these nonzero columns. To satisfy equation (4.2.58) we normalize q^i. Thus we can choose

$$p_1^i = \frac{\pm q^i}{\|q^i\|^{1/2}}$$

QED

Let us now find the inequalities in the U coordinates which correspond to the inequalities in (4.2.56).

Theorem 4.2.8 If a sequence of time-optimal controls exists, it is a solution of the following set of inequalities.

$$\frac{-M_i}{\sqrt{\gamma_i}} + r_i \leq p_{1,i}u_1(0) + p_{2,i}u_2(0) + \dots + p_{m,i}u_m(0)$$

$$+ p_{m+1,i}u_1(T) + p_{m+2,i}u_2(T) + \dots + p_{2m,i}u_m(T)$$

$$+ \dots + p_{Nm-m+1,i}u_1[(N-1)T] + \dots$$

$$+ p_{Nm,i}u_m[(N-1)T] \leq \frac{M_i}{\sqrt{\gamma_i}} + r_i \qquad i=1,2,\dots,n$$

where $r_i = (p_1^i)^T \dfrac{Q_{i,N}}{\gamma_i^2} d_{i,N}$ and $p_1^i = (p_{1,i}, p_{2,i}, \dots, p_{Nm,i})^T$ is the first column of the $P_{i,N}$ matrix and is given by

$$p_1^i = \frac{\pm q^i}{\|q^i\|^{1/2}}$$

where q^i is any nonzero column of $Q_{i,N}$. Moreover, the solution does not depend on which nonzero column of $Q_{i,N}$

we choose or whether we choose the plus or minus sign in the expression for p_1^i. If Nm=1, we get $\gamma_i = Q_{i,N}$ and $p_1^i = \pm 1$. It is assumed that $Q_{i,N} \neq 0$.

<u>Proof</u> Let $P_{i,N} = (p_1^i, p_2^i, \ldots, p_{Nm}^i)$. Then

$$P_{i,N} = \begin{bmatrix} (p_1^i)^T \\ (p_2^i)^T \\ \vdots \\ (p_{Nm}^i)^T \end{bmatrix} \qquad (4.2.63)$$

From equation (4.2.54), $U = P_{i,N}Y_i + \frac{Q_{i,N}}{\gamma_i^2} d_{i,N}$. Thus $Y_i = P_{i,N}^{-1}(U - \frac{Q_{i,N}}{\gamma_i^2} d_{i,N})$. Using equation (4.2.58), this becomes

$$\begin{bmatrix} y_{0,i} \\ y_{1,i} \\ \vdots \\ y_{Nm-1,i} \end{bmatrix} = P_{i,N}^T(U - \frac{Q_{i,N}}{\gamma_i^2} d_{i,N}) \qquad (4.2.64)$$

From (4.2.7) and (4.2.11), we require that $f(U) \leq 0$. This in turn leads to $g(Y_i) \leq 0$ where $g(Y_i)$ is given by equation (4.2.13). From Theorem 4.2.6 we then must have

$$\frac{-M_i}{\sqrt{\gamma_i}} \leq y_{0,i} \leq \frac{M_i}{\sqrt{\gamma_i}} \qquad (4.2.65)$$

From (4.2.64) and (4.2.65), we then have

$$-\frac{M_i}{\sqrt{\gamma_i}} \leq (p_1^i)^T(U - \frac{Q_{i,N}}{\gamma_i^2} d_{i,N}) \leq \frac{M_i}{\sqrt{\gamma_i}}$$

where p_1^i is the first column of $P_{i,N}$. Therefore,

$$-\frac{M_i}{\sqrt{\gamma_i}} + (p_1^i)^T \frac{Q_{i,N}}{\gamma_i^2} d_{i,N} \leq (p_1^i)^T U \leq \frac{M_i}{\sqrt{\gamma_i}} + (p_1^i)^T \frac{Q_{i,N}}{\gamma_i^2} d_{i,N} \quad (4.2.66)$$

From Theorem 4.2.7 we know that $p_1^i = \frac{\pm q^i}{||q^i||^{1/2}}$ where q_1^i is any nonzero column of $Q_{i,N}$. The value of p_1^i does not depend upon which nonzero column we choose. This follows from the fact that the $Q_{i,N}$ matrix is of rank 1 so that the nonzero columns of $Q_{i,N}$ are related by the equation $q^j = \alpha q^i$ where q^j and q^i are nonzero columns of $Q_{i,N}$ and α is a nonzero scalar. Then

$$p_1^j = \frac{\pm q^j}{||q^j||^{1/2}} = \frac{\pm \alpha q^i}{||\alpha q^i||^{1/2}} = \frac{\pm q^i}{||q^i||^{1/2}} \text{sgm}(\alpha) = p_1^i \text{sgm}(\alpha) \quad (4.2.67)$$

Thus p_1^j and p_1^i can differ at most only in sign. To show that Theorem 4.1.8 does not depend on which sign we choose for p_1^j in equation (4.2.67), let us substitute $(-p_1^i)$ for p_1^i in equality (4.2.66). This gives

$$-\frac{M_i}{\sqrt{\gamma_i}} + (-p_1^i)^T \frac{Q_{i,N}}{\gamma_i^2} d_{i,N} \leq (-p_1^i)^T U \leq \frac{M_i}{\sqrt{\gamma_i}} + (-p_1^i)^T \frac{Q_{i,N}}{\gamma_i^2} d_{i,N}$$

or after multiplying through by (-1), we obtain

$$-\frac{M_i}{\sqrt{\gamma_i}} + (p_1^i)^T \frac{Q_{i,N}}{\gamma_i^2} d_{i,N} \leq (p_1^i)^T U \leq \frac{M_i}{\sqrt{\gamma_i}} + (p_1^i)^T \frac{Q_{i,N}}{\gamma_i^2} d_{i,N}$$

This inequality is the same as inequality (4.2.66), and therefore, the result is independent of whether we choose the plus or minus sign in equation (4.2.67).

For Nm=1, $Q^{+}_{i,N} = Q^{-1}_{i,N}$ since the pseudoinverse equals the inverse when the latter exists. If we choose $p^{i}_{1} = \pm 1$ and set $u_1(0) = \pm y_{0,i} + d_{i,1}/Q_{i,1}$, then the inequality $Q^2_{i,1}u_1(0) - 2d_{i,1}u_1(0) + e_{i,1} \leq 0$ transforms into $y^2_{0,i} \leq M^2_i/Q_{i,1}$. This in turn implies that

$$-\frac{M_i}{\sqrt{Q_{i,1}}} + \frac{d_{i,1}}{Q_{i,1}} \leq u_1(0) \leq \frac{M_i}{\sqrt{Q_{i,1}}} + \frac{d_{i,1}}{Q_{i,1}}$$

With $\gamma_i = tr(Q_{i,N}) = Q_{i,N}$, the same result as given in the statement of the theorem follows.

QED

Theorem 4.2.8 provides a means of determining a sequence of time-optimal controls. Before proceeding further it should be noted, however, that the $P_{i,N}$ matrix which diagonalizes the $Q_{i,N}$ matrix is not unique. To show this, let

$$\Lambda = P^T_{i,N}Q_{i,N}P_{i,N}$$

where $P^T_{i,N}P_{i,N} = I$ and Λ is a diagonal matrix with the eigenvalues of $Q_{i,N}$ along the diagonal. Let R be any matrix such that $R^TR = I$ and $R\Lambda = \Lambda R$. The matrix R always exists since we can choose R to be a diagonal matrix. Then $P_{i,N}R$ will also diagonalize $Q_{i,N}$ since

$$Q_{i,N} = P_{i,N} \Lambda P_{i,N}^T = P_{i,N} \Lambda R R^T P_{i,N}^T = P_{i,N} R \Lambda R^T P_{i,N}^T$$

$$= (P_{i,N} R) \Lambda (P_{i,N} R)^T$$

Because $P_{i,N}$ is not unique, it may seem that the solution to the inequalities given in Theorem 4.2.8 depends on how we choose $P_{i,N}$. This is not the case since we are only interested in the first column of $P_{i,N}$ which corresponds to the single nonzero eigenvalue γ_i of $P_{i,N}$. This column is equal to a normalized eigenvector corresponding to γ_i. Thus it can assume only two values, the one differing from the other only in sign. That is, the only possible values of p_1^i are given in Theorem 4.2.8. By this same theorem we know that we get the same set of inequalities regardless of whether we choose the plus or minus sign. The conclusion, therefore, is that the solution to the time-optimal control problem does not depend on how we choose the non-unique $P_{i,N}$ matrix; Theorem 4.2.8 is the same for any choice of $P_{i,N}$ which satisfies equations (4.2.57)-(4.2.58).

Theorem 4.2.8 is the main result when it is assumed that $Q_{i,N} \neq 0$. The following theorem gives the corresponding result when $Q_{i,N} = 0$.

Theorem 4.2.9 If $Q_{i,N} = 0$, then the following is true.

a) If a solution exists, then a time-optimal sequence of controls is a solution of the following set of inequalities.

$$\alpha_{1,i} u_1(0) + \alpha_{2,i} u_2(0) + \ldots + \alpha_{m,i} u_m(0) + \alpha_{m+1,i} u_1(T)$$

$$+\alpha_{m+2,i}u_2(T)+\dots+\alpha_{2m,i}u_m(T)+\dots$$

$$+\alpha_{Nm-m+1,i}u_1[(N-1)T]+\dots+\alpha_{Nm}u_m[(N-1)T]$$

$$\leq M_i^2 - x_0^T\Psi_{i,N}x_0 \quad i=1,2,\dots,n \tag{4.2.72}$$

where $(\alpha_{1,i},\alpha_{2,i},\dots,\alpha_{Nm,i}) = -2x_0^T\Psi_{i,N}\bar{F}_N$ (1xMm vector)

(4.2.73)

b) If $N \geq n/m$ and the nxNm matrix $\bar{F}_N$ is of rank n, then we have the following two cases.

1) If $x_0^T\Psi_{i,N}x_0 \leq M_i^2$, i=1,2,...,n, then the sequence of time-optimal control is arbitrary. In particular, the sequence U = 0 is energy and fuel optimal.

2) If $x_0^T\psi_{i,N}x_0 > M_i^2$ for some i, then no solution exists for this value of N.

<u>Proof</u> From inequality (4.2.7), we are looking for a sequence of controls U such that $U^TQ_{i,N}U - 2d_{i,N}^TU + e_{i,N} \leq 0$ for i=1,2,...,n. With $Q_{i,N} = 0$, this reduces to

$$-2_{i,N}^TU + e_{i,N} \leq 0$$

Using equations (4.2.9) and (4.2.10), this inequality becomes

$$-2x_0^T\psi_{i,N}\bar{F}_NU + x_0^T\psi_{i,N}x_0 - M_i^2 \leq 0 \tag{4.2.74}$$

From (4.2.73) and (4.2.74), inequality (4.2.72) follows.

b) For $N \geq n/m$, inequality (4.2.74) still holds. From equation (4.2.8) $Q_{i,N}= \bar{F}_N^T\psi_{i,N}\bar{F}_N = 0$. By assumption the nxNm matrix $\bar{F}_N$ is of rank n. From Theorem 2.1.20, it follows that $\bar{F}_N$ has right inverse $\bar{F}_N^R$. Thus $\bar{F}_N \psi_{i,N} \bar{F}_N \bar{F}_N^R = 0$ or $\bar{F}_N^T \psi_{i,N} = 0$ which implies that

$$\psi_{i,N} \bar{F}_N = 0 \qquad (4.2.75)$$

Substituting equation (4.2.75) into (4.2.74) yields

$$x_0^T \psi_{i,N} x_0 \leq M_i^2 \qquad (4.2.76)$$

If x_0, $\psi_{i,N}$ and M_i, $i=1,2,\ldots,n$ are such that inequality (4.2.76) is satisfied, then U is arbitrary and U = 0 is the minimum fuel and minimum energy sequence of controls. On the other hand, if for some i we have $x_0^T \psi_{i,N} x_0 > M_i^2$, then no solution exists because inequality (4.2.76) must be satisfied.

QED

Theorems 4.2.8 and 4.2.9 are the main results. They provide a means of determining a solution if it exists. Thus far no systematic method of determining the fewest number of samples N to reach the target has been given. The constraint on the amplitude of the controllers given by equation (4.1.4) has not been taken into consideration. However, both of these problems shall be overcome by defining new variables and using the linear programming method of solution.

4.3 Solution of Control Problem Using Linear Programming

In the previous section the optimal control problem with no constraints on the amplitude of the controller was reduced to finding a solution to a set of linear inequalities. In this section the control problem is formulated as a linear programming problem. Theorems from the Simplex Method [HAD1] of linear programming are used to determine if a solution exists for a particular value of N, and whether the solution is unique. If the solution is not unique, a trajectory is chosen that minimizes the total fuel to reach the target.

From Theorem 4.2.8 we have that the solution to the optimal control problem is a solution to the following set of inequalities.

$$\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,1} u_i(kT) \leq \frac{M_1}{\sqrt{\gamma_1}} + r_1$$

$$\vdots$$

$$\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,n} u_i(kT) \leq \frac{M_n}{\sqrt{\gamma_n}} + r_n \qquad (4.3.1)$$

$$\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,1} u_i(kT) \geq \frac{-M_1}{\sqrt{\gamma_1}} + r_1$$

$$\vdots$$

$$\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,n} u_i(kT) \geq \frac{-M_n}{\sqrt{\gamma_n}} + r_n$$

From equation (4.1.4), we also require that

$$|u_i(kT)| \leq G_{i,k} \quad \text{for } i=1,2,\ldots,m \quad k=0,1,\ldots,N-1 \qquad (4.3.2)$$

The constraint given in equation (4.3.2) can be taken into consideration in two ways [TOR1]. The first method consists of letting

$$u_i(kT) = u_{i,p}(kT) - u_{i,q}(kT) \tag{4.3.3}$$

If we substitute equation (4.3.3) into (4.3.1), we obtain the following set of inequalities.

$$\begin{aligned}
&\sum_{k=0}^{N-1} \sum_{i-1}^{m} p_{km+i,1}(u_{i,p}(kT) - u_{i,q}(kT)) \leq \frac{M_1}{\sqrt{\gamma_1}} + r_1 \\
&\qquad\vdots \\
&\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,n}(u_{i,p}(kT) - u_{i,q}(kT)) \leq \frac{M_n}{\sqrt{\gamma_n}} + r_n \\
&\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,1}(u_{i,p}(kT) - u_{i,q}(kT)) \geq \frac{-M_1}{\sqrt{\gamma_1}} + r_1 \\
&\qquad\vdots \\
&\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,n}(u_{i,p}(kT) - u_{i,q}(kT)) \geq \frac{-M_n}{\sqrt{\gamma_n}} + r_n
\end{aligned} \tag{4.3.4}$$

The constraint given in (4.3.2) is equivalent to

$$0 \leq u_{i,p}(kT) \leq G_{i,k} \quad i=1,2,\ldots,n \quad k=0,1,\ldots,N-1$$

$$0 \leq u_{i,q}(kT) \leq G_{i,k}$$

Once N is known and a linear objective function is specified, the inequalities given in (4.3.4)-(4.3.5) represent a standard linear programming problem. This formulation has the disadvantage that it doubles the number of unknowns. For purposes of determining the value of N, we make the following substitution. Let

$$u_i(kT) = u_{i,k} - G_{i,k} \tag{4.3.6}$$

The inequalities in (4.3.1) then become the following.

$$\begin{aligned}
&\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,1} u_{i,k} \leq b_1 \\
&\quad\vdots \\
&\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,n} u_{i,k} \leq b_n \\
&\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,1} u_{i,k} \geq \beta_1 \\
&\quad\vdots \\
&\sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{,m+i,n} u_{i,k} \geq \beta_n
\end{aligned} \tag{4.3.7}$$

where

$$b_j = \frac{M_j}{\sqrt{\gamma_j}} + r_j + \sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,j} G_{i,k}$$

$$\beta_j = \frac{-M_j}{\sqrt{\gamma_i}} + r_j + \sum_{k=0}^{N-1} \sum_{i=1}^{m} p_{km+i,j} G_{i,k}$$

and r_j is given in Theorem 4.2.8.

The inequalities given in (4.3.2) transform into the following:

$$0 \leq u_{i,k} \leq 2G_{i,k} \tag{4.3.8}$$

This method does not increase the number of unknowns to be determined. In order to solve the problem using linear programming, the next step is to convert the inequalities to equalities by adding slack or surplus variables q_{im+i}.

If we add Nm slack variables to the inequalities in (4.3.8), we have

$$\begin{aligned} u_{1,k} + q_{km+1} &= 2G_{1,k} \\ &\vdots \\ u_{m,k} + q_{km+1} &= 2G_{m,k} \end{aligned} \qquad k=0,1,\ldots,N-1 \qquad (4.3.9)$$

where $q_{km+i} \geq 0$ for $i=1,2,\ldots,m$

In general we do not know which of the b_i and β_i in inequalities (4.3.7) are positive. If a b_i or β_i is negative, we can multiply both sides of the inequality by (-1) to put it in the proper form. We then add the necessary slack and surplus variables to convert the inequalities into equalities. To form an initial basic feasible solution we add the necessary artificial variables $\bar{z}_i$, say q of them, to form an identity submatrix. The value of N is then found by using Phase I of linear programming [HAD1]. This consists of maximizing the objective function

$$J_a = -\sum_{i=1}^{q} \bar{z}_i \qquad (4.3.10)$$

By using the simplex technique and Theorems 2.4.1 and 2.4.2, we can tell whether a solution to the linear programming problem exists for that value of N. If not, we increase N by one and try again. When a solution is found, the Simplex Method also indicates whether the solution is unique. Assuming the solution is not unique, we return to the formulation given by inequalities (4.3.4) and (4.3.5) to determine the time-optimal sequence that requires minimum fuel. The objective function is chosen to be

$$J_b = \sum_{k-0}^{N-1} \sum_{i=1}^{m} (u_{i,p}(kT) + u_{i,q}(kT)) \qquad (4.3.11)$$

It is claimed that if $u_{i,p}(kT)$ and $u_{i,q}(kT)$ minimize J_b, then they also minimize the fuel, that is,

$$J = \sum_{k=0}^{N-1} \sum_{i=1}^{m} |u_i(kT)|$$

It suffices to show that if $\hat{u}_{i,p}(kT)$, $\hat{u}_{i,q}(kT)$ and $\hat{u}_i(kT)$ minimize J_b, then

$$\hat{u}_{i,p}(kT) = \begin{cases} \hat{u}_i(kT) & \text{if } \hat{u}_i(kT) > 0 \\ 0 & \text{if } \hat{u}_i(kT) \leq 0 \end{cases}$$

$$\hat{u}_{i,q}(kT) = \begin{cases} 0 & \text{if } \hat{u}_k(kT) \geq 0 \\ -\hat{u}_i(kT) & \text{if } \hat{u}_i(kT) < 0 \end{cases}$$

where $\hat{u}_{i,p}(kT)$, $\hat{u}_{i,q}(kT)$, and $\hat{u}_i(kT)$ are the values of $u_{ip}(kT)$, $u_{i,q}(kT)$ and $u_i(kT)$ that minimize J_b. To show this, it is sufficient to show that both $\hat{u}_{i,p}(kT)$ and $\hat{u}_{i,q}(kT)$ cannot simultaneously be positive. The proof is by contradiction. Assume an optimal value of J_b has been found. Suppose for some k, $\hat{u}_{i,p}(kT) > 0$ and $\hat{u}_{i,q}(kT) > 0$. First assume that $\hat{u}_{i,p}(kT) > \hat{u}_{i,q}(kT)$. Define two new variables $\overline{u}_{i,p}(kT)$ and $\overline{u}_{i,q}(kT)$ by

$$\overline{u}_{i,p}(kT) = \hat{u}_{i,p}(kT) - \hat{u}_{i,q}(kT) > 0 \text{ and } \overline{u}_{i,q}(kT) = 0$$

Then

$$\bar{u}_{i,p}(kT) - \bar{u}_{i,q}(kT) = \hat{u}_{i,p}(kT) - \hat{u}_{i,q}(kT) = \hat{u}_i(kT)$$

Then if we substitute $\bar{u}_{i,p}(kT)$ for $\hat{u}_{i,p}(kT)$ and $\bar{u}_{i,q}(kT)$ for $\hat{u}_{i,q}(kT)$ into the optimal solution, the constraints are still satisfied. However,

$$\bar{u}_{i,p}(kT) + \bar{u}_{i,q}(kT) = \hat{u}_{i,p}(kT) - \hat{u}_{i,q}(kT) < \hat{u}_{i,p}(kT) + \hat{u}_{i,q}(kT)$$

since by assumption $\hat{u}_{i,q}(kT) > 0$. This is a contradiction of the assumption that $\hat{u}_{i,p}(kT)$ and $\hat{u}_{i,q}(kT)$ are optimal. The case where $\hat{u}_{i,p}(kT) < \hat{u}_{i,q}(kT)$ leads to the same result. It then follows that the minimizing J_b minimizes the fuel J.

The discussion given above was for the case when $Q_{i,N} \neq 0$. However, by Theorem 4.2.9 the solution to the optimal control problem for $Q_{i,N} = 0$ is again a solution to a set of linear inequalities. Thus the same arguments can be repeated for this case also.

If we assume that the constraint on the amplitude of the controller as given by (4.1.4) does not apply, the only modification is to not include equations (4.3.9) in the set of linear equations to be solved.

<u>Example 4.3.1</u> Given the discrete-time system

$$x(k+1) = Cx(k) + Du(k) \tag{4.3.13}$$

where

$$C = \begin{bmatrix} 1 & 1-e^{-1} \\ 0 & c^{-1} \end{bmatrix} \tag{4.3.14}$$

$$D = (e^{-1} \quad 1-e^{-1})^T \tag{4.3.15}$$

$$x(0) = (10 \quad -12)^T \tag{4.3.16}$$

It is assumed that

$$|u(kT)| \leq 0.5 \quad k=0,1,\ldots,N-1 \tag{4.3.17}$$

We wish to find the minimum number of samples N and a corresponding sequence of controls that drive the system to the target described by

$$-1 \leq x_j \ (kT) \leq 1 \qquad j=1,2$$

If the sequence of controls is not unique, we want to choose a sequence that minimizes the fuel, that is,

$$J = \sum_{i=0}^{N-1} |u(iT)|$$

<u>Solution</u> Let N=1. From equations (4.2.2) and (4.3.14)

$$\psi_{1,1} = c_1^T c_1 = \begin{bmatrix} 1 \\ 1-e^{-1} \end{bmatrix} \begin{bmatrix} 1 & 1-e^{-1} \end{bmatrix} = \begin{bmatrix} 1 & 0.63212 \\ 0.63212 & 0.39958 \end{bmatrix} \tag{4.3.18}$$

$$c_2^T c_2 = \begin{bmatrix} 0 \\ e^{-1} \end{bmatrix} \begin{bmatrix} 0 & e^{-1} \end{bmatrix} \quad \begin{bmatrix} 0 & 0 \\ 0 & 0.13534 \end{bmatrix} \tag{4.3.19}$$

The expression for $\bar{F}_1$ is found from equations (4.2.4) and (3.2.4)

$$F_0 = -C^{-1}D = \begin{bmatrix} 0.71828 \\ -1.71828 \end{bmatrix} \tag{4.3.20}$$

$$F_1 = -C^{-2}D = \begin{bmatrix} 3.67077 \\ -4.67077 \end{bmatrix} \tag{4.3.21}$$

$$F_2 = -C^{-3}D = \begin{bmatrix} 11.6965 \\ -12.6965 \end{bmatrix} \tag{4.3.22}$$

From equations (4.2.9), (4.3.16), (4.3.18) and (4.3.20)

$$d_{1,1} = x_0^T \psi_{1,1} \bar{F}_1 = x_0^T \psi_{1,1} F_0 = -0.88826 \tag{4.3.23}$$

$$d_{2,1} = x_0^T \psi_{2,1} \bar{F}_1 = x_0^T \psi_{2,1} F_0 = \ \ 2.7905 \tag{4.3.24}$$

Equation (4.2.8) gives

$$Q_{1,1} = \bar{F}_1^T \psi_{1,1} \bar{F}_1 = F_0^T \psi_{1,1} F_0 = 0.135335 \tag{4.3.25}$$

$$Q_{2,1} = \bar{F}_1^T \psi_{2,1} \bar{F}_1 = F_0^T \psi_{2,1} F_0 = 0.39958 \tag{4.3.26}$$

From equations (4.3.23) and (4.3.25),

$$r_1 = Q_{1,1}^{-1} d_{1,1} = -6.5634$$

Similarly, by equations (4.3.24) and (4.3.26)

$$r_2 = Q_{2,1}^{-1} d_{2,1} = 6.9836$$

Thus by Theorem 4.2.8 and the fact that $M_i = 1$, we have

$$-\frac{1}{\sqrt{0.135335}} - 6.5634 \leq u(0) \leq \frac{1}{\sqrt{0.135335}} - 6.5634$$

$$-\frac{1}{\sqrt{0.39958}} + 6.9836 \leq u(0) < \frac{1}{\sqrt{0.39958}} + 6.9836$$

These inequalities simplify to

$$-9.282 \leq u(0) \leq -3.845 \tag{4.3.27}$$

$$5.402 \leq u(0) \leq 8.566 \tag{4.3.28}$$

In addition, by (4.3.17) we require that

$$|u(0)| \leq 0.5 \tag{4.3.29}$$

By inspection there is no value of u(0) that will satisfy (4.3.27)-(4.3.29). We then increase N by 1. For N=2 we have

$$C^2 = \begin{bmatrix} 1 & 0.63212 \\ 0 & 0.36788 \end{bmatrix}^2 = \begin{bmatrix} 1 & 0.86466 \\ 0 & 0.13534 \end{bmatrix}$$

$$\psi_{i,2} = (c_1^2)^T c_1^2 = \begin{bmatrix} 1 \\ 0.86466 \end{bmatrix} \begin{bmatrix} 1 & 0.86466 \end{bmatrix} = \begin{bmatrix} 1 & 0.86466 \\ 0.86466 & 0.74764 \end{bmatrix} \tag{4.3.30}$$

$$\psi_{2,1} = (c_2^2)^T c_2^2 = \begin{bmatrix} 1 \\ 0.13534 \end{bmatrix} \begin{bmatrix} 0 & 0.13534 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0.018316 \end{bmatrix} \tag{4.3.31}$$

$$\bar{F}_2 = [F_0, F_1] = \begin{bmatrix} 0.71828 & 3.67077 \\ -1.71828 & -4.67077 \end{bmatrix} \tag{4.3.32}$$

$$Q_{1,2} = \bar{F}_2^T \psi_{1,2} \bar{F}_2 = \begin{bmatrix} 0.58899 & 0.28233 \\ 0.28233 & 0.13534 \end{bmatrix} \tag{4.3.33}$$

$$Q_{2,2} = \bar{F}_2^T \psi_{2,2} \bar{F}_2 = \begin{bmatrix} 0.05408 & 0.14700 \\ 0.14700 & 0.39958 \end{bmatrix} \tag{4.3.34}$$

$$d_{1,2}^T = x_0^T \psi_{1,2} \bar{F}_2 = (0.28855 \quad 0.13831) \tag{4.3.35}$$

$$d_{2,2}^T = x_0^T \psi_{2,2} \bar{F}_2 = (0.37766 \quad 1.02658) \tag{4.3.36}$$

$$\gamma_1 = tr(Q_{1,2}) = 0.72433 \tag{4.3.37}$$

$$\gamma_2 = tr(Q_{2,2}) = 0.45366 \tag{4.3.38}$$

Therefore, by equations (4.3.33), (4.3.35) and (4.3.37)

$$\frac{Q_{1,2}}{\gamma_1^2} d_{1,2} = \begin{bmatrix} 0.39837 \\ 0.19096 \end{bmatrix} \tag{4.3.39}$$

Similarly, by equations (4.3.34), (4.3.36) and (4.3.38)

$$\frac{Q_{2,2}}{\gamma^2} d_{2,2} = \begin{bmatrix} 0.83248 \\ 2.26291 \end{bmatrix} \tag{4.3.40}$$

$$p_1^1 = \frac{q^1}{\|q^1\|^{1/2}} = \frac{1}{[(0.28233)^2+(0.13534)^2]^{1/2}} \begin{bmatrix} 0.28233 \\ 0.13534 \end{bmatrix}$$

$$= \begin{bmatrix} 0.90175 \\ 0.43227 \end{bmatrix} \qquad (4.3.41)$$

In a similar fashion

$$p_1^2 = \begin{bmatrix} 0.34526 \\ 0.93851 \end{bmatrix} \qquad (4.3.42)$$

Substituting equations (4.3.37)-(4.3.42) into Theorem 4.2.8 gives

$$-0.7332 \leq 0.90175u(0) + 0.43227u(1) \leq 1.6168 \qquad (4.3.43)$$

$$0.9265 \leq 0.34526u(0) + 0.93851u(1) \leq 3.8959 \qquad (4.3.44)$$

If we ignore the constraint on the amplitude of the controller, a solution to inequalities (4.3.43) and (4.3.44) (if it exists) is a sequence of time-optimal controls. The region described by these inequalities is shown as the region enclosed by the parallelogram in Figure 4.3.1. For example, if we choose sequences of controls corresponding to points A, B, C, and D in this figure, we get

$u(0) = -3.4032$ $u(1) = 5.4032$	Point A	$u(0) = 1.6022$ $u(1) = 0.3978$	Point C
$u(0) = -0.2392$ $u(1) = 4.2392$	Point B	$u(0) = -1.5617$ $u(1) = 1.5618$	Point D

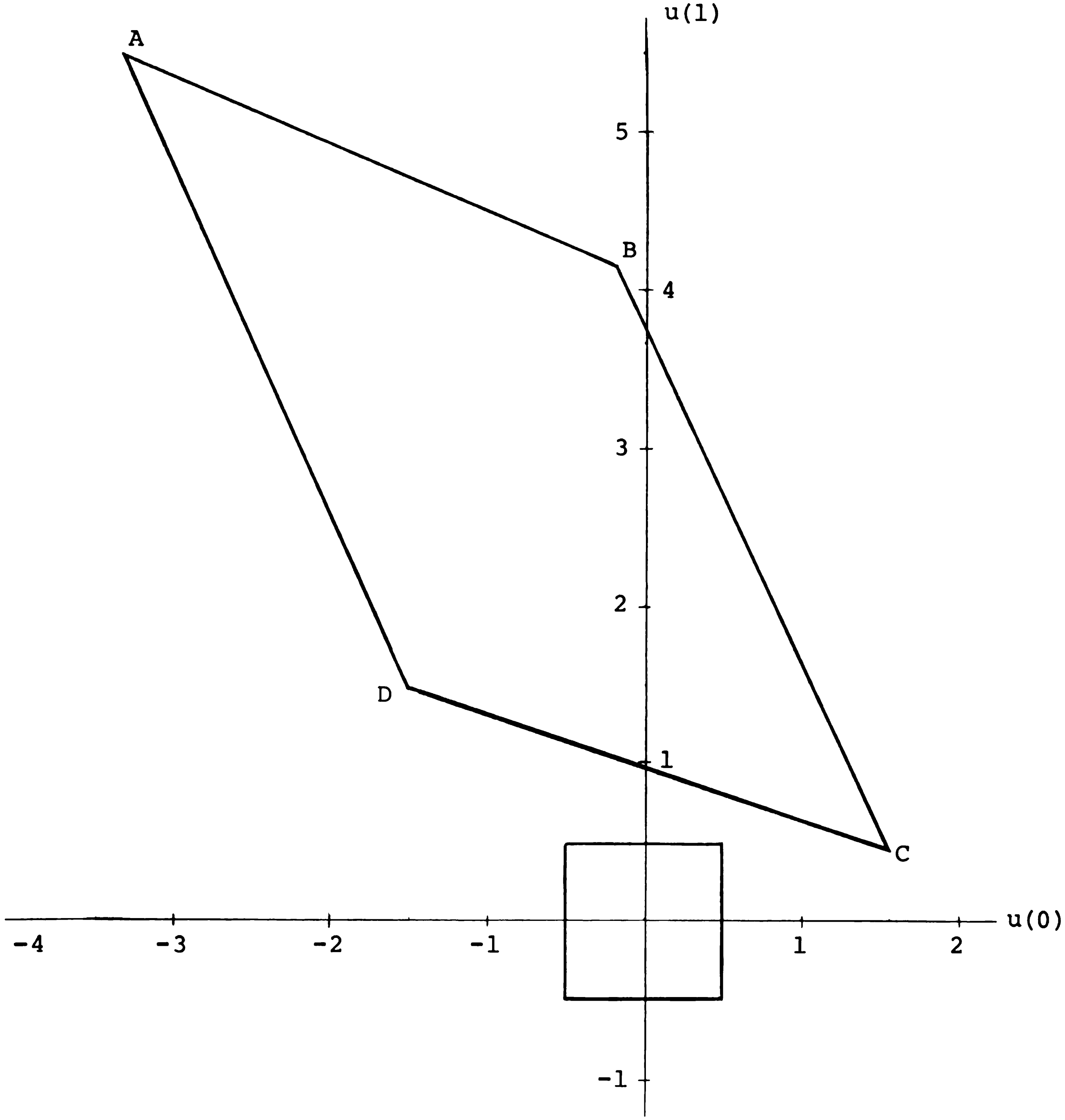


Fig. 4.3.1. Loci of unconstrained time-optmal controls and constraint set for Example 4.3.1

Using the state equation (4.3.13), these sequences of controls give the following sequences of states:

$$\begin{aligned} x(0) &= (10 \quad -12)^T \\ x(1) &= (1.1626 \; -6.5658)^T \\ x(2) &= (-1.000 \; 1.000)^T \\ \text{Fuel} &= 8.806 \end{aligned} \quad \text{Point A}$$

$$\begin{aligned} x(0) &= (10 \quad -12)^T \\ x(1) &= (3.004 \; -3.4018)^T \\ x(2) &= (0.9999 \quad -0.9999)^T \\ \text{Fuel} &= 2.000 \end{aligned} \quad \text{Point C}$$

$$\begin{aligned} x(0) &= (10 \quad -12)^T \\ x(1) &= (2.3266 \quad -4.5658)^T \\ x(2) &= (0.9999 \quad 1.0000)^T \\ \text{Fuel} &= 4.4784 \end{aligned} \quad \text{Point B}$$

$$\begin{aligned} x(0) &= (10 \quad -12)^T \\ x(1) &= (1.840 \quad -5.4017)^T \\ x(2) &= (0.9999 \; -0.999)^T \\ \text{Fuel} &= 3.1235 \end{aligned} \quad \text{Point D}$$

The constraint on the amplitude of the controller given by inequality (4.3.17) is also shown in Figure 4.3.1. In order for a solution to exist, the two sets shown in this figure must share a common point. Since this is not the case, we could conclude that no solution exists for this value of N also. However, in order to illustrate the linear programming technique, we shall not use Figure 4.3.1 and derive the same result.

To determine if N=2 is sufficiently large by using linear programming, we use the formulation given by inequalities (4.3.7) and (4.3.8). Since this system has only a single input, we simplify the notation by letting

$$u_k = u_{i,k} \qquad k=0,1,\ldots,N-1$$

From (4.3.43)-(4.3.44) and (4.3.17), we have

$$0.90175u_0 + 0.43225u_1 \leq 1.6168 + 0.5(0.90175 + 0.43225)$$
$$0.90175u_0 + 0.43225u_1 \geq -0.7332 + 0.5(0.90175 + 0.43225)$$
$$0.34525u_0 + 0.93851u_1 \leq 3.8959 + 0.5(0.34525 + 0.93851)$$
$$0.34525u_0 + 0.93851u_1 \geq 0.9265 + 0.5(0.34525 + 0.93851)$$
$$0 \leq u_0 \leq 2(0.5)$$
$$0 \leq u_1 \leq 2(0.5)$$

After simplifying the above inequalities, adding slack and surplus variables q_i and an artificial variable $\bar{z}_1$, these inequalities are converted into the following equalities.

$$\begin{array}{lllllllll}
u_0 & & +q_1 & & & & & & = 1 \\
 & u_1 & & +q_2 & & & & & = 1 \\
0.90175u_0 & +0.43225u_1 & & & +q_3 & & & & = 2.28376 \\
0.34525u_0 & +0.93851u_1 & & & & +q_4 & & & = 4.53776 \\
-0.90175u_0 & -0.43225u_1 & & & & & +q_5 & & = 0.06622 \\
0.34525u_0 & +0.93851u_1 & & & & & & -q_6 + \bar{z}_1 & = 1.56837
\end{array} \qquad (4.3.45)$$

The purpose of the artificial variable is to help form an initial basic feasible solution. The initial tableau is then constructed as shown in Table 4.3.1. Since we are trying to determine the value of N, we use Phase I of linear programming.

TABLE 4.3.1

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

	c_j		0	0	0	0	0	0	0	0	-1
c_B	Vectors in Basis	b	u_0	u_1	q_1	q_2	q_3	q_4	q_5	q_6	$\bar{z}_1$
0	q_1	1	1	0	1	0	0	0	0	0	0
0	q_2	1	0	(1)	0	1	0	0	0	0	0
0	q_3	2.28376	0.90175	0.43225	0	0	1	0	0	0	0
0	q_4	4.53776	0.34525	0.93851	0	0	0	1	0	0	0
0	q_5	0.06622	-0.90175	-0.43225	0	0	0	0	1	0	0
-1	$\bar{z}_1$	1.56837	0.34525	0.93851	0	0	0	0	0	-1	1
		-1.56837	-0.34525	-0.93851	0	0	0	0	0	1	0

r=2
k=2

TABLE 4.3.2

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

	c_j		0	0	0	0	0	0	0	0	-1
c_B	Vectors in Basis	b	u_0	u_1	q_1	q_q	q_3	q_4	q_5	q_6	$\bar{z}_1$
0	q_1	1	1	0	1	0	0	0	0	0	0
0	u_1	1	0	1	0	1	0	0	0	0	0
0	q_3	1.85151	0.90175	0	0	-0.43225	1	0	0	0	0
0	q_4	3.59925	0.34525	0	0	-0.93851	0	1	0	0	0
0	q_5	0.49847	-0.90175	0	0	0.43225	0	0	1	0	0
-1	$\bar{z}_1$	0.62986	0.34525	0	0	-0.93851	0	0	0	-1	1
		-0.62986	-0.34525	0	0	0.93851	0	0	0	1	0

r=1
k=1

TABLE 4.3.3

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

	c_j		0	0	0	0	0	0	0	0	-1
c_B	Vectors in Basis	b	u_0	u_1	q_1	q_2	q_3	q_4	q_5	q_6	$\bar{z}_1$
0	u_0	1	1	0	1	0	0	0	0	0	0
0	u_1	1	0	1	0	1	0	0	0	0	0
0	q_3	0.94976	0	0	-0.90175	-0.43225	1	0	0	0	0
0	q_4	3.2540	0	0	-0.34525	-0.93851	0	1	0	0	0
0	q_5	0.40328	0	0	0.90175	0.43225	0	0	1	0	0
-1	$\bar{z}_1$	0.28461	0	0	-0.34525	-0.93851	0	0	0	-1	1
		-0.28461	0	0	0.34525	0.93851	0	0	0	1	0

For this reason we assign a "price" of (-1) to the artificial variables and zero "price" to the other variables. This is indicated in the first row and first column of Table 4.3.1. In columns "b" through "$\bar{z}_1$" we place the coefficients given in equation (4.3.45). The last row of the table is found as follows. The last quantity in the "b" column (-1.56837) is found from equation (2.4.4), that is,

$$z = c_B b = (0\ 0\ 0\ 0\ 0\ \ -1) \begin{bmatrix} 1 \\ 1 \\ 2.28376 \\ 4.53776 \\ 0.06622 \\ 1.56837 \end{bmatrix} = -1.56837$$

The remaining terms in the last row are found by computing $z_j - c_j$ as given by equation (2.4.4). For example, the quantity (-0.34525) at the bottom of the column u_0 is found from

$$z_1 - c_1 = c_B a_1 - c_1$$

$$= (0\ 0\ 0\ 0\ 0\ -1) \begin{bmatrix} 1 \\ 0 \\ 0.90175 \\ 0.34525 \\ -0.90175 \\ 0.34525 \end{bmatrix} - 0 = -0.34525$$

where a_1 is the column headed by u_0.

Similarly, the quantity (0) at the bottom of the column headed $\bar{z}_1$ is given by

$$z_{10} - c_{10} = c_B a_{10} - c_{10}$$

$$= (0\ 0\ 0\ 0\ 0\ -1) \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} - (-1) = 0$$

where a_{10} denotes the column headed by $\bar{z}_1$.

By examining the last row of Table 4.3.1 (excluding the term in the column headed by "b"), we see that some of the quantities $z_j - c_j$ are negative. By Step 2) of the Simplex Method described in Section 2.4, we know that an optimal basic feasible solution has not been found. From Step 3) we see that u_1 should be added to the basis because (-0.93851) has the most negative value. From Step 4) the vector to be removed is determined by

$$\frac{x_{Br}}{y_{rk}} = \min\left\{\frac{1}{1}, \frac{2.28376}{0.43225}, \frac{4.53776}{0.93851}, \frac{1.56837}{0.93851}\right\} = 1$$

Thus q_2 is to be removed from the basis. The "pivot element" $y_{rk} = 1$ is circled for convenience. By Step 5) the new table is determined by equations (2.4.7)-(2.4.9).

Sample calculations:

$$\hat{y}_{rj} = \hat{y}_{2j} = \frac{y_{2j}}{y_{22}} = y_{2j}$$

In particular,

$$\hat{y}_{20} = 1 \quad \text{and} \quad \hat{y}_{21} = 0$$

Also,

$$\hat{y}_{ij} = y_{ij} - \frac{y_{i2}}{y_{22}} y_{2j} = y_{ij} - y_{i2}y_{2j} \qquad i \neq 2$$

Therefore,

$$\hat{y}_{10} - y_{10} - y_{12}y_{20} = 1 - (0)1 = 1$$

$$\hat{y}_{70} = y_{70} - y_{72}y_{20} = -1.56837 - (-0.93851)1 = -0.62986$$

In a similar fashion the rest of Table 4.3.2 is completed. By examining the last row of this table we see that we should add u_0 to the basis. From Step 4) the vector to be removed is q_1. After another set of calculations similar to those above, we obtain Table 4.3.3. By examining the last row (excluding to the column headed by "b") we see that all terms are nonnegative, that is, the "optimality condition" has been satisfied. Moreover, the artificial variable $\bar{z}_1$ is in the basis at a nonzero level (0.28461). By Theorem 2.4.1, part c), this problem has no feasible solution. This means that the target can not be reached in N=2 samples.

We now set N=3 and proceed as before. For this case

$$Q_{1,3} = \begin{bmatrix} 0.83622 & 0.70180 & 0.33641 \\ 0.70180 & 0.58899 & 0.28233 \\ 0.33641 & 0.28233 & 0.13534 \end{bmatrix}$$

$$Q_{2,3} = \begin{bmatrix} 0.007318 & 0.019894 & 0.054077 \\ 0.019894 & 0.054077 & 0.146996 \\ 0.054077 & 0.146996 & 0.399576 \end{bmatrix}$$

$$d_{1,3} = (1.28257 \quad 1.07640 \quad 0.515971)^T$$

$$d_{2,3} = (0.05111 \quad 0.13893 \quad 0.377657)^T$$

$$\gamma_1 = tr(Q_{1,3}) = 1.56055$$

$$\gamma_2 = tr(Q_{2,3}) = 0.46097$$

$$Q_{1,3}^{+} d_{1,3} = \frac{Q_{1,3}}{\gamma_1^2} d_{1,3} = \begin{bmatrix} 0.821872 \\ 0.689758 \\ 0.330635 \end{bmatrix} \qquad (4.3.46)$$

$$Q^{+}_{2,3}d_{2,3} = \frac{Q_{2,3}}{\gamma_2^2} d_{2,3} = \begin{bmatrix} 0.110875 \\ 0.301390 \\ 0.819263 \end{bmatrix} \qquad (4.3.47)$$

$$p_1^1 = \frac{q^1}{\|q^1\|^{1/2}} = \frac{1}{[(0.83622)^2+(0.70180)^2+(0.33641)^2]^{1/2}} \begin{bmatrix} 0.83622 \\ 0.70180 \\ 0.33641 \end{bmatrix}$$

$$= \begin{bmatrix} 0.73202 \\ 0.614349 \\ 0.294488 \end{bmatrix} \qquad (4.3.48)$$

Similarly,

$$p_1^2 = \begin{bmatrix} 0.126000 \\ 0.342506 \\ 0.931028 \end{bmatrix} \qquad (4.3.49)$$

From Theorem 4.2.8 and equations (4.3.46)-(4.3.49),

$$r_1 = (p_1^1)^T Q^{+}_{1,3} d_{1,3} = 1.12275 \qquad (4.3.50)$$

$$r_2 = (p_1^2)^T Q^{+}_{2,3} d_{2,3} = 0.879955 \qquad (4.3.51)$$

Thus

$$r_1 + \frac{M_1}{\sqrt{\gamma_1}} = 1.12275 + \frac{1}{\sqrt{1.56055}} = 1.92325 \qquad (4.3.52)$$

$$r_1 - \frac{M_1}{\sqrt{\gamma_1}} = 0.32246 \qquad (4.3.53)$$

Similarly,

$$r_2 + \frac{M_2}{\sqrt{\gamma_2}} = 0.879955 + \frac{1}{\sqrt{0.46097}} = 2.35282 \qquad (4.3.54)$$

$$r_2 - \frac{M_2}{\sqrt{\gamma_2}} = -0.592909 \qquad (4.3.55)$$

By Theorem 4.2.8 we require that

$$-\frac{M_i}{\sqrt{\gamma_i}} + r_i \le p_{1,i}u(0)+p_{2,i}u(1)+p_{3,i}u(2) \le \frac{M_i}{\sqrt{\gamma_i}} + r_i \quad i=1,2$$

Substituting (4.3.48), (4.3.49), (4.3.54) and (4.3.55) into this set of inequalities gives

$$0.322246 \le 0.73202u(0)+0.614349u(1)+0.294488u(2) \le 1.92325 \tag{4.3.56}$$

$$-0.592909 \le 0.12600u(0)+0.342506u(1)+0.931028u(2) \le 2.35382 \tag{4.3.57}$$

To determine if N is large enough we make the substitution given by equation (4.3.6), that is,

$$u(k) = u_k - 0.5 \quad k=0,1,2 \tag{4.3.58}$$

Substituting (4.3.58) into (4.3.56) and (4.3.57), we obtain the following inequalities.

$$1.142674 \le 0.73202u_0 + 0.614349u_1 + 0.294488u_2 \le 2.74368$$

$$0.106859 \le 0.12600u_0 + 0.342506u_1 + 0.931028u_2 \le 3.05259$$

In addition by (4.3.8) we require that

$$0 \le u_j \le 1 \qquad \text{for } j=0,1,2$$

By adding slack, surplus and artificial variables, these inequalities are converted to the following equalities.

$$\begin{array}{llllllllll}
u_0 & & & +q_1 & & & & & & =1. \\
 & u_1 & & & +q_2 & & & & & =1. \\
 & & u_2 & & & +q_3 & & & & =1. \\
\multicolumn{3}{l}{0.73202u_0+0.614349u_1+0.294488u_2} & & & & +q_4 & & & =2.74368 \\
\multicolumn{3}{l}{0.12600u_0+0.342506u_1+0.931028u_2} & & & & & +q_5 & & =3.05259 \\
\multicolumn{3}{l}{0.73202u_0+0.614349u_1+0.294488u_2} & & & & & & -q_6+\bar{z}_1 & =1.142674 \\
\multicolumn{3}{l}{0.12600u_0+0.342506u_1+0.931028u_2} & & & & & & -q_7+\bar{z}_2 & =0.106859
\end{array}$$

The linear programming initial tableau is then constructed as shown in Table 4.3.4. In this case we attempt to maximize

$$J = -\sum_{i=1}^{2} \bar{z}_i$$

As the bottom row of the tableau indicates, we have not found the optimal basic feasible solution. By using the simplex technique discussed previously, we continue to change bases until we arrive at the one indicated in Table 4.3.8. Examination of the last row shows that all terms are nonnegative. Furthermore, the artificial variables $\bar{z}_1$ and $\bar{z}_2$ have been removed from the basis. By Theorem 2.4.1 this means that we have found an optimal basic feasible solution. Thus the target can be reached in N=3 samples. From Table 4.3.8 we can also obtain a set of time-optimal controls:

$$u_0 = 1 \qquad u_1 = 0.6684 \qquad u_2 = 0 \text{ (since } u_2 \text{ is nonbasic)}$$

Substituting these quantities into (4.3.58), we get the following sequence of time-optimal controls.

$$u(0) = 0.5000 \quad u(1) = 0.1684 \quad u(2) = -0.5000 \tag{4.3.59}$$

These controls are substituted into the state equation (4.3.13) to give the following sequence of states.

$$x(0) = (10 \quad -12)^T$$

$$x(1) = (2.5985 \quad -4.0985)^T$$

$$x(2) = (0.0697 \quad -1.4013)^T$$

$$x(3) = (-1.000 \quad -0.8316)^T$$

TABLE 4.3.4

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

	c_j		0	0	0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	u_2	q_1	q_2	q_3	q_4	q_5	q_6	q_7	$\bar{z}_1$	$\bar{z}_2$
0	q_1	1	1	0	0	1	0	0	0	0	0	0	0	0
0	q_2	1	0	1	0	0	1	0	0	0	0	0	0	0
0	q_3	1	0	0	1	0	0	1	0	0	0	0	0	0
0	q_4	2.74368	0.73202	0.61435	0.29449	0	0	0	1	0	0	0	0	0
0	q_5	3.05259	0.12600	0.34251	0.93103	0	0	0	0	1	0	0	0	0
-1	$\bar{z}_1$	1.142674	0.73202	0.61435	0.29449	0	0	0	0	0	-1	0	1	0
-1	$\bar{z}_2$	0.106859	0.12600	0.34251	0.93103	0	0	0	0	0	0	-1	0	1
		-1.24953	-0.85802	-0.95686	-1.22552	0	0	0	0	0	1	1	0	0

r=7
k=3

TABLE 4.3.5

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

	c_j		0	0	0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	u_2	q_1	q_2	q_3	q_4	q_5	q_6	q_7	$\bar{z}_1$	$\bar{z}_2$
0	q_1	1	1	0	0	1	0	0	0	0	0	0	0	0
0	q_2	1	0	1	0	0	1	0	0	0	0	0	0	0
0	q_3	0.88523	-0.135334	-0.36788	0	0	0	1	0	0	0	1.07408	0	-1.07408
0	q_4	2.70988	0.692166	0.50601	0	0	0	0	1	0	0	0.316306	0	-0.316306
0	q_5	2.94573	0	0	0	0	0	0	0	1	0	1	0	-1
-1	$\bar{z}_1$	1.10887	0.692166	0.50601	0	0	0	0	0	0	-1	0.316306	1	-0.316306
0	u_2	0.114775	0.135334	0.36788	1	0	0	0	0	0	0	-1.07408	0	1.07408
		-1.108871	-0.692166	-0.50601	0	0	0	0	0	0	1	-0.316306	0	1.31631

r=7
k=1

TABLE 4.3.6

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

	c_j		0	0	0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	u_2	q_1	q_2	q_3	q_4	q_5	q_6	q_7	$\bar{z}_1$	$\bar{z}_2$
0	q_1	0.151913	0	-2.71831	-7.38913	1	0	0	0	0	0	7.93651	0	-7.93651
0	q_2	1	0	1	0	0	1	0	0	0	0	0	0	0
0	q_3	1	0	0	1	0	0	1	0	0	0	0	0	0
0	q_4	2.12286	0	-1.37551	-5.11450	0	0	0	1	0	0	5.80969	0	-5.80969
0	q_5	2.94573	0	0	0	0	0	0	0	1	0	1	0	-1
-1	$\bar{z}_1$	0.521853	0	-1.37551	-5.11450	0	0	0	0	0	-1	5.80969	1	-5.80969
0	u_0	0.848087	1	2.71831	7.3891	0	0	0	0	0	0	-7.83651	0	7.93651
		-0.52185	0	1.37551	5.11450	0	0	0	0	0	1	-5.80969	0	6.80969

r=1
k=10

TABLE 4.3.7

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

c_B	c_j / Vectors in Basis	b	u_0	u_1	u_2	q_1	q_2	q_3	q_4	q_5	q_6	q_7	$\bar{z}_1$	$\bar{z}_2$
	c_j		0	0	0	0	0	0	0	0	0	0	-1	-1
0	q_7	0.191410	0	-0.342507	-0.931030	0.12600	0	0	0	0	0	1	0	-1
0	q_2	1	0	1	0	0	1	0	0	0	0	0	0	0
0	q_3	1	0	0	1	0	0	1	0	0	0	0	0	0
0	q_4	2.01166	0	0.614350	0.294498	-0.73202	0	0	1	0	0	0	0	0
0	q_5	2.92659	0	0.342507	0.931030	0.12600	0	0	0	1	0	0	0	0
-1	$\bar{z}_1$	0.410645	0	0.614350	0.294498	-0.73202	0	0	0	0	-1	0	1	0
0	u_0	1	1	0	0	1	0	0	0	0	0	0	0	0
		-0.410645	0	-0.614350	-0.294498	0.73202	0	0	0	0	1	0	0	1

r=6
k=2

TABLE 4.3.8

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.3.1

	c_j		0	0	0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	u_2	q_1	q_2	q_3	q_4	q_5	q_6	q_7	$\bar{z}_1$	$\bar{z}_2$
0	q_7	0.24808	0	0	-0.76684	-0.28211	0	0	0	0	-0.55751	1	0.55751	-1
0	q_2	0.33158	0	0	-0.47936	1.19154	1	0	0	0	1.62774	0	-1.62774	0
0	q_3	1	0	0	1	0	0	1	0	0	0	0	0	0
0	q_4	1.6010	0	0	0	0	0	0	1	0	1	0	-1	0
0	q_5	2.69765	0	0	0.76684	0.53411	0	0	0	1	0.55751	0	-0.55751	0
0	u_1	0.66842	0	1	0.47936	-1.19154	0	0	0	0	-1.62774	0	1.62774	0
0	u_0	1	1	0	0	1	0	0	0	0	0	0	0	0
			0	0	0	0	0	0	0	0	0	0	1	1

$$\text{Fuel} = 0.5000 + 0.1684 + 0.5000 = 1.1684$$

Thus the sequence of controls given by (4.3.59) does drive the initial state to the target in three samples.

Returning again to Table 4.3.8, we note that the terms in the last row are zero, that is, $z_j - c_j = 0$ for $j=2,3,\ldots,10$. Also, there are $y_{ij} > 0$ terms in these columns. By Theorem 2.4.2, part b) this implies that the basis can be changed, and again we will have an optimal basic feasible solution. In other words, the time-optimal sequence of controls is not unique. We now consider the problem of determining a sequence of controls that minimizes the fuel. Now that N=3 is known we can use the formulation given by inequalities (4.3.4) and (4.3.5). After adding slack, surplus, and artificial variables, these inequalities are converted to the following equalities. Because the system in this example has only one input, the notation is simplified by setting

$$u_p(kT) = u_{i,p}(kT)$$

$$u(kT) = u_{i,q}(kT)$$

$$
\begin{aligned}
&u_p(0) + q_1 = 0.5\\
&u_p(1) + q_2 = 0.5\\
&u_p(2) + q_3 = 0.5\\
&u_q(0) + q_4 = 0.5\\
&u_q(1) + q_5 = 0.5\\
&u_q(2) + q_6 = 0.5\\
&0.73202u_p(0)+0.61435u_p(1)+0.29449u_p(2)-0.73202u_q(0)-0.61435u_q(1)-0.29449u_q(2) + q_7 = 1.92325\\
&0.12600u_p(0)+0.34251u_p(1)+0.93103u_p(2)-0.12600u_q(0)-0.34251u_q(1)-0.93103u_q(2) + q_8 = 2.35282\\
&-0.12600u_p(0)-0.34251u_p(1)-0.93103u_p(2)+0.12600u_q(0)+0.34251u_q(1)+0.93103u_q(2) + q_9 = 0.59291\\
&0.73202u_p(0)+0.61435u_p(1)+0.29449u_p(2)-0.73202u_q(0)-0.61435u_q(1)-0.29449u_q(2) - q_{10} + \bar{z}_1 = 0.32225
\end{aligned}
$$

The objective function is given by equation (4.3.11), that is, we wish to maximize

$$J = -\sum_{i=0}^{2}(u_p(i)+u_q(i))$$

$$u_p(i) \geq 0$$

$$u_q(i) \geq 0$$

The above equations represent a well-formulated linear programming problem; the solution of which is the following:

$$u_p(0) = 0.4402 \qquad u_q(0) = 0.0000$$

$$u_p(1) = 0.0000 \qquad u_q(1) = 0.0000$$

$$u_p(2) = 0.0000 \qquad u_q(2) = 0.0000$$

Substituting these quantities into (4.3.3), we get the following sequence of fuel-optimal controls:

$$u(0) = 0.4402 \quad u(1) = 0.0000 \quad u(2) = 0.0000$$

The corresponding sequence of states is

$$x(0) = (10 \quad -12)^T$$

$$x(1) = (2.576 \quad -4.136)^T$$

$$x(2) = (-0.0381 \quad -1.522)^T$$

$$x(3) = (-1.000 \quad -0.5598)^T$$

$$\text{Fuel} = 0.4402$$

For purposes of comparison, a sequence of controls which requires maximum fuel is the following.

$$u(0) = u(1) = u(2) = 0.5$$

The corresponding sequence of states is

$$x(0) = (10 \quad -12)^T$$

$$x(1) = (2.598 \quad -4.098)^T$$

$$x(2) = (0.1917 \quad -1.917)^T$$

$$x(3) = (-0.3777 \quad -0.1223)^T$$

$$\text{Fuel} = 1.5$$

The maximum fuel trajectory requires approximately 3.4 times as much fuel as the minimum fuel trajectory.

4.4 Statement and Solution of Stochastic Time-Optimal Control Problem

This section is devoted to the study of a stochastic version of the problem considered in the previous section.

Problem Statement Given a system described by

$$x[(k+1)T] = Cx(kT) + Du(kT) + Ew(kT) \qquad k=0,1,\ldots,N-1 \tag{4.4.1}$$

where

$x(kT)$ is a $n \times 1$ state vector
C is a $n \times n$ nonsingular matrix
D is a $n \times m$ matrix
E is a $n \times r$ matrix
$u(kT)$ is a $m \times 1$ nonrandom vector to be determined
$w(kT)$ is a sequence of $r \times 1$
zero mean independent random vectors

The initial state is

$$x(0) = x_0 + v$$

where x_0 is a known $n \times 1$ vector and v is a random vector whose components are independent of $w(kT)$ for $k=0,1,\ldots,N-1$. We want to find

1) the smallest value of N such that

$$E(x_i^2(NT)) \leq M_i^2 \qquad i=1,2,\ldots,n. \tag{4.4.2}$$

where the $x_i(NT)$ are the components of $x(NT)$ and

2) The sequence of open-loop controls that drive the initial state to the target described by inequality (4.4.2). By open-loop we mean that the controls are assumed nonrandom and can be precomputed. If the sequence of controls is not unique, we want to choose one that minimizes the fuel, that is, minimizes

$$J = \sum_{k=0}^{N-1} \sum_{i=1}^{m} |u_i(kT)|$$

where the $u_i(kT)$ are the components of $u(kT)$. It is assumed that the amplitude of the controls is constrained, that is,

$$|u_i(kT)| \leq G_{i,k} \quad i=1,2,\ldots,m \quad k=0,1,\ldots,N-1 \qquad (4.4.3)$$

<u>Solution</u> As in the problem with the hyperspherical target, we make the following preliminary definitions.

<u>Definition</u> Let $W = (w(0), w(T), \ldots, w[(N-1)T])^T$. The matrix R_N is defined by

$$R_N = E(WW^T) \quad (Nr \times Nr \text{ matrix}) \qquad (4.4.4)$$

<u>Definition</u> The covariance matrix V, is defined by

$$V = E(vv^T) \quad (n \times n \text{ matrix}) \qquad (4.4.5)$$

The solution of equation (4.4.1) is

$$x(NT) \; C^N x(0) - C^N \sum_{j=0}^{N-1} F_j u(jT) - C^N \sum_{j=0}^{N-1} H_j w(jT) \qquad (4.4.6)$$

where

$$F_j = -C^{-(j+1)} D \qquad j=0,1,\ldots,N-1 \qquad (4.4.7)$$

$$H_j = -C^{-(j+1)} E \qquad j=0,1,\ldots,N-1 \qquad (4.4.8)$$

The solution is obtained in the same way as that for the deterministic system described in Section 3.2.

If $x(NT)=(x_1(NT), x_2(NT),\ldots,x_n(NT))^T$, then

$$x_i(NT)=c_i^N[x(0)-\sum_{j=0}^{N-1} F_j u(jT) - \sum_{j=0}^{N-1} H_j w(jT)]$$

where c_i^N is the i-th row of C^N. Thus

$$x_i^2(NT) = [x(0) - \sum_{j=0}^{N-1} F_j u(jT) - \sum_{j=0}^{N-1} H_j w(jT)]\psi_{i,N}[x(0)-\sum_{j=0}^{N-1} F_j u(jT) - \sum_{j=0}^{N-1} H_j w(jT)] \quad (4.4.9)$$

where

$$\psi_{i,N} = (c_i^N)^T c_i^N \quad \text{(n x n matrix)} \quad (4.4.10)$$

Expanding equation (4.4.9), we get

$$x_i^2(NT)=x^T(0)\psi_{i,N}x(0)-2x^T(0)\psi_{i,N}\sum_{j=0}^{N-1} F_j u(jT)$$

$$-2x^T(0)\psi_{i,N}\sum_{j=0}^{N-1} H_j w(jT)+(\sum_{j=0}^{N-1} F_j u(jT))^T\psi_{i,N}(\sum_{j=0}^{N-1} F_j u(jT))$$

$$+2(\sum_{j=0}^{N-1} F_j u(jT))^T\psi_{i,N}(\sum_{j=0}^{N-1} H_j w(jT))$$

$$+(\sum_{j=0}^{N-1} H_j w(jT))^T\psi_{i,N}(\sum_{j=0}^{N-1} H_j w(jT)) \quad (4.4.11)$$

Taking the expected value of both sides of equation (4.4.11) while noting the assumptions of independence and zero mean, we get

$$E(x_i^2(NT))=x_0^T\psi_{i,N}x_0 + E(v^T\psi_{i,N}v) - 2x(0)\psi_{i,N}\sum_{j=0}^{N-1}F_ju(jT)$$

$$+(\sum_{j=0}^{N-1}F_ju(jT))^T\psi_{i,N}(\sum_{j=0}^{N-1}F_ju(jT))$$

$$+ E\{(\sum_{j=0}^{N-1}H_jw(jT))^T\psi_{i,N}(\sum_{j=0}^{N-1}H_jw(jT))\}$$

$$= x_0^T\psi_{i,N}x_0 + E(v^T\psi_{i,N}v) - 2\bar{d}_{i,N}^TU$$

$$+ U^TQ_{i,N}U + E(W^TS_{i,N}W) \qquad (4.4.12)$$

where

$$\bar{d}_{i,N} = x_0^T\psi_{i,N}\bar{F}_N \qquad (4.4.13)$$

$$\bar{F}_N = [F_0,F_1,\ldots,F_{N-1}] \qquad (4.4.14)$$

$$S_{i,N} = [H_0,H_1,\ldots,H_{N-1}]\psi_{i,N}[H_0,H_1,\ldots,H_{N-1}] \qquad (4.4.15)$$

$$Q_{i,N} = \bar{F}_N^T\psi_{i,N}\bar{F}_N \qquad (4.4.16)$$

Using the same argument as that used in obtaining equations (3.6.18) and (3.6.19), equation (4.4.12) can be written in the following way.

$$E(x_i^2(NT))=U^T Q_{i,N}U-2\bar{d}_{i,N}^TU+x_0^T\psi_{i,N}x_0+tr(S_{i,N}R_N)$$

$$+ tr(\psi_{i,N}V)-M_i^2 \qquad (4.4.17)$$

Since we require that $E(x_i^2(NT))-M_i \leq 0$, we are looking for N and U such that

$$f(U) = U^T Q_{i,N} U - 2\bar{d}_{i,N}^T U + \bar{e}_{i,N} \leq 0 \tag{4.4.18}$$

where

$$\bar{e}_{i,N} = x_0^T \psi_{i,N} x_0 + tr(S_{i,N} R_N) + tr(\psi_{i,N} V) - M_i^2 \tag{4.4.19}$$

By comparing inequality (4.4.18) with the corresponding deterministic inequality (4.2.7), we see they are the same except that $d_{i,N}$ is replaced by $\bar{d}_{i,N}$ and $e_{i,N}$ is replaced by $\bar{e}_{i,N}$. Thus the form of the solution in the stochastic case is similar to that of the deterministic case. Starting with inequality (4.4.18), we can then repeat the steps used in obtaining the sequence of controls in the deterministic system after substituting $\bar{d}_{i,N}$ for $d_{i,N}$ and $\bar{e}_{i,N}$ for $e_{i,N}$. In particular, we simplify inequality (4.4.18) by substituting

$$U = P_{i,N} Y_i + C_{i,N} \tag{4.4.20}$$

where

$$P_{i,N}^T P_{i,N} = I$$

into (4.4.18) to give

$$g(Y_i) = Y_i^T P_{i,N}^T Q_{i,N} P_{i,N} Y_i + \bar{L}_{i,N} Y_i + \bar{g}_{i,N} \leq 0 \tag{4.4.21}$$

where

$$\bar{L}_{i,N} = P_{i,N}^T (Q_{i,N} C_{i,N} - \bar{d}_{i,N}) \tag{4.4.22}$$

$$\bar{g}_{i,N} = C_{i,N}^T Q_{i,N} C_{i,N} - 2\bar{d}_{i,N}^T C_{i,N} + \bar{e}_{i,N} \tag{4.4.23}$$

By using the same steps as those used to obtain Theorem 4.2.6, we have the following theorem.

<u>Theorem 4.4.1</u> The expression for $g(Y_i)$ in (4.4.21) reduces to

$$g(Y_i)=y_{0,i}^2\gamma_i+tr(S_{i,N}R_N)+tr(\psi_{i,N}V)-M_i^2 \leq 0 \qquad (4.4.24)$$

where $\gamma_i = tr(Q_{i,N})$ and $Y_i = (y_{0,i},y_{1,i},\ldots,y_{Nm-1,i})^T$

Since $\gamma_i \geq 0$, it follows from inequality (4.4.24) that in order for a solution to exist, it is necessary that

$$tr(S_{i,N}R_N)+tr(\psi_{i,N}V) \leq M_i^2 \quad \text{for } i=1,2,\ldots,n$$

For the stochastic case, Theorem 4.2.8 is replaced by the following theorem.

<u>Theorem 4.4.2</u> Let $Q_{i,N} \neq 0$. If a sequence of time-optimal controls exists, it is a solution to the following set of inequalities.

$$-\frac{\bar{M}_i}{\sqrt{\gamma_i}} + r_i \leq p_{1,i}u_1(0)+p_{2,i}u_2(0)+\ldots+p_{m,i}u_m(0)+p_{m+1,i}u_1(T)$$

$$+\ldots+p_{2m,i}u_m(T)+\ldots+p_{Nm-m+1,i}u_1[(N-1)T]$$

$$+\ldots+p_{Nm,i}u_m[(N-1)T] \leq \frac{\bar{M}_i}{\sqrt{\gamma_i}} + r_i \qquad i=1,2,\ldots,n$$

where

$$r_i = (p_1^i)^T \frac{Q_{i,N}}{\gamma_i^2} \bar{d}_{i,N}$$

$$p_1^i = \frac{+q^i}{\|q^i\|^{1/2}} \text{ and } q^i \text{ is any nonzero column of } Q_{i,N}.$$

$$\bar{M}_i = (M_i^2 - tr(S_{i,N}R_N) - tr(\psi_{i,N}V))^{\frac{1}{2}}$$

The solution does not depend on which nonzero column of $Q_{i,N}$ we choose or whether we choose the plus or minus sign in the expression for p_1^i. If Nm=1, we set $\gamma_i = Q_{i,N}$ and $p_1^i = \pm 1$.

Similarly, Theorem 4.2.9 for the deterministic system becomes the following theorem.

Theorem 4.4.3 If $Q_{i,N} = 0$, the following is true.

a) If a solution exists, a time-optimal sequence of controls is a solution of the following set of inequalities.

$$\bar{\alpha}_{1,i}u_1(0)+\bar{\alpha}_{2,i}u_2(0)+\ldots+\bar{\alpha}_{m,i}u_m(0)+\bar{\alpha}_{m+1,i}u_1(T)+\ldots$$

$$+\bar{\alpha}_{2m,i}u_m(T)+\ldots+\bar{\alpha}_{Nm-m+1,i}u_1[(N-1)T]+\ldots$$

$$+\bar{\alpha}_{Nm,i}u_m[(N-1)T] \leq M_i^2-x_0^T\psi_{i,N}x_0-tr(S_{i,N}R_N)$$

$$-tr(\psi_{i,N}V) \qquad i=1,2,\ldots n$$

where $(\bar{\alpha}_{1,i},\bar{\alpha}_{2,i},\ldots,\bar{\alpha}_{Nm,i}) = 2x_0^T\psi_{i,N}\bar{F}_N$ (1xNm vector)

b) If $N \geq n/m$ and the nxNm matrix $\bar{F}_N$ is of rank n, then we have the following two cases.

1) If $x_0^T\psi_{i,N}x_0+tr(S_{i,N}R_N)+tr(\psi_{i,N}V) \leq M_i^2$, $i=1,2,\ldots,n$, then the sequence of time-optimal controls is arbitrary. In particular, the sequence U = 0 is the minimum energy and fuel solution.

2) If $x_0^T\psi_{i,N}x_0+tr(S_{i,N}R_N)+tr(\psi_{i,N}V) > M_i^2$ for some i, no solution exists for this value of N.

By comparing Theorem 4.4.2 with Theorem 2.4.8, we see that the only difference is that in the stochastic case $\bar{M}_i$ replaces M_i. From Theorem 4.4.2 we see that $\bar{M}_i \leq M_i$. In other words, the effect of the noise is the same as that of reducing the dimensions (M_i) of the target set of the corresponding deterministic system.

<u>Example 4.4.1</u> Consider the system described by

$$x(k+1) = Cx(k) + Du(k) + Ew(k) \qquad k=0,1,\ldots,N-1 \qquad (4.4.25)$$

$$x(0) = x_0 + v \qquad (4.4.26)$$

where

$$C = \begin{bmatrix} 1 & 1-e^{-1} \\ 0 & e^{-1} \end{bmatrix} \qquad (4.4.27)$$

$$D = E = (e^{-1} \quad 1-e^{-1})^T \qquad (4.4.28)$$

$$x_0 = (10 \quad -12)^T \qquad (4.4.29)$$

C, D, and x_0 have the same values as those given in Example 4.3.1 for the deterministic system. It is assumed that v, w(k) for k=0,1,...,N-1 are independent Gaussian random variables such that

$$E(v) = E(w(k)) = 0 \qquad k=0,1,\ldots,N-1 \qquad (4.4.30)$$

$$V = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \qquad (4.4.31)$$

$$R_N = E(WW^T) = 0.5I_N \quad (I_N \text{ is an NxN identity matrix}) \qquad (4.4.32)$$

$$M_i = 1 \qquad i=1,2 \qquad (4.4.33)$$

We wish to determine the smallest value of N and a corresponding sequence of controls such that $E(x_i^2(NT)) \leq 1$.

If the sequence is not unique, we want to choose one that minimizes

$$J = \sum_{i=0}^{N-1} |u(iT)| \qquad (4.4.34)$$

It is assumed that the amplitude of the controls is constrained, that is,

$$|u(iT)| \leq 1 \quad i=0,1,\ldots,N-1 \qquad (4.4.35)$$

<u>Solution</u> Setting N=1, Theorem 4.4.2 requires that

$$-\frac{\bar{M}_i}{\sqrt{\gamma_i}} + r_i \leq u(0) \leq \frac{\bar{M}_i}{\sqrt{\gamma_i}} + r_i \quad i=1,2 \qquad (4.4.35)$$

where

$$r_i = \frac{Q_{i,1}}{\gamma_i^2} \bar{d}_{i,1} \qquad (4.4.37)$$

$Q_{i,N}$ and $\psi_{i,N}$ are the same for the deterministic and stochastic systems since C and D are the same. Thus, from Example 4.3.1,

$$\gamma_1 = Q_{1,1} = 0.135335 \qquad (4.4.38)$$

$$\gamma_2 = Q_{2,1} = 0.39958 \qquad (4.4.39)$$

$$\psi_{1,1} = \begin{bmatrix} 1 & 0.63212 \\ 0.63212 & 0.39958 \end{bmatrix} \qquad (4.4.40)$$

$$\psi_{2,1} = \begin{bmatrix} 0 & 0 \\ 0 & 0.13534 \end{bmatrix} \qquad (4.4.41)$$

Also, $\bar{d}_{1,1} = x_0^T \psi_{1,1} F_0 = (10 \quad -12) \begin{bmatrix} 1 & 0.63212 \\ 0.63212 & 0.39958 \end{bmatrix} \begin{bmatrix} 0.71828 \\ -1.71828 \end{bmatrix}$

$$= -0.88826 \qquad (4.4.42)$$

$$\bar{d}_{2,1}=x_0^T\psi_{2,1}F_0 = 2.7905 \qquad (4.4.43)$$

Thus by (4.4.37)

$$r_1 = \frac{-0.88826}{0.135335} = -6.5634 \qquad (4.4.44)$$

and

$$\bar{M}_1=(M_1^2-tr(S_{1,1}R_1)-tr(\psi_{1,1}V))^{\frac{1}{2}}$$

$$=\left\{1-tr(0.135335)(0.5)-tr\begin{bmatrix}1 & 0.63212\\ 0.63212 & 0.39958\end{bmatrix}\begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix}\right\}^{1/2}$$

$$= 0.89015 \qquad (4.4.45)$$

Similarly,

$$r_2 = 6.98358 \qquad (4.4.46)$$

$$\bar{M}_2 = 0.88695 \qquad (4.4.47)$$

Substituting equations (4.4.38)-(4.4.39) and (4.4.44)-(4.4.47) into (4.4.36) gives

$$-\frac{0.89015}{0.135335} - 6.5634 \leq u(0) \leq \frac{0.89015}{0.135335} - 6.5634$$

$$\frac{0.88695}{0.39958} + 6.98358 \leq u(0) \leq \frac{0.88695}{0.39958} + 6.98358$$

After simplification, the above inequalities become

$$-8.9831 \leq u(0) \leq -4.1437 \qquad (4.4.48)$$

$$5.5800 \leq u(0) \leq 8.3867 \qquad (4.4.49)$$

In addition we require that

$$-1 \leq u(0) \leq 1 \qquad (4.4.50)$$

By inspection, there is no value of u(0) which satisfies (4.4.48)-(4.4.50) so we must increase the value of N by 1. For N=2 we have

$$\psi_{1,2} = \begin{bmatrix} 1 & 0.86467 \\ 0.86467 & 0.74765 \end{bmatrix} \quad (4.4.51)$$

$$\psi_{2,2} = \begin{bmatrix} 0 & 0 \\ 0 & 0.18316 \end{bmatrix} \quad (4.4.52)$$

$$Q_{1,2} = \begin{bmatrix} 0.58899 & 0.28233 \\ 0.28233 & 0.13534 \end{bmatrix} \quad (4.4.53)$$

$$Q_{2,2} = \begin{bmatrix} 0.05408 & 0.14700 \\ 0.14700 & 0.39958 \end{bmatrix} \quad (4.4.54)$$

$$\bar{d}_{1,2} = (0.28855 \quad 0.13831)^T \quad (4.4.55)$$

$$\bar{d}_{2,2} = (0.37766 \quad 1.02658)^T \quad (4.4.56)$$

$$\gamma_1 = tr(Q_{1,2}) = 0.72433 \quad (4.4.57)$$

$$\gamma_2 = tr(Q_{2,2}) = 0.45366 \quad (4.4.58)$$

$$p_1^1 = \begin{bmatrix} 0.90175 \\ 0.43227 \end{bmatrix} \quad (4.4.59)$$

$$p_1^2 = \begin{bmatrix} 0.34526 \\ 0.93851 \end{bmatrix} \quad (4.4.60)$$

Then

$$r_1 = 0.44177 \quad \text{and} \quad \bar{M}_1 = 0.68049 \quad (4.4.61)$$

Similarly,

$$r_2 = 2.39535 \quad \text{and} \quad \bar{M}_2 = 0.86882 \quad (4.4.62)$$

Therefore, by Theorem 4.4.2 and equations (4.4.57)-(4.4.62),

$$-\frac{0.68049}{\sqrt{0.72433}} + 0.44177 \le 0.90175u(0) + 0.43227u(1) \le \frac{0.68049}{\sqrt{0.72433}} + 0.44177$$

$$-\frac{0.86882}{\sqrt{0.45366}} + 2.39535 \le 0.34526u(0) + 0.93851u(1) \le \frac{0.86882}{\sqrt{0.45366}} + 2.39535$$

After simplification, these inequalities become

$$-0.35780 \le 0.90175u(0) + 0.43227u(1) \le 1.24133 \quad (4.4.63)$$

$$1.10542 \le 0.34526u(0) + 0.93851u(1) \le 3.68528 \quad (4.4.64)$$

To determine if N is large enough, we use the linear programming technique discussed in Section 4.3. From (4.3.6) we define

$$u(0) = u_0 - 1 \quad (4.4.65)$$

$$u(1) = u_1 - 1 \quad (4.4.66)$$

Substituting equations (4.4.65)-(4.4.66) into (4.4.63)-(4.4.64) and adding the necessary slack, surplus, and artificial variables, we obtain the following equations.

$$\begin{array}{llllll}
u_0 & & +q_1 & & & = 2. \\
& u_1 & +q_2 & & & = 2. \\
0.90175u_0 & + 0.43225u_1 & +q_3 & & & = 2.57533 \\
0.34526u_0 & + 0.93851u_1 & +q_4 & & & = 4.96905 \\
0.90175u_0 & + 0.43225u_1 & & -q_5+\bar{z}_1 & & = 0.97620 \\
0.34526u_0 & + 0.93851u_1 & & & -q_6+\bar{z}_2 & = 2.38919
\end{array} \quad (4.4.67)$$

For a performance index we choose to maximize

$$J = - \sum_{i=1}^{2} \bar{z}_i \qquad (4.4.68$$

where the $\bar{z}_i$ are the artificial variables.

The initial tableau for the linear programming solution is shown in Table 4.4.1. Examination of the bottom row indicates we have not found an optimal basic feasible solution. By repeatedly changing bases we arrive at Table 4.4.4. Since the bottom row is nonnegative and no artificial variables appear in the basis, we have satisfied the optimality criterion and a solution exists for N=2. From Table 4.4.4 we see that

$$u_0 = 1.4834 \qquad u_1 = 2.$$

From equations (4.4.65) and (4.4.66) we have that

$$u(0) = 0.4834 \qquad (4.4.69)$$

$$u(1) = 1. \qquad (4.4.70)$$

$$\text{Fuel} = 1.4834 \qquad (4.4.71)$$

Table 4.4.4 indicates that the solution is not unique. To determine a sequence that requires minimum fuel we use the formulation given in equations (4.3.3) and (4.3.4)(with M_i replaced by $\bar{M}_i$). That is, we make the substitution

$$u(0) = u_p(0) - u_q(0) \qquad (4.4.72)$$

$$u(1) = u_p(1) - u_q(1) \qquad (4.4.73)$$

in the inequalities given by (4.4.63) and (4.4.64). After adding slack, surplus, and artificial variables, these inequalities become the following equalities.

TABLE 4.4.1

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.4.1

	c_j		0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	q_1	q_2	q_3	q_4	q_5	q_6	$\bar{z}_1$	$\bar{z}_2$
0	q_1	2	1	0	1	0	0	0	0	0	0	0
0	q_2	2	0	1	0	1	0	0	0	0	0	0
0	q_3	2.57533	0.90175	0.43225	0	0	1	0	0	0	0	0
0	q_4	4.96905	0.34526	0.93851	0	0	0	1	0	0	0	0
-1	$\bar{z}_1$	0.97620	0.90175	0.43225	0	0	0	0	-1	0	1	0
-1	$\bar{z}_2$	2.38919	0.34526	0.93851	0	0	0	0	0	-1	0	1
		-3.36539	-1.24701	-1.37076	0	0	0	0	1	1	0	0

k=2
r=2

TABLE 4.4.2

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.4.1

	c_j		0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	q_1	q_2	q_3	q_4	q_5	q_6	$\bar{z}_1$	$\bar{z}_2$
0	q_1	2	1	0	1	0	0	0	0	0	0	0
0	u_1	2	0	1	0	1	0	0	0	0	0	0
0	q_3	1.7108	0.90175	0	0	-0.43225	1	0	0	0	0	0
0	q_4	3.0920	0.34526	0	0	-0.93851	0	1	0	0	0	0
-1	$\bar{z}_1$	0.11170	0.90175	0	0	-0.43225	0	0	-1	0	1	0
-1	$\bar{z}_2$	0.51217	0.34526	0	0	-0.93851	0	0	0	-1	0	1
		-0.62387	-1.24701	0	0	1.37076	0	0	1	1	0	0

k=1
r=5

TABLE 4.4.3

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.4.1

	c_j		0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	q_1	q_2	q_3	q_4	q_5	q_6	$\bar{z}_1$	$\bar{z}_2$
0	q_1	1.8761	0	0	1	0.47935	0	0	1.1090	0	-1.1090	0
0	u_1	2	0	1	0	1	0	0	0	0	0	0
0	q_3	1.5991	0	0	0	0	1	0	1	0	-1	0
0	q_4	3.0492	0	0	0	-0.77301	0	1	0.38288	0	-0.38288	0
0	u_0	0.12397	1	0	0	-0.4794	0	0	-1.1090	0	1.1090	0
-1	$\bar{z}_2$	0.46940	0	0	0	-0.77301	0	0	0.38288	-1	-0.38288	1
		-0.46940	0	0	0	0.77301	0	0	-0.38288	1	1.38288	0

k=7
r=6

TABLE 4.4.4

LINEAR PROGRAMMING TABLEAU FOR EXAMPLE 4.4.1

	c_j		0	0	0	0	0	0	0	0	-1	-1
c_B	Vectors in Basis	b	u_0	u_1	q_1	q_2	q_3	q_4	q_5	q_6	$\bar{z}_1$	$\bar{z}_2$
0	q_1	0.5165	0	0	1	2.7183	0	0	0	2.8965	0	-2.8965
0	u_1	2	0	1	0	1	0	0	0	0	0	0
0	q_3	0.3731	0	0	0	2.0189	1	0	0	2.6118	0	-2.6118
0	q_4	2.5798	0	0	0	0	0	1	0	1	0	-1
0	u_0	1.4834	1	0	0	-2.7184	0	0	0	-2.8965	0	2.8965
0	q_5	1.2260	0	0	0	-2.0189	0	0	1	-2.6118	-1	2.6118
		0	0	0	0	0	0	0	0	0	1	1

$$
\begin{aligned}
&u_p(0) + q_1 = 1\\
&u_p(1) + q_2 = 1\\
&u_q(0) + q_3 = 1\\
&u_q(1) + q_4 = 1\\
&0.90175u_p(0)+0.43225u_p(1)-0.90175u_q(0)-0.43225u_q(1) + q_5 = 1.24133\\
&0.34526u_p(0)+0.93851u_p(1)-0.34526u_q(0)-0.93851u_q(1) + q_6 = 3.68528\\
&-0.90175u_p(0)-0.43225u_p(1)+0.90175u_q(0)+0.43225u_q(1) + q_7 = 0.35780\\
&0.34526u_p(0)+0.93851u_p(1)-0.34526u_q(0)-0.93851u_q(1) + q_8+\bar{z}_1 = 1.10542
\end{aligned}
\tag{4.4.74}
$$

By equation (4.3.11) we wish to maximize

$$J = -\sum_{i=1}^{2} (u_p(i) + u_q(i)) \qquad (4.4.75)$$

subject to

$$u_p(i), u_q(i) \geq 0 \quad i=1,2 \qquad (4.4.76)$$

The expressions in (4.4.74)-(4.4.76) represent a well-formulated linear programming problem; the solution of which is

$$u_p(0) = 0.4834 \quad u_q(0) = 0.0000$$

$$u_p(1) = 1.0000 \quad u_q(1) = 0.0000$$

By equation (4.4.72)-(4.4.73) these equations imply that

$$u(0) = 0.4834 \qquad (4.4.77)$$

$$u(1) = 1.0000 \qquad (4.4.78)$$

u(0) and u(1) given by equation (4.4.77) and (4.4.78) are the time-optimal controls that require minimum fuel. By comparing these values with those of (4.4.69) and (4.4.70), we see that they are the same. This is not true in general.

To check the above results the discrete-time system was simulated using Gaussian noise with statistics given by equations (4.4.31) and (4.4.32). Using different noise sequences, the system was simulated 500 times to determine the following sample means.

$$E(x(0)) = (10.0 \quad -11.99)^T$$

$$E(x(1)) = (2.58 \quad -4.14)^T$$

$$E(x(2)) = (0.344 \quad -0.873)^T$$

$$E(x_1^2(2)) = 0.69 \tag{4.4.79}$$

$$E(x_2^2(2)) = 1.00 \tag{4.4.80}$$

Equations (4.4.79) and (4.4.80) indicate that $E(x_i^2(2)) \leq 1$, $i=1,2$ as desired.

CHAPTER V

SUMMARY AND EXTENSIONS

In Section 5.1 some of the main results obtained in Chapters III and IV are briefly summarized. Certain possible extensions of these results are given in Section 5.2.

5.1 Summary

For the case when the target set is a hypersphere (Chapter III), the procedure for determining a sequence of open-loop time-optimal controls is most easily described by using Figures 3.2.1, 3.5.1 and 3.6.1. Depending on the parameters of the system and the initial state, we may or may not have a unique solution. In general, the solution is not unique if $N \leq n/m$ and is never unique if $N > n/m$. Because of the nonuniqueness of the solution, we have an opportunity to optimize the system according to another criterion. The criterion chosen here is to minimize the total energy required to drive the initial state to the target set. It is shown in Section 3.3 that except for the case when all the controls are zero the minimum energy sequence that drives the system to the boundary of the target also is the minimum energy sequence to the entire target. A method of determining the minimum energy seqence of controls is given in Section 3.3. The method is readily adaptable to computer solution, and the problem reduces to finding the roots of a polynomial of order less than or equal to 2n where n is the order of the system. It is shown in Section 3.4 that the time-optimal controller can be synthesized as a feedback controller. The sequence of optimal controls is proportional to the state of the system, plus a bias term proportional to the radius of the target. Using arguments similar to that for the open-loop system, the time-optimal control problem for a system with a single delay in the input is solved. In Section 3.6 a stochastic version of the original problem is solved. Since the controller is assumed to be open-loop, it is shown that the sequence of controls can be found in a manner similar to that of the deterministic system. It is also shown that for the undelayed deterministic system the target can also be reached in a number of samples equal to the smallest integer greater than or equal to n/m but that this need not be true for the stochastic system.

In Chapter IV the target set is changed to that of a hyperrectangle, and constraints on the amplitude of the controller are added. By means of a linear transformation it is shown that this problem reduces to finding the solution to a set of linear inequalities. By defining new variables, the original problem is reduced to a linear programming problem for which the method of solution is well-known. The procedure for determining a sequence of time-optimal controls begins by assuming N=1 and using Phase I of linear programming to determine if a solution exists for this value of N. If no solution exists, N is increased by 1 and the procedure repeated. Assuming a solution exists for some value of N, the theory of linear programming indicates whether the solution is unique. If it is not, the linear programming problem is reformulated with a different cost functional which is taken to be the total fuel required to reach a point contained in the target. By solving this new linear programming problem we determine the sequence of time-optimal controls that require minimum total fuel required to reach the target. Corresponding to this deterministic problem is the stochastic system described in Section 4.5. Because of the assumption that the controller be open-loop, this problem is reduced to a problem quite similar to that of the deterministic system.

5.2 Extensions

There are several extensions to the problems considered in Chapters III and IV which can be considered. These extensions are given here.

(1) Time-Varying Linear Systems

It is assumed that the system to be considered is described by the following difference equation.

$$x[(k+1)T] = C_{k+1,k}x(kT) + D_k u(kT) \qquad (5.2.1)$$

$$x(0) = x_o \qquad (5.2.2)$$

where $C_{k+1,k}$ has the properties of a transition matrix, that is,

$$C_{k,k} = I \text{ for all } k$$

$$C_{k,j}C_{j,i} = C_{k,i}$$

$$C_{k,j}^{-1} = C_{j,k}$$

These assumptions on $C_{i,k}$ are automatically satisfied if the state difference equation is derived from a linear system described by a time-varying differential equation. Such is the case in a sampled-data system.

The solution of equations (5.2.1) and (5.2.2) is

$$x(NT) = C_{N,0}x(0) + \sum_{i=0}^{N-1} C_{N,i+1}D_{i+1}u(iT)$$

Using the second property of $C_{j,k}$ given above, this can be written as

$$x(NT) = C_{N,0}\left[x(0) + \sum_{i=0}^{N-1} C_{0,i+1}D_{i+1}u(iT)\right]$$

$$= C_{N,0}\left[x(0)+(C_{0,1}D_1u(0)+C_{0,2}D_2u(T)+\ldots+C_{0,N}D_Nu[(N-1)T])\right]$$

$$= C_{N,0}\left[x(0)-\left(-C_{0,1}D_1,-C_{0,2}D_2,\ldots,-C_{0,N}D_N,\ldots,-C_{0,N}D_N\right)\right]\begin{bmatrix} u(0) \\ u(T) \\ \cdot \\ \cdot \\ \cdot \\ u[N-1)T] \end{bmatrix}$$

$$= C_{N,0}[x(0) - \underline{\overline{F}}_N U]$$

where

$$\underline{\overline{F}}_N = \left[-C_{0,1}D_1,-C_{0,2}D_2,\ldots,-C_{0,N}D_N\right] \qquad (5.2.3)$$

and

$$U = (u(0),u(T),\ldots,u[(N-1)T])^T$$

Then

$$x^T(NT)x(NT) = U^T\underline{\overline{F}}_N^T\Psi_N\underline{\overline{F}}_N U - 2x_o^T\Psi_N\underline{\overline{F}}_N U + x_o^T\Psi_N x_o$$

and

$$x^T(NT)x(NT) - R^2 = \underline{f}(U) = U^T\underline{Q}_N U - 2\underline{d}_N^T U + \underline{e}_N \qquad (5.2.4)$$

where

$$\Psi_N = (C_{N,0})^T C_{N,0} \qquad (5.2.5)$$

$$\underline{Q}_N = \underline{\overline{F}}_N^T\Psi_N\underline{\overline{F}}_N \qquad (5.2.6)$$

$$\underline{d}_N^T = x_o^T\Psi_N\underline{\overline{F}}_N \qquad (5.2.7)$$

$$\underline{e}_N = x_o^T\Psi_N x_o - R^2 \qquad (5.2.8)$$

Equations (5.2.3)-(5.2.8) are similar to equations (3.2.11), (3.2.14), (3.2.7), (3.2.10), (3.2.12) and (3.2.13). The only significant difference is that $\underline{\overline{F}}_N$ replaces $\overline{F}_N$ in the time-invariant system. Thus the steps used in solving the optimal control problem for the time varying system are the same as those given in Chapter III. The two assumptions given in Section 3.1 for the time-invariant system are replaced by the following assumptions.

(i) It is assumed that

$$\text{rank } (\underline{\overline{F}}_N) = \text{rank } (D_1, C_{1,2}D_2, \ldots, C_{1,n}D_n) = n$$

(ii) It is assumed that

$$\text{rank}(D_1, C_{1,2}D_2, \ldots, C_{1,N}D_N) = \text{maximum for}$$

all $N > 0$

Assumption (i) is the same as the assumption of complete controllability of a time varying system [SOR1]. With these assumptions the optimal control problem for the time varying system reduces to replacing C by $C_{k,k-1}$ and D by D_k in the time-invariant system of Chapter III.

With regard to the problem of Chapter IV (hyper-rectangular target set), no assumptions of controllability were made. In this case we can substitute $C_{k,k-1}$ for C and D_k for D, and all the results given in Chapter IV can be used.

(2) Cost functional of the form $J = \sum_{i=0}^{N-1} u^T(iT)Su(iT)$ where

S is positive definite and symmetric.

Section 3.3 was concerned with minimizing the total energy required to reach the boundary of the hyperspherical target set. In some case it may be desirable to assign a heavier cost to certain components of u(kT). One possible way to do this is to express the cost in the form given by J above. To solve this problem it is noted by Theorem 2.1.9 that S is positive definite if and only if there exists a nonsingular matrix L such that $s = L^T L$. Therefore,

$$J = \sum_{k=0}^{N-1} Y^T(kT)y(kT) \qquad (5.2.9)$$

where y(kT) = Lu(kT). We can rewrite the state equation (5.2.1) as

$$y[(k+1)T] = Cy(kT) + D'y(kT) \qquad (5.2.10)$$

where $D' = DL^{-1}$. We then have the problem of minimizing J given by equation (5.2.9) subject to equation (5.2.10). This is the same type of problem as that solved in Section 3.3 except that D' replaces D, and y(kT) replaces u(kT). Once y(kT) is determined, we can find u(kT) by the equation $u(kT) = L^{-1}y(kT)$.

(3) Minimum Energy Solution for Hyperrectangular Target Set.

In Chapter IV it was shown that minimization of the fuel when the solution to the time-optimal control problem was not unique led to a linear programming problem. In equation (4.3.3) we let

$$u_i(kT) = u_{i,p}(kT) - u_{i,q}(kT) \quad i=1,2,\ldots,n \quad k=0,1,\ldots,N-1$$

If we wish to minimize the energy, the performance index could be taken to be

$$J = \sum_{k=0}^{N-1} \sum_{i=1}^{m} u_i(kT)u_i(kT)$$

$$= \sum_{k-0}^{N-1} \sum_{i=1}^{m} \left[u_{i,p}(kT)-u_{i,q}(kT)\right] \left[u_{i,p}(kT)-u_{i,q}(kT)\right]$$

$$= \sum_{k=0}^{N-1} \sum_{i=1}^{m} \left[u_{i,p}(kT)u_{i,q}(kT)\right] \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} u_{i,p}(kT) \\ u_{i,q}(kT) \end{bmatrix} \qquad (5.2.11)$$

The performance index given by equation (5.2.11) and the inequality constraints given by (4.3.4)-(4.3.5) represent a well-formulated quadratic programming problem; the method of solution of which is well-known [BOO1], [CAN1], [GUE1].

(4) Moving target Set

It was assumed in Chapter III that the target remained centered at the origin. The modification for a moving target set is straight forward. Let the target set be described by

$$\{x(NT): \quad (x(NT) - x_c(NT))^T(x(NT) - x_c(NT)) \leqslant R^2\}$$

Then by the same argument that was used to obtain equation (3.2.8), we have

$$(x(NT) - x_c(NT))^T(x(NT) - x_c(NT)) = U^TQ_NU - 2(d_N^T$$

$$+ x_c^T(NT)C^N\overline{F}_N)U + x_o^T\Psi_Nx_o - 2x_c^T(NT)C^Nx_o$$

$$+ x_c^T(NT)x_c(NT)$$

where Q_N, $\bar{\bar{F}}_N$ and d_N are defined by equations (3.2.10)-(3.2.12). By letting $\underline{d}_N^T = d_N^T + x_c^T(NT)C^N\bar{\bar{F}}_N$ and $\underline{e}_N = x_o^T\Psi_N x_o - 2x_c^T(NT)C^N x_o + x_c^T(NT)x_c(NT)$ and restricting the problem to finding the minimum number of samples to reach the boundary of the target, we have

$$f'(U) \equiv (x(NT)-x_c(NT))^T(x(NT)-x_c(NT))-R^2 = U^TQ_NU - 2\underline{d}_N^TU + \underline{e}_N = 0 \qquad (5.2.12)$$

Equations (5.2.12) is similar to equation (3.2.14) except that $\underline{d}_N$ replaces d_N and $\underline{e}_N$ replaces e_N. We can then repeat the same arguments and derive theorems analogous to those of Section 3.2 after replacing $\underline{d}_N$ by d_N and $\underline{e}_N$ by e_N. The same type of argument can be applied to the problem considered in Chapter IV.

REFERENCES

REFERENCES

[AOK1] M. Aoki, Optimization of Stochastic Systems, Academic Press, New York, 1967.

[ATH1] M. Athans, P. Falb, Optimal Control, An Introduction to the Theory and Its Applications, McGraw-Hill Book Company, New York, 1966.

[AYR1] F. Ayres, Theory and Problems of Matrices, Schaum Publishing Co., New York, 1962.

[BER1] J. Bertram, P. Sarachik, "On Optimal Computer Control", Proc. First International Congress of Automatic Control, Butterworth, pp. 419-422, 1960.

[BOO1] J. C. Boot, Quadratic Programming, North-Holland Publishing Co., Amsterdam, 1964.

[CAD1] J. A. Cadzow, H. R. Martens, Discrete-Time and Computer Control Systems, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1970.

[CAD2] J. A. Cadzow, "An Extension of the Minimum Norm Controller for Discrete Systems", IEEE Trans. on Automatic Control, Vol. AC-12, No. 2, pp. 202-203, April, 1967.

[CAD3] J. A. Cadzow, "Nilpotency Property of the Discrete Regulator", IEEE Trans. on Automatic Control, pp. 734-735, December, 1968.

[CAN1] M. D. Canon, J. H. Eaton, "A New Algorithm for a Class of Quadratic Programming Problems with Application to Control", J. SIAM Control, Vol. 4, No. 1, pp. 34-45, 1966.

[CUL1] C. Cullen, Matrices and Linear Transformations, Addison-Wesley Publishing Co., Reading, Massachusetts, 1966.

[DER1] P. Derusso, J. Rob, C. Close, State Variables for Engineers, John Wiley & Sons, Inc., New York, 1967.

[DES1] C. A. Desoer, J. Wing, "The Minimal Time Regulator Problem for Linear Sampled-Data Systems: General Theory", J. Franklin Inst., Vol. 272, No. 9, pp. 208-228, 1961.

[DEU1] R. Deutsch, System Analysis Techniques, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1969.

[DEU2] R. Deutsch, Estimation Theory, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1965.

[FAR1] J. B. Farison, F. Fu, "The Matrix Properties of Minimum-Time Discrete Linear Regulator Control", IEEE Trans. on Automatic Control, pp. 390-391, June, 1970.

[FEG1] K. A. Fegley, M. I. Hsu, "Optimum Discrete Control by Linear Programming", IEEE Trans. Auto. Control, Vol. AC-10, No. 1, pp. 114-115, January, 1965.

[FRA1] J. S. Frame, "Matrix Functions and Applications", IEEE Spectrum, pp. 208-220, 102-108, 100-109, 123-131, March-July, 1964.

[FRA2] J. S. Frame, Matrix Theory and Linear Algebra with Applications", (Mathematics 831 Class Notes), Department of Mathematics, Michigan State University.

[GAN1] F. R. Gantmacher, The Theory of Matrices, Chelsea Publishing Co., Vol. 1, New York, 1960.

[GRA1] J. J. Grainger, K. G. Pandy, "Time-Optimal Control of Sampled-Data Systems", Proc. of IEEE, pp. 1295-1297, August, 1970.

[GRE1] T. Greville, "The Pseudoinverse of a Rectangular or Singular Matrix and its Applications to the Solution of Systems of Linear Equations", SIAM Review, Vol. 1, No. 1, pp. 38-43, January, 1959.

[GRE2] T. Greville, "Some Applications of the Pseudoinverse of a Matrix", SIAM Review, Vol. 2, No. 1, pp. 15-22, January, 1960.

[GUE1] R. L. Gue, M. E. Thomas, Mathematical Methods in Operations Research, The Macmillan Company, London, 1968.

[HAD1] G. Hadley, Linear Programming, Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, 1962.

[HIL1] F. B. Hilderbrand, Methods of Applied Mathematics, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1965.

[HO1] Y. C. Ho, "Solution Space Approach to Optimal Control Problems", Trans. ASME, Vol. 83, pp. 53-58, March, 1961.

[IBM1] IBM Application Program, 1130 Scientific Subroutine Package (1130-CM-02X), Programmer's Manual, International Business Machine Corp., White Plains, New York, 1968.

[JAM1] M. James, G. Smith, J. Wolford, Applied Numerical Methods for Digital Computation with FORTRAN, International Textbook Company, Scranton, Pennsylvania, 1967.

[KAL1] R. E. Kalman, "Optimal Non-Linear Control of Saturating Servo-Mechanisms by Intermittent Action", IRE Wescon Record, Part 4, pp. 130-135, 1957.

[KAL2] R. E. Kalman, "On the General Theory of Control Systems", Proc. First International Congress of Automatic Control, pp. 481-492, Moscow, 1960.

[KOE1] H. E. Koenig, Y. Tokad, H. K. Kesavan, Analysis of Discrete Physical Systems, McGraw-Hill Book Company, New York, 1967.

[KOP1] R. W. Koepcke, "On the Control of Linear Systems with Pure Time-Delay", J. Basic Engineering, pp. 74-80, March, 1965.

[KUO1] S. S. Kuo, Numerical Methods and Computers, Addison-Wesley Publishing Co., Reading, Massachusetts, 1965.

[KUO2] B. C. Kuo, Discrete-Data Control Systems, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970.

[KUR1] F. Kurzweil, "The Control of Multivariable Processes in the Presence of Pure Transport Delays', IEEE Trans. on Automatic Control, Vol. AC-8, pp. 27-34, January, 1963.

[PED1] D. Pedoe, A Geometric Introduction to Linear Algebra, John Wiley & Sons, Inc., New York, 1963.

[PEN1] R. Penrose, "A Generalized Inverse of Matrices", Proc. Cambridge Philos. Soc., Vol. 51, pp. 406-413, 1955.

[PLA1] J. B. Plant, M. Athans, "An Iterative Technique for the Computation of Time Optimal Controls", Paper 13D, IFAC Congress, London, 1966.

[PLA2] J. B. Plant, "An Iterative Procedure for the Computation of Fixed-Time Fuel-Optimal Controls", IEEE Trans. on Automatic Control, Vol. AC-11, No. 4, pp. 652-660, October, 1966.

[RAL1] A. Ralston, H. Wilf, Mathematical Methods for Digital Computers, Chpt. 7, John Wiley & Sons, New York, 1962.

[SAR1] P. E. Sarachik, G. M. Kranc, "Optimal Control of Discrete Systems with Constrained Inputs", J. Franklin Institute, Vol. 277, No. 3, pp. 237-255, March, 1964.

[SOR1] H. W. Sorenson, "Controllability and Observability of Linear, Stochastic, Time-Discrete Control Systems", Advances in Control Systems, Vol. 6, pp. 95-158, 1968.

[TOR1] H. C. Torng, "Optimization of Discrete Control Systems Through Linear Programming", J. Franklin Institute, Vol. 278, No. 1, pp. 28-44, July, 1964.

[TOU1] J. T. Tou, "Synthesis of Discrete Systems Subject to Control-Signal Saturation", J. Franklin Institute, Vol. 277, No. 5, pp. 401-413, May, 1964.

[TOU2] J. T. Tou, Optimal Design of Digital Control Systems, Academic Press, New York, 1963.

[TOU3] J. T. Tou, Modern Control Theory, McGraw-Hill Co., New York, 1964.

[TUE1] W. G. Tuel, Jr., "Comments on 'The Matrix Properties of Minimum-Time Discrete Linear Regulator Control'", IEEE Trans. on Automatic Control, Vol. AC-16, No. 1, p. 105, February, 1971.

[ZAD1] L. Zadeh, C. Desoer, Linear System Theory, McGraw-Hill Book Co., Inc., New York, 1963.

[ZAD2] L. Zadeh, B. H. Whalen, "On Optimal Control and Linear Programming", IRE Trans. Automatric Control, Vol. AC-7, pp. 45-46, 1962.

APPENDIX

APPENDIX

In this appendix the conversion from continuous to discrete-time systems is given for a system with white noise input.

Let the system be described by the following equations

$$\dot{x} = A(t)x + B(t)u + v(t) \tag{1}$$

$$x(0) = x_o$$

$$E(v(t)) = 0$$

$$E(v(s)v^T(\tau)) = R\delta(s - \tau)$$

where $\delta(t)$ is the dirac delta function. With $\Phi(t,\tau)$ as the transition matrix, the solution of (1) is

$$x(t) = \Phi(t,t_o)x(t_o) + \int_{t_o}^{t} \Phi(t,\tau)B(\tau)u(\tau)d\tau + \int_{t_o}^{t} \Phi(t,\tau)v(\tau)d\tau \tag{2}$$

Setting $t = t_{k+1}$ and $t_o = t_k$, equation (2) becomes

$$x(t_{k+1}) = \Phi(t_{k+1},t_k)x(t_k) + \int_{t_k}^{t_{k+1}} \Phi(t_{k+1},\tau)B(\tau)u(\tau)d\tau + \int_{t_k}^{t_{k+1}} \Phi(t_{k+1},\tau)v(\tau)d\tau \tag{3}$$

Since it is assumed that the system is of the sampled-data type, $u(\tau) = u(t_k)$ for $t_k \leq \tau < t_{k+1}$. Equation (3) then becomes

$$x_{k+1} = \Phi_k x_k + D_k u_k + w_k \qquad k=0,1,\ldots,N-1 \tag{4}$$

where

$$x_k = x(t_k) \qquad \text{and } \Phi_k = \Phi(t_{k+1},t_k)$$

$$D_k = \int_{t_k}^{t_{k+1}} \Phi(t_{k+1},\tau)B(\tau)d\tau \qquad w_k = \int_{t_k}^{t_{k+1}} \Phi(t_{k+1},\tau)v(\tau)d\tau$$

Also,

$$E(w_k w_j^T) = E\left\{\int_{t_k}^{t_{k+1}} \Phi(t_{k+1},\tau)v(\tau)d\tau \int_{t_j}^{t_{j+1}} v^T(\gamma)\Phi^T(t_{j+1},\gamma)d\gamma\right\}$$

$$= \int_{t_k}^{t_{k+1}} \int_{t_j}^{t_{j+1}} \Phi(t_{k+1},\tau)R\delta(\tau-\gamma)\Phi^T(t_{j+1},\gamma)d\tau d\gamma$$

$$= \int_{t_k}^{t_{k+1}} \Phi(t_{k+1},\tau)R\Phi^T(t_{k+1},\tau)d\tau\delta_{k,j}$$

Therefore,

$$E(w_k w_j^T) = Q_k \delta_{k,j} \tag{5}$$

where

$$Q_k = \int_{t_k}^{t_{k+1}} \Phi(t_{k+1},\tau) R \Phi^T(t_{k+1},\tau) d\tau \qquad \delta_{k,j} = \begin{cases} 1 \text{ if } k=j \\ 0 \text{ if } k \neq j \end{cases}$$

and

$$E(w_k) = 0 \tag{6}$$

Equations (4), (5) and (6) represent the sampled-data system corresponding to the continuous time system given in (1). If A(t) = A and B(t) = B, then Φ_k and D_k also become constants, that is, $\Phi_k = C$ and $D_k = D$. Equation (4) then becomes

$$x_{k+1} = Cx_k + Du_k + w_k$$