APPLICATIONS OF FUNCTIONAL ANALYSIS TO TIME OPTIMAL CONTROL OF LINEAR SYSTEMS WITH OUTPUT CONSTRAINTS

by

MICHAEL BERNARD SHILMAN

A dissertation submitted to the Graduate Faculty in Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1970

This manuscript has been read and accepted for the Graduate Faculty in Engineering in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

4/6/1970
date

Chairman of Examining Committee

6 April 1970
date

Executive Officer

Prof. Stanley Katz

Prof. Donald Goldfarb

Prof. Ralph Mekel

Supervisory Committee

The City University of New York

ACKNOWLEDGMENTS

I wish to thank my advisor, Professor George Kranc, for suggesting the course of the research pursued in this dissertation. His continuing advice and encouragement were particularly beneficial.

I am also indebted to the members of my guidance committee, Professors Stanley Katz, Ralph Mekel, and Donald Goldfarb, for their critical comments and suggestions which enriched this work considerably.

Special thanks are due to Dean Egon Brenner for his active interest in the progress of this work and for his guidance and support which were necessary ingredients for its completion.

For the careful typing of this thesis, I am indebted to the Grumman Aerospace Corporation for permission to use their facilities and in particular, Dr. Richard Kopp, Mr. James Compton, and Mrs. Mildred Sudwischer.

Last, but not least, I wish to express my deepest gratitude to my wife Marilyn for her continuing patience and understanding without which this work would not have been possible.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

ABSTRACT

This research is concerned with time optimal control of linear systems with control and output constraints. Although this class of problems has been treated by other researchers the approach taken in this dissertation, which employs results from functional analysis, yields an efficient means for obtaining numerical solutions. Approximations for the optimal solution are obtained directly from the results of a finite dimensional minimization of a convex function. These approximations may be successively improved and furthermore are shown to converge in a specified sense to an optimal solution. Upper and lower bounds for the optimal time are also obtained. A method is presented which, in many cases, reduces the dimension of the minimization necessary to obtain an acceptable approximation.

It is also shown how these methods are modified so that discrete systems may be treated. For this case, the optimal solution is obtained.

CHAPTER 1. INTRODUCTION

The classical time-optimal control problem may be stated as follows: Find the control variables, constrained in some manner, which bring the output of the controlled plant from some initial value to a desired final one in the shortest time.[1] This problem received much attention in the past two decades and presently there are a variety of methods one can use to obtain its solution.

In this decade much research has been devoted to the time-optimal control problem with constraints on the output variables in addition to those on the control variables. A precise statement of this problem will be given in the next section. This problem may arise, for example, in the control of chemical processes which require constraints on pressures and temperatures at critical points in a chemical plant or in flight control problems where an aircraft may be required to stay within a specified flight envelope to prevent excessive heating rates. It is interesting to note that the bounded output problem can arise even though the motion of a controlled object itself is not restricted.[2] For example, if one wants to change the course of a ship as rapidly as possible by positioning its rudder then one must take into account not only the limit on the displace-

ment of the rudder but also the limit on its rate of change due to inertia. One way to remove this multiple limit on the movements on the control is to consider the velocity of the rudder as a control of the original system augmented by an integrator as shown in Figure 1-1.

The rudder displacement is now considered as an output variable of the augmented system and the original problem with multiple saturation limits reduces to one of time optimal control with control and output constraints.

We remark that the problem of time optimal control with control and output constraints subsumes the problem of time optimal control with control and state constraints (which is the problem generally considered in the literature) as can be seen from Eq. (1-1) with $C(t)$ taken as the identity matrix.

1.1 Mathematical Statement of the Problem

The linear system to be controlled, which we shall call the plant, can be described by the superposition integral

$$x(t) = \underline{x}_o(t) + \int_{t_0}^{t} H(t,\tau)\underline{u}(\tau)\,d\tau \qquad (1\text{-}1)$$

where $\underline{x}(t)$ is an m-vector representing the output of the system, $\underline{u}(t)$ is a p-vector representing the system

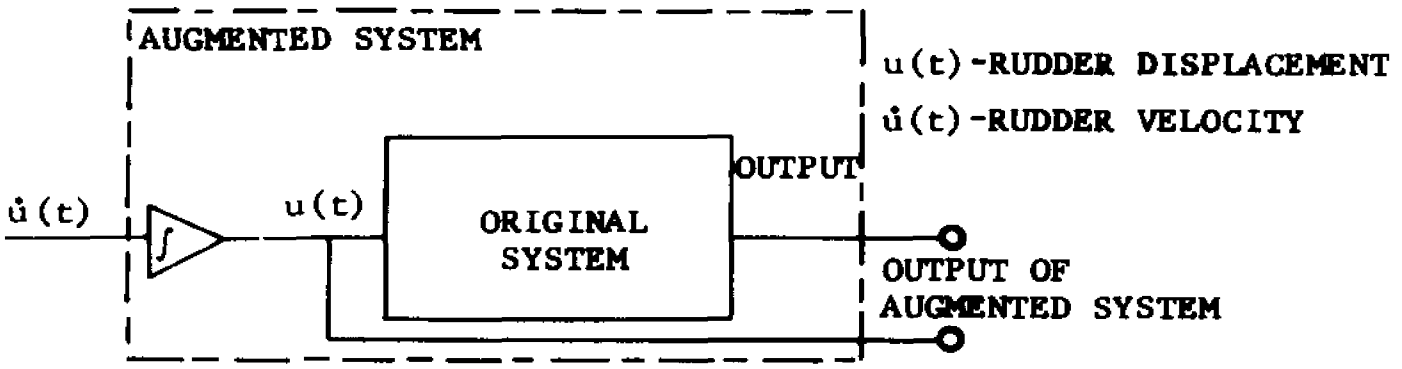


Fig. 1-1 Formulation of a Problem with Multiple Saturation Limits as a Problem of Control with Output Constraints

input, $\underline{x}_o(t)$ is an m-vector expressing the effect of initial conditions in the plant at $t = t_o$ and is assumed to be known and continuous and $H(t,\tau)$ is an $m \times p$ matrix whose elements are bounded piecewise continuous functions of their arguments.

Equation (1-1) results if $\underline{x}(t)$ is the output of a system satisfying the linear vector differential equation[3]

$$\begin{aligned} \dot{\underline{w}}(t) &= A(t)\underline{w}(t) + B(t)\underline{u}(t) \\ \underline{x}(t) &= C(t)\underline{w}(t) \end{aligned} \tag{1-2}$$

where $\underline{w}(t)$ is an n-vector representing the state of the system and $A(t)$, $B(t)$, and $C(t)$ are matrices of appropriate dimensions which are piecewise continuous functions of t. In this case $\underline{x}_o(t)$ and $H(t,\tau)$ are given by

$$\underline{x}_o(t) = C(t)\varphi(t,t_o)\underline{x}(t_o) \quad , \quad H(t,\tau) = C(t)\varphi(t,\tau)B(\tau)$$

and $\varphi(t,\tau)$ is the state transition matrix of (1-2).

The class of admissible control functions consists of all functions in some Banach space satisfying

$$\|\underline{u}\| \leq C_1$$

where C_1 is some fixed positive constant and the Banach space and its norm are determined by physical considerations. For a single input system the Banach space will be taken as

the function space $L_p[t_o,T]$ with the norm of a function $u(t)$ in this space given by

$$\|u\| = \left(\int_{t_0}^{T} |u(t)|^p dt\right)^{1/p} \qquad 1 < p \leq \infty$$

where T is the terminal time.

The problem is to steer the output $\underline{x}(t)$ from some given initial point $\underline{x}_o(t_o)$ to a desired terminal point $\underline{x}_d$ by an admissible controller in the least time subject to amplitude constraints on some or all of the components of the output vector, i.e.,

$$\begin{aligned} |x^{i_1}(t)| &< C'_{i_1} \\ &\vdots \quad\; \vdots \\ |x^{i_{r_1}}(t)| &\leq C'_{i_{r_1}} \end{aligned} \qquad \begin{aligned} &1 \leq i_1 < \ldots < i_{r_1} \leq m \qquad r_1 \leq m \\ &C'_{i_1}, \ \ldots \ C'_{i_{r_1}} > 0 \end{aligned}$$

1.2 Historical Background of the Development of Problems with State Constraints

The early investigators were primarily concerned with the theoretical aspects of the problem and developed necessary conditions which an optimal pair (the control and its trajectory) must satisfy. Necessary conditions have been derived using both the calculus of variations[4,5,9,10] and by modifications of the maximum principle.[2,6,7,8,11] If we

denote by Δ the admissible region in the state space, then the basic assumption is that if $u^*(t)$, $0 \leq t \leq t^*$, is an optimal control and $x^*(t)$ is the corresponding trajectory joining $\underline{x}_o(t_o)$ and $\underline{x}_d$ both interior to Δ with $\underline{x}^*(t) \subset \Delta$ for each t on $0 \leq t \leq t^*$, then there exists a finite subdivision into subintervals

$$0 = t_o < t_1 \leq t_2 < t_3 \leq t_4 \ldots < t_{2k+1} = t^*$$

such that

$$\underline{x}^*(t) \subset \text{interior } \Delta \text{ on } t_i < t < t_{i+1} \quad \text{even } i$$

$$\underline{x}^*(t) \subset \text{boundary } \Delta \text{ on } t_i \leq t \leq t_{i+1} \quad \text{odd } i$$

On each interior segment (even i) the trajectory $\underline{x}^*(t)$ joins $\underline{x}^*(t_i)$ to $\underline{x}^*(t_{i+1})$ without the state constraint and thus satisfies the usual maximum principle. On the boundary segments (odd i) it is found that the trajectory satisfies a modified maximum principle. Since, by assumption, the trajectory is on the boundary in $[t_i, t_{i+1}]$, i odd, there exists no other admissible control which transfers $\underline{x}^*(t_i)$ to $\underline{x}^*(t_{i+1})$ in less time than $u^*(t)$ <u>and</u> has its corresponding trajectory on the boundary in $[t_i, t_{i+1}]$, otherwise this would contradict the optimality of the pair $u^*(t)$, $x^*(t)$. Therefore the modification comes about by

writing the system equations in terms of suitable coordinates on the boundary (since the trajectory must lie on the boundary in this interval), forming the Hamiltonian in the usual way for this reduced set of equations and maximizing the Hamiltonian with respect to all admissible controls having their corresponding trajectories on the boundary.

This still leaves unresolved the determination of the times t_i and if this were not bad enough, we might add that there are certain junction conditions that must be satisfied when an interior segment meets a boundary segment.[6]

Essentially the same results are obtained from the calculus of variations where the modifications occur in the Euler-Lagrange equations during the period when the trajectory is on the boundary.

Recently, researchers have been interested in trying to generate a computational solution for the bounded state problem. Denham and Bryson[12] have devised a direct method of solution for this problem by assuming a nominal solution which need not satisfy the state or terminal constraints and then incrementing the nominal control so as to both reduce the transfer time and satisfy the constraints more closely. This process of incrementing is continued until suitable error bounds are met. However, if there is a simultaneous control and state constraint or if the trajectory is

on the boundary for more than one segment, the computational time is greatly increased.

There are two other types of approaches used to obtain computational solutions for this problem. One is an approach using penalty functions, the other is to transform the problem into a discrete one by assuming a piecewise constant control law.

The penalty method in the calculus of variations was introduced by Courant.[13] It was applied to optimal control problems by Kelley,[14] McGill,[15] Kirin,[16] Lee,[17] and Lasdon et al.,[18] and rigorous mathematical justifications for the procedure were given by Russell,[19] and Okamura.[20] Chang[2,7] used this method to derive necessary conditions for an optimal control and its corresponding trajectory.

The basic idea behind the penalty method is the following:[19] instead of attempting a direct solution of the constrained state optimal control problem, an unconstrained problem is considered wherein the original cost functional is augmented by a nonnegative penalty function which sharply increases the cost associated with the trajectories which violate the state constraints. By using sequences of cost functionals involving more and more severe penalty functions, it is to be expected in many cases that the desired constrained state solution of the original optimization problem

may be approximated to any desired degree of accuracy by solutions to these unconstrained problems.

For the time-optimal bounded state problem the performance functional we wish to minimize is

$$J = \int_{t_0}^{t} dt$$

where the corresponding trajectories must lie in Δ (Δ is the admissible region in the state space). Let $F(x)$ be a continuous piecewise differentiable function such that

$$F(x) = 0 \qquad x \in \Delta$$

$$F(x) > 0 \qquad x \notin \Delta$$

The problem is now put into penalty function form by minimizing the augmented performance functional,

$$J_a = \int_{t_0}^{t} \left[1 + KF\left(x(t)\right)\right] dt$$

where K is a suitably large positive constant, and ignoring the actual state constraint. By increasing K the penalty for leaving Δ becomes more severe and it appears reasonable that if K is large enough the optimal path is completely in Δ.

The above is an example of an exterior penalty constraint. We can also construct interior penalty constraints, that is, constraints for which the function F is defined only on Δ and is 0 in Δ except for x sufficiently close to its boundary. We do not discuss this type of penalty for in the class of problems dealt with in this research (problems with simultaneous constraints on the control and state variables), the generation of a starting point needed for this method is, in many cases, as difficult as solving the original problem.

The other approach of discretizing the system has been employed by Ho and Brentani,[21] Nagata et al.,[22] and Fath.[23] The basic assumption is that the control is piecewise constant over each subinterval of a suitable subdivision of some fixed interval $[t_o, t_1]$. A performance functional is chosen such that by maximizing (or minimizing) this function, which is now a function of the finite number of variables $u\left(\frac{k(t_1 - t_o)}{N}\right)$, $k = 0, 1, \ldots, N - 1$ we can deduce whether or not an admissible control exists which transfers the initial state to the terminal state in the interval $[t_o, t_1]$. The maximization is performed subject to the constraints on the state variables which is carried out by Fath using linear programming techniques and by Ho and Brentani using a modified gradient procedure. A search is then made for

various values of t_1 to find the first value of time for which an admissible transfer can be made.

The techniques of functional analysis were first applied to time optimal bounded state problems in 1962-63 when Gabasov and Kirillova[24,25] obtained solutions to the regulator problem with magnitude constraints on the control for a class of systems known as linked discrete systems. In 1964 they used the properties of adjoint operators to obtain solutions for continuous time systems with an output constraint at discrete instants of time.[26] Recently, Katz and Kranc[27] exploited the "multinorm" formulation developed by Sarachik and Kranc[28] to obtain results for the continuous time problem treated by Gabosov and Kirillova with more general constraints.

1.3 Scope of the Dissertation

In this thesis, we utilize the product space approach of Katz and Kranc to obtain a number of different methods of obtaining an approximation for the time optimal problem with amplitude constraints on the outputs at every instant of time during the transition period. In the next chapter, this approach is presented in a form which is convenient for our purposes. We also present some elements of real variable theory which are necessary for Chapter 5.

Presented in Chapter 3 is the discrete point approximation to the problem. The basic idea is to replace the original problem by a problem in which the output constraint is enforced only at a finite number of discrete instants of time, solve this latter problem exactly and use its solution as an approximate solution for the original problem. The main result is Theorem 3.2 which states that by using enough points we can obtain as good an approximation to the original problem as desired in a sense made precise in that theorem.

One problem that may be encountered when using the discrete point approximation is that, in some cases, the approach may lead to performing minimizations of large dimension. If the dimension is large enough this could lead to computational difficulties. In Chapter 4 we present the solution to the terminal control problem with output constraints and show how it may be used to circumvent the difficulty of performing large dimensional minimizations.

A different means of obtaining an approximation to the time optimal bounded output problem is demonstrated in Chapter 5. It is shown that at the optimal time there exists a function of finite variation $\psi^*(t)$ which minimizes (5-31). Moreover, the function related to $\psi^*(t)$ by (5-35) is an optimal control. We discuss a method of approximating $\psi^*(t)$ and thereby obtain an approximation to the optimal control.

Whereas the discrete point approximation can be considered a means of approximating the problem the present approach differs in the sense that we are obtaining an approximation to the optimal control of the original problem.

In Chapter 6 we consider the time optimal bounded output problem for discrete systems. Although the theoretical development is essentially the same as for continuous time systems, certain difficulties arise which we can manage to avoid in the continuous case. Two schemes are presented for overcoming these difficulties.

A general discussion of the numerical procedures used in the preceding chapters is given in Chapter 7. In addition, we prove the applicability of minimization schemes which require gradient calculations to the minimization problems considered in this thesis.

CHAPTER 2. MATHEMATICAL BACKGROUND

2.1 Plants with Bounded Outputs

This chapter contains the mathematical background upon which most of the work in the succeeding chapters is based. Much of this material can be found in the paper by Katz and Kranc[27] but is included here for completeness.

We begin by considering a specific problem and the mathematical concepts are introduced as needed in developing its solution. It is emphasized that the following development applies only to linear systems, although the systems may be time-varying.

The linear system to be controlled, which we shall call the plant, can be described by the following input-output relations

$$x^1(t) = x_o^1(t) + \int_0^t h_1(t,\tau)u(\tau)d\tau$$

$$x^2(t) = x_o^2(t) + \int_0^t h_2(t,\tau)u(\tau)d\tau \qquad (2\text{-}1)$$

where $x^1(t)$, $x^2(t)$ are outputs of the plant which are to be controlled by the single variable $u(t)$ - the control. The terms $x_o^1(t)$ and $x_o^2(t)$ express the effect on $x^1(t)$

and $x^2(t)$ respectively, of initial conditions in the plant at $t = 0$ and are assumed to be known and continuous. The initial time is assumed to be $t = 0$ without loss of generality. The functions $h_1(t,\tau)$ and $h_2(t,\tau)$ are assumed to be piecewise continuous functions of τ on the interval $[0,t]$ and equal 0 for $\tau > t$ for all finite values of t.

The problem considered is that of finding the smallest T, called the optimal time, and the corresponding control function $u(t)$, called the optimal control, such that

$$x^1(T) = x_o^1(T) + \int_0^T h_1(T,\tau)u(\tau)d\tau = x_d^1 \qquad (2\text{-}2)$$

where x_d^1 is some desired final value. We have a constraint on the control $u(t)$ which we take in the form

$$\left(\int_0^T |u(t)|^p dt\right)^{1/p} \leq C_1 \qquad (2\text{-}3)$$

where C_1 is a positive constant and p is an integer satisfying $1 < p \leq \infty$. For $p = \infty$, this represents a constraint on the essential upper bound of $u(t)$.[29]

In addition, we have a constraint on the output $x^2(t)$ at n instants of time, $t_i = r_i T$, $i = 1, \ldots, n$ where r_i is some real number, $0 \leq r_i \leq 1$

$$|x^2(t_i)| = |x_o^2(t_i) + \int_0^{t_i} h_2(t_i,\tau)u(\tau)d\tau| \leq C_2 \qquad (2\text{-}4)$$

where C_2 is some positive constant.

This problem can be reformulated to postpone the determination of the smallest T by tentatively fixing T and asking for the control function $u(t)$ which makes

$$\max \left\{ \frac{1}{C_1} \left(\int_0^T |u(t)|^p dt \right)^{1/p} , \frac{1}{C_2} \max_i |x^2(t_2)| \right\} = \text{minimum} \qquad (2\text{-}5)$$

while keeping

$$\int_0^T h_1(T,\tau)u(\tau)d\tau = x_d^1 - x_o^1(T) \qquad (2\text{-}6)$$

The smallest T for which the minimum in (2-5) is unity then solves the original problem with constraints (2-2 - 2-4). If there is no T for which the minimum in (2-5) is unity then the problem posed has no solution.

As will be seen, the technique developed for the solution to this problem can be directly extended to handle the case of multiple terminal and/or output constraints.

Our approach involves the construction of a particular Banach space (complete normed linear space) such that the

variational problem (2-5 - 2-6) may be viewed in abstract terms as one of finding a bounded linear functional of minimum norm which maps certain elements of the space into given fixed scalars. This abstract problem which is sometimes known as the L-problem in the theory of moments[30] is solved in two stages. The first stage is an existence proof for the desired functional which employs the Hahn-Banach theorem and the second stage consists of finding the actual form of the linear functional (which directly yields the optimal control) from the equality conditions for Holder's inequality.

2.2 Product Spaces and Functionals

We now proceed with the construction of the Banach space mentioned in the preceding section. The parameter q conjugate to p is defined by

$$\frac{1}{p} + \frac{1}{q} = 1 \tag{2-7}$$

and for functions $y(t)$ in $L_q[0,T]$, where T is temporarily fixed, and n-tuples $\underline{a} = [a_1, a_2, \ldots, a_n]$ in $\ell_1^{(n)}$, define composite vectors $\bar{y}$ by pairing a function $y(t)$ with an n-tuple $\underline{a}$:

$$\bar{y} = \left\{ y(t), \underline{a} \right\} \tag{2-8}$$

Using the definitions of the norm of a function in $L_q[0,T]$

$$\|y\|_q = \left(\int_0^T |y(t)|^q dt\right)^{1/q} \tag{2-9}$$

and the norm of a vector in $\ell_1^{(n)}$

$$\|\underline{a}\|_{\ell_1^{(n)}} = \sum_{i=1}^{n} |a_i| \tag{2-10}$$

a norm $(\| \|_{\bar{L}_1})$ can be introduced for composite vectors $\bar{y}$ of the form (2-8) by defining

$$\|\bar{y}\|_{\bar{L}_1} = c_1\|y\|_q + c_2\|\underline{a}\|_{\ell_1^{(n)}} \tag{2-11}$$

It can be shown that with this norm (and of course defining the operations of addition and scalar multiplication in the natural way) that the composite vectors $\bar{y}$ form a Banach space (this follows from the fact that $L_q[0,T]$ and $\ell_1^{(n)}$ are Banach spaces, $\bar{L}_1$ is the product space of $L_q[0,T]$ and $\ell_1^{(n)}$, and the product of a finite number of Banach spaces is a Banach space for a suitably constructed norm).[31]

Also, the following can be shown:

1. Any bounded linear functional f on $\bar{L}_1$, has a representation

$$f(\bar{y}) = \int_0^T y(t)u(t)dt + \sum_{i=1}^{n} a_i b_i \qquad (2\text{-}12)$$

with $u(t)$ in $L_p[0,T]$ and b_i, $i = 1, \ldots, n$, finite scalars.

2. The norm of the functional f of (2-12) defined as

$$\|f\| = \sup_{\bar{y} \neq 0} \frac{|f(\bar{y})|}{\|\bar{y}\|_{\bar{L}_1}} \qquad (2\text{-}13)$$

is given by

$$\|f\| = \max \left\{ \frac{1}{C_1}\|u\|_p \ , \ \frac{1}{C_2} \max_i |b_i| \right\} \qquad (2\text{-}14)$$

The proofs of the preceding statements are given in Appendix I.

To apply these developments to the problem (2-5 - 2-6) we recognize that the functions $h_1(T,\tau)$ and $h_2(t_i,\tau)$, being piecewise continuous in $[0,T]$ are in $L_q[0,T]$ for every T so that we may construct from them composite vectors of the form (2-8):

$$\bar{\varphi} = \left\{ h_1(T,\tau), \ \underline{0} \right\} \qquad \underline{0} = [0,0,\ldots,0]$$

$$\bar{\theta}_k = \left\{ h_2(t_k,\tau), \ -\underline{e}_k \right\} \quad \underline{e}_k = [0,0,\ldots,1,\ldots,0], \quad k = 1,\ldots,n \qquad (2\text{-}15)$$

where the one occurring in $\underline{e}_k$ is the k^{th} column. If now, we require that a functional f of the form (2-12) satisfy

$$f(\bar{\varphi}) = x_d^1 - x_o^1(T) \tag{2-16}$$

we are requiring that the $u(t)$ component of f satisfy (2-6). If further, we require that f satisfy

$$f(\bar{\theta}_k) = - x_o^2(t_k) \qquad k = 1, \ldots, n \tag{2-17}$$

we are requiring that the b_k component of f be given by

$$b_k = x_o^2(t_k) + \int_0^T h_2(t_k,\tau)u(\tau)d\tau = x^2(t_k) \tag{2-18}$$

where the last of the equalities in (2-18) follows from (2-1) and the fact that $h_2(t_k,\tau) = 0$ for $\tau > t_k$. The norm of f, as given by (2-14), then becomes

$$\|f\| = \max \left\{ \frac{1}{C_1}\|u\|_p \ , \ \frac{1}{C_2} \max_k |x^2(t_k)| \right\} \tag{2-19}$$

which is the left side of Eq. (2-5).

It is now clear that the abstract problem of finding the bounded linear functional of minimum norm on $\bar{L}_1$, which maps the given elements $\bar{\varphi}$, $\bar{\theta}_k$, $k = 1, \ldots, n$ of this space into the given fixed scalars specified by (2-16 - 2-17) is equivalent to solving the variational problem (2-5 - 2-6)

and the first value of T for which this functional has its norm equal to unity is the optimal time.

If the scalars specified by (2-16 - 2-17) are arbitrary then the above problem is well posed only if the vectors $\bar{\varphi}$ and $\bar{\theta}_k$, $k = 1, \ldots, n$, are linearly independent and although this is obviously true in the present case for any nontrivial $h_1(T,\tau)$, suitable restrictions (such as total controllability)[1] must be placed on the plant when there are multiple terminal constraints to ensure this condition.

Since the vectors $\bar{\varphi}$ and $\bar{\theta}_k$, $k = 1, \ldots, n$, are linearly independent, they span an $n + 1$ dimensional linear space contained in $\bar{L}_1$ and we can define a linear functional f_1 on this space satisfying (2-16 - 2-17) by

$$f_1\Big(\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k\bar{\theta}_k\Big) = \alpha f_1(\bar{\varphi}) + \sum_{k=1}^{n} \gamma_k f(\bar{\theta}_k) = \alpha c + \sum_{k=1}^{n} \gamma_k d_k \qquad (2\text{-}20)$$

where

$$c = x_d^1 - x_o^1(T) \quad , \quad d_k = - x_o^2(t_k) \qquad (2\text{-}21)$$

The norm of f_1 on this $n + 1$ dimensional space is found from (2-13)

$$\|f_1\| = \sup_{\substack{\alpha,\gamma_k \\ \text{not all} = 0}} \frac{|f_1(\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k d_k)|}{\|\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k \bar{\theta}_k\|_{\bar{L}_1}}$$

$$= \sup_{\substack{\alpha,\gamma_k \\ \text{not all} = 0}} \frac{|\alpha c + \sum_{k=1}^{n} \gamma_k d_k|}{\|\alpha\bar{\varphi} + \sum \gamma_k \bar{\theta}_k\|_{\bar{L}_1}} \qquad (2\text{-}22)$$

or, since $\alpha c + \sum_{k=1}^{n} \gamma_k d_k = 0$ clearly does not give the largest value of the ratio,

$$\|f_1\| = \frac{1}{\inf_{\alpha,\gamma_k} \|\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k \bar{\theta}_k\|_{\bar{L}_1}} \qquad (2\text{-}23)$$

$$\alpha c + \sum_{k=1}^{n} \gamma_k d_k = 1$$

or using the explicit form of the norm in L, given by (2-9 - 2-11)

$$\|f_1\| = \frac{1}{\inf_{\alpha,\gamma_k} \left\{C_1 \left[\int_0^T |B(\tau)|^q d\tau\right]^{1/q} + C_2 \sum_{k=1}^{n} |\gamma_k|\right\}} \qquad (2\text{-}24)$$

$$\alpha c + \sum_{k=1}^{n} \gamma_k d_k = 1$$

where

$$B(\tau) = \alpha h_1(T,\tau) + \sum_{k=1}^{n} \gamma_k h_2(t_k,\tau)$$

and the infimum is evaluated over all real values of α and γ_k, $k = 1, \ldots, n$, satisfying $\alpha c + \sum_{k=1}^{n} \gamma_k d_k = 1$.

We now make use of the Hahn-Banach theorem which is stated below. For a proof, see Liusternik and Sabolev.[29]

Hahn Banach Theorem - Let M be a linear subspace of a normed linear space N, and let f be a linear functional defined on M. Then f can be extended to a linear functional defined on the whole space N without increase in norm.

Now, identifying the $n + 1$ dimensional linear space spanned by $\bar{\varphi}$ and $\bar{\theta}_k$, and the space $\bar{L}_1$ with M and N of the above theorem, respectively, the Hahn-Banach theorem asserts that the linear functional f_1 defined by (2-20) can be extended to a linear functional f over L_1 such that $\|f\| = \|f_1\|$. Since any extension of f_1 over L_1 must have its norm equal or greater than that of f_1 we therefore infer the existence of a functional of minimum norm satisfying (2-16 - 2-17) which has its norm equal to (2-24). The first value of T for which (2-24) equals unity is the optimal time hereafter called T_o.

2.3 Form of the Optimal Control

The existence of a solution to the problem (2-16 - 2-17) and therefore to (2-5 - 2-6) at time T_o has just been es-

tablished and we now turn to a derivation of the actual form of the optimal control. Let the infimum in the expression (2-24) $(T = T_o)$ be attained at the values $\alpha = \alpha^*$, $\gamma_k = \gamma_k^*$, $k = 1, \ldots, n$ with $\alpha^* c + \sum_{k=1}^{n} \gamma_k^* d_k = 1$. Also, let f be any functional of minimum norm satisfying (2-16 - 2-17). By what has just preceded, there exists at least one, its norm equals unity and the $u(t)$ component of the functional is the optimal control. From the definition of the norm of a functional, (2-13),

$$|f(\bar{y})| \leq \|f\| \|\bar{y}\|_{\bar{L}_1} \tag{2-25}$$

But

$$1 = |f(\alpha^* \bar{\varphi} + \sum_{k=1}^{n} \gamma_k^* \bar{\theta}_k)| \leq \|f\| \|\alpha^* \bar{\varphi} + \sum_{k=1}^{n} \gamma_k^* \bar{\theta}_k\|_{\bar{L}_1} = 1 \tag{2-26}$$

where the last equality is obtained from (2-24) using the fact that the infimum is attained at the values $\alpha = \alpha^*$, $\gamma_k = \gamma_k^*$, $k = 1, \ldots, n$ and that $\|f_1\| = \|f\|$. Therefore

$$|f(\alpha^* \bar{\varphi} + \sum_{k=1}^{n} \gamma_k^* \bar{\theta}_k)| = \|f\| \|\alpha^* \bar{\varphi} + \sum_{k=1}^{n} \gamma_k^* \bar{\theta}_k\|_{\bar{L}_1} \tag{2-27}$$

or more explicitly, using (2-9 - 2-12) and letting $\alpha^* h_1(T_o, \tau) + \sum_{k=1}^{n} \gamma_k^* h_2(t_k, \tau) = \sigma^*(\tau)$ we have

$$\left|\int_0^{T_o} \sigma^*(\tau)u(\tau)d\tau + \sum_{k=1}^{n} \gamma_k^* b_k\right| = \tag{2-28}$$

$$\max\left\{\frac{1}{C_1}\|u\|_p \ , \ \frac{1}{C_2}\max_k|b_k|\right\}\left[C_1\|\sigma^*\|_q + C_2\sum_{k=1}^{n}|\gamma_k^*|\right]$$

Since $|a + b| \leq |a| + |b|$, it follows that

$$\left|\int_0^{T_o} \sigma^*(\tau)u(\tau)d\tau + \sum_{k=1}^{n} \gamma_k^* b_k\right| \leq \left|\int_0^{T_o} \sigma^*(\tau)u(\tau)d\tau\right| + \left|\sum_{k=1}^{n} \gamma_k^* b_k\right| \tag{2-29}$$

Using various forms of the Holder inequality which is derived in Appendix II,

$$\left|\int_0^{T_o} \sigma^*(\tau)u(\tau)d\tau\right| \leq \|\sigma^*\|_q\|u\|_p = C_1\|\sigma^*\|_q\left[\frac{1}{C_1}\|u\|_p\right] \tag{2-30}$$

$$\left|\sum_{k=1}^{n} \gamma_k^* b_k\right| \leq \left[\sum_{k=1}^{n}|\gamma_k^*|\right]\left[\max_k|b_k|\right] = \left[C_2\sum_{k=1}^{n}|b_k^*|\right]\left[\frac{1}{C_2}\max_k|b_k|\right] \tag{2-31}$$

Adding the above two inequalities

$$\left|\int_0^{T_o} \sigma^*(\tau)u(\tau)\,d\tau\right| + \left|\sum_{k=1}^{n} \gamma_k^* b_k\right| \le C_1\|\sigma^*\|_q\left[\frac{1}{C_1}\|u\|_p\right] + \left[C_2 \sum_{k=1}^{n} |\gamma_k^*|\right]\left[\frac{1}{C_2} \max_k |b_k|\right] \tag{2-32}$$

From

$$ab + cd \le \left[\max\left\{b,d\right\}\right][a + c] \qquad a, b, c, d > 0 \tag{2-33}$$

it follows that

$$C_1\|\sigma^*\|_q\left[\frac{1}{C_1}\|u\|_p\right] + \left[C_2 \sum_{k=1}^{n} |\gamma_k^*|\right]\left[\frac{1}{C_2} \max_k |b_k|\right] \le \max\left\{\frac{1}{C_1}\|u\|_p\,,\ \frac{1}{C_2} \max_k |b_k|\right\}\left[C_1\|\sigma^*\|_q + C_2 \sum_{k=1}^{n} |\gamma_k^*|\right] \tag{2-34}$$

Comparing (2-28) and (2-29 - 2-34), one finds that all inequalities in (2-29 - 2-34) must be satisfied by equality. A necessary condition for equality in (2-30) is

$$u(t) = K|\sigma^*(t)|^{q-1} \operatorname{sgn} \sigma^*(t) \tag{2-35}$$

where $\operatorname{sgn} a = 1$ if $a > 0$, $\operatorname{sgn} a = -1$ if $a < 0$ and we assume $\sigma^*(t)$ vanishes only on a set of isolated points. K is

an arbitrary nonzero constant. A necessary condition for equality in (2-33) is $b = d$, so for equality in (2-34)

$$\frac{1}{C_1}\|u\|_p = \frac{1}{C_2}\max_k |b_k| \tag{2-36}$$

(2-36) is true provided $\sum_{k=1}^{n} |\gamma_k^*| > 0$ and this condition holds whenever the output constraints are active, i.e., the constrained output solution differs from the unconstrained output solution. Therefore

$$\frac{1}{C_1}\|u\|_p = \max\left\{\frac{1}{C_1}\|u\|_p\ ,\ \frac{1}{C_2}\max_k |b_k|\right\} = \|f\| = 1 \tag{2-37}$$

or $\|u\|_p = C_1$ and the control operates on the boundary of its constraint. From (2-35), (2-37)

$$\begin{aligned}\|u\|_p &= K\left[\int_0^{T_o}\left(|\sigma^*(t)|^{q-1}\right)^p\right]^{1/p} \\ &= K\left[\int_0^{T_o}|\sigma^*(t)|^q\right]^{1/p} = K\|\sigma^*\|^{q/p} = K\|\sigma^*\|^{q-1} = C_1\end{aligned} \tag{2-38}$$

or

$$K = C_1\|\sigma^*\|^{1-q} \tag{2-39}$$

and

$$u(t) = C_1 \|\sigma^*\|^{1-q} |\sigma^*(t)|^{q-1} \operatorname{sgn} \sigma^*(t) \qquad (2\text{-}40)$$

(2-40) is the explicit form of the optimal control.

2.4 Concepts from Real Variable Theory

Some concepts from real variable theory which will be useful in the next two chapters will now be presented. Standard references for this material are Natanson[32] and Riesz and Nagy.[33]

Let a function $f(t)$ be defined and finite on the interval $[a,b]$. Subdivide $[a,b]$ into parts by means of the points

$$a = t_o < t_1 < \ldots < t_n = b \qquad (2\text{-}41)$$

and form the sum

$$V = \sum_{k=0}^{n-1} |f(t_{k+1}) - f(t_k)| \qquad (2\text{-}42)$$

Definition: The least upper bound of the set of all possible sums V is called the total variation of the function $f(t)$ on $[a,b]$ and is designated by $\overset{b}{\underset{a}{V}}(f)$. If $\overset{b}{\underset{a}{V}}(f) < \infty$, then $f(t)$ is said to be a function of finite variation on $[a,b]$.

Theorem 1: Let an infinite family of functions $F = \{f_n(t)\}$ be defined on the segment $[a,b]$. If all functions of the family and the total variation of all functions of the family are bounded by a single number K, i.e.,

$$|f_n(t)| \leq K \qquad \overset{b}{\underset{a}{V}}(f_n) \leq K \tag{2-43}$$

then there exists a sequence $\{f_{n_k}(t)\}$ in the family F which converges at every point of $[a,b]$ to some function $f(t)$ of finite variation.

The proof of the above theorem, which is known as Helly's first theorem can be found in Natanson.

Functions of finite variation are important when working with a generalization of the Riemann integral known as the Stieltjes integral. Let $f(t)$ and $g(t)$ be finite functions defined on the closed interval $[a,b]$. Subdivide $[a,b]$ into parts by means of the points

$$a = t_o < t_1 < \ldots < t_n = b \tag{2-44}$$

and choose a point λ_k in $[t_k, t_{k+1}]$ for $k = 0, 1, n - 1$, and form the sum

$$\sigma = \sum_{k=0}^{n-1} f(\lambda_k)[g(t_{k+1}) - g(t_k)] \tag{2-45}$$

If the sum tends to a finite limit I as $\max(t_{k+1} - t_k) \to 0$ independently of both the method of subdivision and the choice of the points λ_k, this limit is called the Stieltjes integral of the function $f(t)$ with respect to the function $g(t)$ and is designated by $\int_q^b f(t)dg(t)$.

The exact meaning of the definition is this: The number I is the Stieltjes integral of the function $f(t)$ with respect to the function $g(t)$ if, for every $\epsilon > 0$ there exists a $\delta > 0$ such that for an arbitrary method of subdivision for which $\max(t_{k+1} - t_k) < \delta$, the inequality $|\sigma - I| < \epsilon$ holds, for all choices of the points λ_k. The Riemann integral is a special case of the Stieltjes integral, obtained by setting $g(t) = t$.

A sufficient condition for the existence of $\int_q^b f(t)dg(t)$ is that the function $f(t)$ be continuous on $[a,b]$ and $g(t)$ have finite variation on $[a,b]$. Furthermore, if $g(t)$ has a Riemann integrable derivative $g'(t)$ at every point of $[a,b]$, then

$$\int_a^b f(t)dg(t) = \int_a^b f(t)g'(t)dt \qquad (2\text{-}46)$$

where the right side of (2-46) is an ordinary Riemann integral.

An important theorem which gives sufficient conditions for the passage to the limit under the Stieltjes integral sign is Helly's second theorem which is stated below. Again, for a proof see Natanson.

Theorem 2: Let $f(t)$ be a continuous function defined on the interval $[a,b]$ and let $\{g_n(t)\}$ be a sequence of functions which converges to a finite function $g(t)$ at every point of $[a,b]$. If, for some positive constant K,

$$\overset{b}{\underset{a}{V}}(g_n) < K \tag{2-47}$$

for all n, then

$$\lim_{n \to \infty} \int_a^b f(t)\,dg_n(t) = \int_a^b f(t)\,dg(t) \tag{2-48}$$

We conclude with the following two lemmas.

Lemma 1: Let $\{f_n(t)\}$ be a sequence of functions which converges to a finite function $f(t)$ of finite variation at every point of $[a,b]$. If $\overset{b}{\underset{a}{V}}(f_n) \leq K$ for all n, where K is some positive constant, then an infinite subsequence $\{f_{n'}\}$ can be chosen such that

$$\lim_{n' \to \infty} \overset{b}{\underset{a}{V}}(f_{n'}) \geq \overset{b}{\underset{a}{V}}(f) \tag{2-49}$$

Proof: $\left\{\overset{b}{\underset{a}{V}}(f_n)\right\}$ is a sequence of real numbers bounded by K and therefore contains a convergent subsequence $\left\{\overset{b}{\underset{a}{V}}(f_{n'})\right\}$ with

$$\lim_{n' \to \infty} \overset{b}{\underset{a}{V}}(f_{n'}) = M \leq K \tag{2-50}$$

It is now shown that $M \geq \overset{b}{\underset{a}{V}}(f)$. In fact, since f has finite variation, there exists a finite subdivision of $[a,b]$ such that for arbitrary $\epsilon > 0$

$$\overset{b}{\underset{a}{V}}(f) \geq \sum_{k=0}^{m-1} |f(t_{k+1}) - f(t_k)| > \overset{b}{\underset{a}{V}}(f) - \frac{\epsilon}{2} \tag{2-51}$$

Obviously $\{f_{n'}(t)\}$ converges pointwise to $f(t)$ on $[a,b]$, therefore there exists some N such that

$$\sum_{k=0}^{m-1} |f(t_{k+1}) - f(t_k)| - \sum_{k=0}^{m-1} |f_{n'}(t_{k+1}) - f_{n'}(t_k)| < \frac{\epsilon}{2} \tag{2-52}$$

for all $n' \geq N$ and it then follows that

$$\sum_{k=0}^{m-1} |f_{n'}(t_{k+1}) - f_{n'}(t_k)| > \overset{b}{\underset{a}{V}}(f) - \epsilon \tag{2-53}$$

But

$$\overset{b}{\underset{a}{V}}(f_{n'}) \geq \sum_{k=0}^{m-1} |f_{n'}(t_{k+1}) - f_{n'}(t_k)| > \overset{b}{\underset{a}{V}}(f) - \epsilon \tag{2-54}$$

for all $n' > N$. Therefore, taking limits as $n' \to \infty$

$$\lim_{n' \to \infty} \overset{b}{\underset{a}{V}}(f_{n'}) = M \geq \overset{b}{\underset{a}{V}}(f) - \epsilon \tag{2-55}$$

and since ϵ is arbitrary, the lemma is proved.

Lemma 2: If $f(t)$ is any function of finite variation on $[a,b]$, there exists a sequence of step functions $\{f_n(t)\}$ converging pointwise to $f(t)$ on $[a,b]$, such that $\overset{b}{\underset{a}{V}}(f_n) \leq \overset{b}{\underset{a}{V}}(f)$ for all n and

$$\lim_{n \to \infty} \overset{b}{\underset{a}{V}}(f_n) = \overset{b}{\underset{a}{V}}(f) \tag{2-56}$$

Proof: Since $f(t)$ has finite variation on $[a,b]$, it has a countable number of points of discontinuity[32] and at every discontinuity point t_o, the limits

$$f(t_o + 0) = \lim_{\substack{t \to t_o \\ t > t_o}} f(t) \qquad f(t_o - 0) = \lim_{\substack{t \to t_o \\ t < t_o}} f(t) \tag{2-57}$$

exist.

If at a discontinuity point of $f(t)$ we define $f(t)$ as being equal to its right hand limit, then this function is now defined on $[a,b]$.

Define the sequence of step functions $\{f_n(t)\}$ as follows: let the n^{th} subdivision of $[a,b]$ be

$$a,\ a + \frac{b - a}{n},\ a + \frac{2(b - a)}{n},\ \ldots,\ b \tag{2-58}$$

and define

$$f_n\left(a + \frac{k(b - a)}{n}\right) = f\left(a + \frac{k(b - a)}{n}\right) \tag{2-59}$$

$$f_n(t) = f\left(a + \frac{k(b - a)}{n}\right)\ t \epsilon \left(a + \frac{(k - 1)(b - a)}{n},\ a + \frac{k(b - a)}{n}\right)$$

We let $n = 2, 4, 8, \ldots$ and show that $f_n(t)$ converges everywhere to $f(t)$ on $[a,b]$. Let t_1 be any point of $[a,b]$. If t_1 is one of the denumerable points of the subdivision scheme then there exists an N and a $k \leq N$ such that $t_1 = a + \frac{k(b - a)}{N}$. It is clear from the way the sequence $\{f_n(t)\}$ was defined that $f_n(t_1) = f(t_1)$ for all $n \geq N$ and therefore $\lim_{n \to \infty} f_n(t_1) = f(t_1)$.

Let t_1 be a point not in the subdivision scheme. Then t_1 is either a continuity point or a discontinuity point. If t_1 is a continuity point then corresponding to any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(t_2) - f(t_1)| < \epsilon$ for all t_2 such that $|t_2 - t_1| < \delta$. If N is chosen large enough so that $\frac{b - a}{N} < \delta$, then for all $n \geq N$, $|f_n(t_1) - f(t_1)| < \epsilon$, which is exactly what we mean when we say $f_n(t_1)$ converges to $f(t_1)$.

If t_1 is a discontinuity point which is not a point of the subdivision scheme, then it is always contained in

the interior of any subdivision. Since the right hand limit of $f(t_1)$ exists, then for any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(t_2) - f(t_1)| < \delta$ for all t_2 such that $0 \leq t_2 - t_1 < \delta$. If N is chosen large enough so that $\frac{b-a}{N} < \delta$, then for all $n \geq N$, $|f_n(t_1) - f(t_1)| < \epsilon$ and therefore $f_n(t)$ converges to $f(t)$ on $[a,b]$.

Also, $\lim_{n \to \infty} \overset{b}{\underset{a}{V}}(f_n) = \overset{b}{\underset{a}{V}}(f)$. In fact

$$\overset{b}{\underset{a}{V}}(f_n) = \sum_{k=0}^{n-1} \left| f\left(a + \frac{(k+1)(b-a)}{n}\right) - f\left(a + \frac{k(b-a)}{n}\right) \right| \qquad (2\text{-}60)$$

because $f_n(t)$ is constant in $\left(a + \frac{k(b-a)}{n},\ a + \frac{(k+1)(b-a)}{n}\right)$. Employing the inequality $|a - c| \leq |a - b| + |b - c|$ it is seen that $\left\{\overset{b}{\underset{a}{V}}(f_n)\right\}$ is a monotone increasing sequence of real numbers and therefore has a well defined limit which may be finite or infinite. But by inspection of (2-60), $\overset{b}{\underset{a}{V}}(f) \geq \overset{b}{\underset{a}{V}}(f_n)$ for all n and therefore

$$\lim_{n \to \infty} \overset{b}{\underset{a}{V}}(f_n) \leq \overset{b}{\underset{a}{V}}(f) \qquad (2\text{-}61)$$

Utilizing Lemma 1, the reverse inequality also holds and it follows that

$$\lim_{n \to \infty} \overset{b}{\underset{a}{V}}(f_n) = \overset{b}{\underset{a}{V}}(f) \qquad \text{Q.E.D.} \qquad (2\text{-}62)$$

CHAPTER 3. AN APPROXIMATION TO THE PROBLEM FOR PLANTS WITH CONTINUOUSLY BOUNDED OUTPUTS

The preceding chapter suggests a natural way of obtaining an approximate solution for problems with constraints on some or all of the outputs at every instant of time during the transition period. This method of approximation which we term "an approximation to the problem" is developed in this chapter.

3.1 Formulation of an Equivalent Problem

For ease of presentation, we consider a double output-single input plant with a terminal constraint on one output and a magnitude constraint imposed at every instant of time during the transition period on the other. The results obtained extend directly to the case with multiple terminal and/or output constraints and multi-input systems as shown in 3.6.

The plant input-output relations are

$$x^1(t) = x_o^1(t) + \int_0^t h_1(t,\tau)u(\tau)d\tau$$

$$x^2(t) = x_o^2(t) + \int_0^t h_2(t,\tau)u(\tau)d\tau \qquad (3\text{-}1)$$

where $x^1(t)$ and $x^2(t)$ are outputs of the plant and $x_o^1(t)$ and $x_o^2(t)$ express the effect on $x^1(t)$ and $x^2(t)$ respectively, of initial conditions in the plant at $t = 0$ and are assumed to be known and continuous. The initial time is assumed to be $t = 0$ without loss of generality. The functions $h_1(t,\tau)$ and $h_2(t,\tau)$ are assumed to be bounded piecewise continuous functions of τ on the interval $[0,t]$ and equal 0 for $\tau > t$ for all finite values of t.

The problem considered is that of finding the smallest T, the optimal time, and the corresponding control function u(t), the optimal control, such that

$$x^1(T) = x_o^1(T) + \int_0^T h_1(T,\tau)u(\tau)d\tau = x_d^1 \qquad (3\text{-}2)$$

where x_d^1 is some desired final value. We have a constraint on the control $u(t)$ of the form

$$\left(\int_0^T |u(t)|^p dt\right)^{1/p} \leq C_1 \qquad C_1 > 0 \qquad 1 < p \leq \infty \qquad (3\text{-}3)$$

and a constraint on the output $x^2(t)$ of the form

$$|x^2(t)| \leq C_2 \qquad C_2 > 0 \qquad 0 \leq t \leq T \qquad (3\text{-}4)$$

A problem equivalent to (3-1) through (3-4) which is amenable to solution by the product space approach of Katz and Kranc is now formulated. The output $x^2(t)$ is defined by an (Lebesgue) integral and is, therefore, a continuous function of t. If the magnitude of $x^2(t)$ is constrained to be equal or less than C_2 on any countable dense set of times $\{t_i\}$ in $[0,T]$, it follows from the continuity of $x^2(t)$ that its magnitude is equal or less than C_2 at all instants of time in $[0,T]$. Therefore, for any T, any control satisfying (3-1) through (3-4) must also satisfy (3-1) through (3-3) and

$$|x_2(t_i)| \leq C_2 \qquad i = 1, 2, \ldots \tag{3-5}$$

and conversely. If we now ask for the smallest T and the corresponding control $u(t)$ such that (3-1) through (3-3), (3-5) are satisfied, then this control must also be the optimal control satisfying (3-1) through (3-4). In fact if we assume the existence of a control $u_1(t)$ satisfying (3-1) through (3-4) which yields a transition time less than that obtained when $u(t)$ is applied, then from what has preceded $u_1(t)$ satisfies (3-1) through (3-3), (3-5) and this contradicts the fact that $u(t)$ is the optimal control satisfying these relationships.

We note that any countable dense set of times $\{t_i\}$ in $[0,T]$ has the form $t_i = r_i T$ where $\{r_i\}$ is a countable dense set of real numbers (e.g., the rationals) in $[0,1]$.

3.2 Discrete Point Approximation to the Equivalent Problem

As in Chapter 2, we tentatively fix T and ask for the control function $u(t)$ which makes

$$\max \left\{ \frac{1}{C_1} \|u\|_p, \frac{1}{C_2} \sup_i |x^2(t_i)| \right\} = \text{minimum} \tag{3-6}$$

while keeping

$$x_o^1(T) + \int_0^T h_1(T,\tau)u(\tau)\,d\tau = x_d^1 \tag{3-7}$$

The smallest T for which the minimum in (3-6) is equal or less than unity then solves the equivalent problem with constraints, (3-2), (3-3), (3-5).

Our product space $\bar{L}_1$ consists of composite vectors $\bar{y} = \{y(t), \underline{a}\}$ pairing functions $y(t)$ in $L_q[0,T]$ with vectors $\underline{a} = \{a_1, a_2, \ldots\}$ in ℓ_1 with norm

$$\|\underline{a}\|_{\ell_1} = \sum_{i=1}^{\infty} |a_i| \tag{3-8}$$

The norm in $\bar{L}_1$ is

$$\|\bar{y}\|_{\bar{L}_1} = C_1\|y\|_q + C_2\|\underline{a}\|_{\ell_1} \tag{3-9}$$

and with this norm $\bar{L}_1$ is a Banach space. Analogous to (2-12), (2-14), we have that any bounded linear functional on $\bar{L}_1$ has the representation

$$f(\bar{y}) = \int_0^T y(t)u(t)dt + \sum_{i=1}^{\infty} a_i b_i \tag{3-10}$$

with $u(t)$ in $L_p[0,T]$ and $\underline{b} = \{b_1, b_2, \ldots\}$ an infinite vector with a uniform bound on its components

The norm of the functional defined by (2-13) is

$$\|f\| = \max \left\{ \frac{1}{C_1} \|u\|_p, \frac{1}{C_2} \sup_i |b_i| \right\} \tag{3-11}$$

We now recognize that the functions $h_1(T,\tau)$ and $h_2(t_i,\tau)$ being bounded and piecewise continuous in $[0,T]$ are in $L_q[0,T]$ for every T so that we may construct from them composite vectors in $\bar{L}_1$

$$\bar{\varphi} = \left\{ h_1(T,\tau), \quad \underline{0} \right\} \quad \underline{0} = [0,0,\ldots]$$

$$\bar{\theta}_k = \left\{ h_2(t_k,\tau), -\underline{e}_k \right\} \quad \underline{e}_k = [0,0,\ldots,1,\ldots,0,\ldots] k = 1,2,\ldots \tag{3-12}$$

where the one occurring in $\underline{e}_k$ is in the k^{th} column. As in Chapter 2, it is clear that by requiring

$$f(\bar{\varphi}) = x_d^1 - x_o^1(T)$$

$$f(\bar{\theta}_k) = - x_o^2(t_k) \qquad k = 1, 2, \ldots \tag{3-13}$$

that the norm of f, (the u(t) component of f satisfies the terminal condition) as given by (3-11) becomes

$$\|f\| = \left\{\frac{1}{C_1}\|u\|_p, \frac{1}{C_2}\sup_k |x^2(t_k)|\right\} \tag{3-14}$$

which is the left side of (3-6).

A straightforward application of the methods of Chapter 2 yields that the functional of minimum norm satisfying (3-13) (which has as its u(t) component the control satisfying (3-6) and (3-7)) has its norm equal to

$$\|f\| = \frac{1}{\inf_{\alpha,\gamma_k,n} \left\{C_1\left[\int_0^T |B(\tau)|^q d\tau\right]^{1/q} + C_2 \sum_{k=1}^{n} |\gamma_k|\right\}} \tag{3-15}$$
$$\alpha c + \sum_{k=1}^{n} \gamma_k d_k = 1$$

where

$$B(\tau) = \alpha h_1(T,\tau) + \sum_{k=1}^{n} \gamma_k h_2(t_k,\tau)$$

and

$$c = x_d^1 - x_o^1(T) \tag{3-16}$$

$$d_k = -x_o^2(t_k) \tag{3-17}$$

The infimum of (3-15) is not only over α and the n variables γ_k, but also over n itself because of the infinite

dimensionality of the linear space spanned by the vectors φ and $\bar{\theta}_k$, $k = 1, 2, \ldots$. It has been assumed that these vectors are linearly independent.

The first value of T for which the right side of (3-15) is equal or less than unity is the optimal time. In order to find the optimal time it is therefore necessary to perform a constrained infinite dimensional minimization, the minimization of the denominator of (3-15)

$$\inf_{\alpha,\gamma_k,n} \left\{ C_1 \left[\int_0^T \left| \alpha h_1(T,\tau) + \sum_{k=1}^{n} \gamma_k h_2(t_k,\tau) \right|^q d\tau \right]^{1/q} + C_2 \sum_{k=1}^{n} |\gamma_k| \right\} \tag{3-18}$$

$$\alpha c + \sum_{k=1}^{n} \gamma_k d_k = 1$$

where n is varying and unbounded. As a means of approximating this quantity, we pick s values of $t_i = r_i T$, $r_i \in [0,1]$, $i = 1, \ldots, s$, and then assume all γ_k equal 0 except for k corresponding to the s values of t_i. The minimization is then performed over α and the s variables γ_k to obtain

$$\min_{\alpha,\gamma_i} \left\{ C_1 \left[\int_0^T \left| \alpha h_1(T,\tau) + \sum_{i=1}^{s} \gamma_i h_2(t_i,\tau) \right|^q d\tau \right]^{1/q} \right.$$

$$\alpha c + \sum_{i=1}^{s} \gamma_i d_i = 1 \qquad (3\text{-}19)$$

$$\left. + C_2 \sum_{i=1}^{s} |\gamma_i| \right\}$$

and we assume the finite dimensional minimization (3-19) equals (3-18). It is noted that we have replaced infimum by minimum in (3-19) as it can be shown that the infimum is always attained.

The nature of this approximation becomes evident upon comparison with the methods of Chapter 2 whereupon it is seen that (3-19) would result if we attempted to find the smallest T and the corresponding control for which the terminal and control constraints and output constraint at the s instants of time t_i are satisfied. The control obtained as the solution to this problem is then used as the approximate solution to the problem with continuous output constraint. The net effect of this approximation then is to replace the problem with continuous output constraint by a problem with output constraints at a finite number of discrete instants of time, solve this latter problem exactly and use its solution as an approximate solution to the original problem.

3.3 Convergence and Error Bounds

Two interesting features of using the above method of approximation are:

1. The existence of a solution to the problem with continuous output constraint guarantees the existence of an approximate solution.
2. The transfer time obtained from the approximate solution is always less than or at most equal to the optimal time for the problem with continuous output constraint.

Both of these statements are consequences of the fact that at the optimal time T_o, expression (3-18) has a value equal or greater than unity. Since expression (3-19) is equal or greater than (3-18), its value at T_o is equal or greater than unity. Furthermore, as shown in Appendix III, expression (3-19) viewed as a function of T is continuous and this implies the existence of a first value of T, say T_o' which is equal or less than T_o, for which this condition is satisfied. It follows that T_o' is the optimal time for the approximate problem.

Any approximate solution for the time optimal problem with continuous output constraint has, in essence, two parameters of importance, the magnitude of the transfer time and the violation of the output constraint. From statement 2

above, it is seen that the transfer time obtained from our approximate solution is certainly no larger than the true optimal time and therefore it is only necessary to concern ourselves with the output constraint violation. It will be shown that, assuming the existence of a solution to the problem with continuous output constraint, it is possible to limit this violation within any prespecified error. We first prove a theorem which yields an error bound for the approximate solution. The theorem is proved assuming the input-output relations are derived from the plant state equations $\left(c(t) = \text{identity matrix}\right)$ but the results are easily modified for general input output relations.

Proposition 3.1: Let the input-output relations of the plant be obtained from the plant state equation (1-2) so that the output at any instant of time t is related to the output at any other instant of time $t_1 < t$ by[3]

$$\underline{x}(t) = \varphi(t,t_1)\underline{x}(t_1) + \int_{t_1}^{t} H(t,\tau)u(\tau)d\tau \qquad (3\text{-}20)$$

where $\underline{x}(t)$ is an n-vector, $\varphi(t,\tau)$ is the $n \times n$ state transition matrix of the plant and $H(t,\tau) = \varphi(t,\tau)B(\tau)$ where $B(t)$ is an $n \times 1$ time-varying matrix, piecewise continuous on all finite values of the real line. If

$$\|u\|_p^{t_1-t} = \left[\int_{t_1}^{t} |u(t)|^p dt\right]^{1/p} \leq C_1 \tag{3-21}$$

and $x^i(t)$ is the i^{th} output of the system, then

$$|x^i(t) - x^i(t_1)| \leq \max_j \left\{\varphi_{ij}(t,t_1) - \delta_{ij}\right\} \sum_{i=1}^{n} |x^i(t_1)| + C_1\left[\int_{t_1}^{t} |h_i(t,\tau)|^q d\tau\right]^{1/q} \tag{3-22}$$

where $\varphi_{ij}(t,t_1)$ is the component in the i^{th} row and j^{th} column of the state transition matrix, $h_i(t,\tau)$ is the component in the i^{th} row of the impulse response matrix $H(t,\tau)$ $\left(H(t,\tau)\right.$ has only one column since we assume a single control$\left.\right)$ and

$$\delta_{ij} = 0, \quad j \neq i \qquad \delta_{ij} = 1 \quad j = i \tag{3-23}$$

Proof: From the i^{th} row of (3-20)

$$x^i(t) = \sum_{j=1}^{n} \varphi_{ij}(t,t_1)x^j(t_1) + \int_{t_1}^{t} h_i(t,\tau)u(\tau)d\tau \tag{3-24}$$

Subtracting $x^i(t_1)$ from both sides

$$x^i(t) - x^i(t_1) = \sum_{j=1}^{n} [\varphi_{ij}(t,t_1) - \delta_{ij}]x^j(t_1) + \int_{t_1}^{t} h_i(t,\tau)u(\tau)d\tau$$

Then

$$|x^i(t) - x^i(t_1)| \leq |\sum_{j=1}^{n} [\varphi_{ij}(t,t_1) - \delta_{ij}]x^j(t_1)| + |\int_{t_1}^{t} h_i(t,\tau)u(\tau)d\tau|$$

$$\leq \max_j \left\{\varphi_{ij}(t,t_1) - \delta_{ij}\right\} \sum_{i=1}^{n} |x^i(t_1)| \qquad (3\text{-}25)$$

$$+ \left[\int_{t_1}^{t} |h_i(t,\tau)|^q d\tau\right]\left[\int_{t_1}^{t} |u(\tau)|^p d\tau\right]^{1/p}$$

$$\leq \max_j \left\{\varphi_{ij}(t,t_1) - \delta_{ij}\right\} \sum_{i=1}^{n} |x^i(t_1)| \qquad (3\text{-}26)$$

$$+ C_1\left[\int_{t_1}^{t} |h_i(t,\tau)|^q d\tau\right]^{1/q}$$

where Holder's inequality was used in (3-25) and the control constraint (3-21) was used in (3-26). Q.E.D.

Theorem 3.1: For the plant described in Proposition 3.1, let t_2 be some time greater than $t_1 > 0$. If $|x^i(t_1)| \leq C_2$, then

$$|x^i(t)| \leq C_2 + \max_{\substack{t \\ t_1 \leq t \leq t_2}} \left\{ \max_j [\varphi_{ij}(t,t_1) - \delta_{ij}] \sum_{i=1}^{n} |x^i(t_1)| \right\} \tag{3-27}$$

$$+ C_1 \max_{\substack{t \\ t_1 \leq t \leq t_2}} \left[\int_{t_1}^{t} |h_i(t,\tau)|^q d\tau \right]^{1/q}$$

for all t satisfying $t_1 \leq t \leq t_2$, where $\|u\|_p^{t_1-t_2} = \left[\int_{t_1}^{t_2} |u(t)|^p dt \right]^{1/p} \leq C_1$. Furthermore, for any $\epsilon > 0$ there exists a $\delta > 0$ such that the right side of inequality (3-27) is equal or less than $C_2 + \epsilon$ for all t satisfying $t_1 \leq t \leq t_2$ with $t_2 - t_1 < \delta$.

Proof: From (3-22) and the fact that $\left[\int_{t_1}^{t} |u(t)|^p dt \right]^{1/p} \leq \left[\int_{t_1}^{t_2} |u(t)|^p dt \right]^{1/p}$ for $t_1 \leq t \leq t_2$ it follows that

$$\max_{\substack{t \\ t_1 \le t \le t_2}} |x^i(t) - x^i(t_1)| \le$$

$$\max_{\substack{t \\ t_1 \le t \le t_2}} \left\{ \max_j [\varphi_{ij}(t,t_1) - \delta_{ij}] \sum_{i=1}^{n} |x^i(t_1)| \right\} \qquad (3\text{-}28)$$

$$+ C_1 \max_{\substack{t \\ t_1 \le t \le t_2}} \left[\int_{t_1}^{t} |h_i(t,\tau)^q d\tau \right]^{1/q}$$

But

$$|x^i(t)| \le |x^i(t) - x^i(t_1)| + |x^i(t_1)|$$

therefore

$$\max_{\substack{t \\ t_1 \le t \le t_2}} |x^i(t)| \le |x^i(t_1)| + \max_{\substack{t \\ t_1 \le t \le t_2}} \left\{ |x^i(t) - x^i(t_1)| \right\} \qquad (3\text{-}29)$$

Combining (3-28) and (3-29) and using $|x^i(t_1)| \le C_2$ we obtain (3-27).

To prove the second assertion, observe that $h_i(t,\tau)$ is a bounded function on finite intervals of t and τ. Let M be a bound for $h_i(t,\tau)$ in a suitably large region containing $t = t_1, \tau = t_1$, then

$$C_1 \max \left[\int_{t_1}^{t} |h_i(t,\tau)|^q d\tau \right]^{1/q} < \frac{\epsilon}{2} \qquad (3\text{-}30)$$

for $|t_2 - t_1| < \frac{1}{M^q}\left(\frac{\epsilon}{2C_1}\right)^q$. Since the functions $\varphi_{ij}(t,t_1)$ satisfy

$$\varphi_{ij}(t_1,t_1) = 0, \qquad i \neq j, \qquad \varphi_{ii}(t_1,t_1) = 1$$

it follows that

$$\max_j [\varphi_{ij}(t,t_1) - \delta_{ij}] \tag{3-31}$$

equals 0 at $t = t_1$. The continuity of expression (3-31) as a function of t follows from the continuity of the $\varphi_{ij}(t,t_1)$. Then by the definition of continuity, corresponding to any ϵ' there exists a δ' such that (3-31) is less than ϵ' if $|t - t_1| < \delta'$. Letting

$\epsilon' = \dfrac{\epsilon}{2\sum_{j=1}^{n} |x^j(t_1)|}$, finding the corresponding δ' and setting

$$\delta = \min\left\{\delta', \frac{1}{M^q}\left(\frac{\epsilon}{2C_1}\right)^q\right\} \tag{3-32}$$

then

$$\max_{\substack{t \\ t_1 \le t \le t_2}} \left\{\max_j [\varphi_{ij}(t,t_1) - \delta_{ij}] \sum_{j=1}^{n} |x^j(t_1)|\right\} + C_1 \max_{\substack{t \\ t_1 \le t \le t_2}} \left[\int_{t_1}^{t} |h_i(t,\tau)|^q d\tau\right]^{1/q} < \epsilon \tag{3-33}$$

for all t such that $t_1 \leq t \leq t_2$ when $|t_2 - t_1| < \delta$ and the second assertion follows.

It is now a straightforward matter to prove the desired result stated below in Theorem 3.2.

Theorem 3.2: Assuming the existence of a solution for the problem with continuous output constraint specified by (3-1) through (3-4) then corresponding to any $\epsilon > 0$ there exists a discrete point approximation to the problem as described in Section 3.2 such that the magnitude of the constrained output $x^2(t)$ is equal or less than $C_2 + \epsilon$ for all instants of time during the transition period. Furthermore, this transfer time is less than or at most equal to the transfer time for the original problem.

Proof: Statement 1 of this section guarantees the existence of a discrete point approximation regardless of the number of points t_i at which the constraint $|x^2(t_i)| \leq C_2$ is enforced. Statement 2 asserts that the transfer time for any discrete point approximation is less than or at most equal to the transfer time of the original problem, which we designate by T_o.

Since $\|u\|_p \leq C_1$, Theorem 3.1 asserts the existence of a δ corresponding to the given ϵ such that

$$|x^2(t)| \leq C_2 + \epsilon \qquad t_1 \leq t \leq t_2 \tag{3-34}$$

if

$$|x^2(t_1)| \leq C_2 , \qquad |t_1 - t_2| < \delta \tag{3-35}$$

Moreover, it follows from the uniform continuity of the $\varphi_{ij}(t,\tau)$ on $[0,T_o] \times [0,T_o]$ that one δ can be chosen regardless of the location of t_1 in $[0,T_o]$.

If we now choose the number of points N used in the discrete point approximation so that

$$\frac{T_o}{N} < \delta$$

and solve the approximate problem with constraint points at $\frac{kT}{N}$, $k = 0, 1, 2, \ldots, N-1$, it follows that, since $T \leq T_o$, that

$$\frac{T}{N} \leq \frac{T_o}{N} < \delta$$

Since any value of t in the transition period $[0,T]$ satisfies $\frac{kT}{N} \leq t \leq \frac{(k+1)T}{N}$ for some k and $|x^2\left(\frac{kT}{N}\right)| \leq C_2$ by the method of construction of the approximate solution, we obtain from (3-34) and (3-35) $\left(\frac{(k+1)T}{N} - \frac{kT}{N} = \frac{T}{N} < \delta\right)$ that

$$|x^2(t)| \leq C_2 + \epsilon$$

for all t in the transition period. Q.E.D.

3.4 Construction of the Approximation

The first problem that arises in attempting to construct a discrete point approximation is that a particular approximation may not exist. If this is the case then a solution to the original problem does not exist and the imposed constraints are too severe. It is interesting to note, however, that the existence of an approximate solution does not imply the existence of an exact solution and conceivably there are cases for which we can construct an approximation although no exact solution exists.

In general, we do not have an upper bound on the transfer times of the approximate solutions and consequently the construction procedure is an iterative one. One possible method is to pick some initial number of points, say 3, and solve a 3-point approximation to the problem where the points are appropriately placed (the placement of points will be discussed below). If the output constraint violation is small enough, then the 3-point approximation is an acceptable approximate solution; if the violation is too large, construct a 4-point approximation, etc. and continue in this manner until an acceptable approximation is obtained. Assuming the existence of a solution to the original problem this procedure will yield an acceptable approximation.

When constructing an n-point approximation, it is desirable to have an optimal placement of the n points in the sense that any other placement yields a larger magnitude violation of the output constraint. Essentially nothing can be said about optimal placement for the general case, however, in many cases it is possible to obtain a satisfactory placement by using the unconstrained output solution as a guide. This is illustrated in the examples, but we now give the motivation behind this procedure. Suppose the given problem is that stated in Section 3.1. We first obtain the unconstrained output solution, i.e., no output constraint on $x^2(t)$, by any of the standard techniques such as the approach given by Kranc and Sarachik.[34] Let us assume that when the optimal control so obtained is applied to the system that the plot of x^2 versus time appears as in Fig. 3-1 where T_o is the optimal time without output constraint. It can be seen from this figure that if we tried to construct an n point approximation which has

$$0 \leq t_k \leq \frac{T_1}{T_o} T' \quad \text{or} \quad \frac{T_2}{T_o} T' \leq t_k \leq T' \qquad k = 1, \ldots, n$$

where t_k, $k = 1, \ldots, n$, are the points at which the output constraint is enforced and T' is the optimal time for the n point approximation, that this approximation would be identical to the unconstrained output solution previously

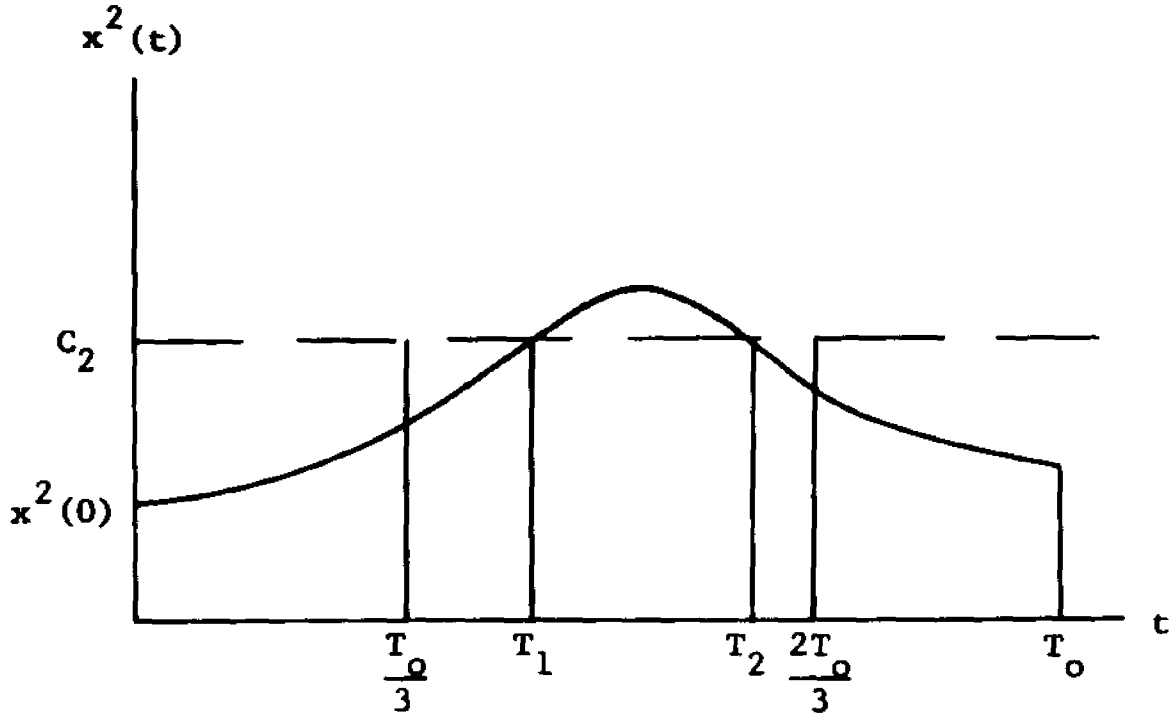


Fig. 3-1 Illustration of Point Placement for Discrete Point Approximation

obtained. For example, if we constructed a two-point approximation with the output constraint enforced at $\frac{T'}{3}$ and $\frac{2T'}{3}$, then since the unconstrained output solution satisfies the output constraint at these two points it will be the solution to the two-point approximation. The obvious remedy is to enforce the output constraint at at least one point for which the trajectory $x^2(t)$ associated with the unconstrained output solution violates the given output constraint. In Fig. 3-1 any point in the interval $\left(\frac{T_1}{T_o} T', \frac{T_2}{T_o} T'\right)$ would be acceptable. Computational experience seems to indicate that best results are obtained when most of the constraint points are placed in the interval of violation of the output constraint. If the resulting approximation yields too large an output constraint violation, then we construct another discrete point approximation with more constraint points using the approximation previously obtained as a guide in establishing the placement of the additional points.

It is possible to use some of the results on error bounds developed in Section 3.3 to aid in the placement of the constraint points. If we can find certain intervals for which it is impossible for the output constraint to be violated then any placement of constraint points in these intervals will be wasted. For example, if the state of the

system at the initial time $t = 0$ is the origin then (3-22) yields for the constrained output $x^2(t)$ that

$$|x^2(t)| \leq C_1\left[\int_0^t |h_2(t-\tau)|^q d\tau\right]^{1/q} = C_1\left[\int_0^t |h_2(\tau)|^q d\tau\right]^{1/q} \qquad (3\text{-}36)$$

where C_1 is the control constraint and we have assumed that the system is time invariant. Let T_1 be the first value of t such that

$$\left[\int_0^t |h_2(\tau)|^q d\tau\right]^{1/q} = \frac{C_2}{C_1}$$

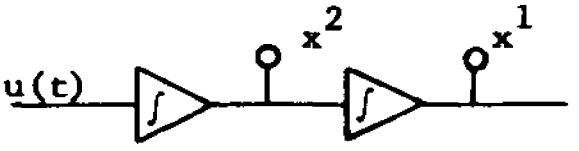
then it follows that $|x_2(t)| \leq C_2$ in $[0,T_1]$ and consequently no constraint points need be placed in this interval.

The discrete point approximation just discussed will in the general case, violate the output constraint and although this violation can be made arbitrarily small this method remains suitable only for problems with "soft" constraints, i.e., constraints which have error tolerances. For problems with "hard" constraints, the procedure must be modified. If the output constraint level is C_2, the modified procedure consists of constraining the output at discrete points to be equal or less than fC_2 where f is a real number between zero and one. Although the existence of a solution

to the original problem does not guarantee the existence of a discrete point approximation for a particular f, in many cases it is reasonable to expect that there exists some value of f such that some discrete point approximation does not violate the original output constraint level C_2. The particular choice of f depends upon the specific problem under investigation and is found by a combination of guessing and experimental trials. If we can construct an approximation which does not violate the original output constraint level then we obtain immediately as a by-product an upper bound on the optimal time of the exact solution. This follows because the solution just obtained satisfies all constraints of the problem (control, terminal, and output) and therefore the first value of T for which the problem constraints can be satisfied must be equal or less than the transfer time of this approximation. Using this upper bound and a lower bound obtained from the transfer time of any discrete point approximation with constraints placed at the original output level, it is possible to get an excellent idea of the output violation-transfer time trade off which enables us to make an intelligent selection among the possible approximations.

3.5 Examples

Consider the following double output single input linear system



where $x^1(0) = x^2(0) = 0$. The control $u(t)$ is constrained in magnitude to be equal or less than 1, the output $x^2(t)$ is constrained in magnitude to be equal or less than 1 and it is desired that $x^1 = 2$ in the shortest possible time. The input output relations are

$$x^1(t) = \int_0^t (t - \tau)u(\tau)d\tau$$

$$x^2(t) = \int_0^t u(\tau)d\tau$$

If we ignore the output constraint on $x^2(t)$, then by any of the standard techniques we find that the optimal time is $T = 2$. The optimal control is shown in Fig. 3-2 and the corresponding trajectory for $x^2(t)$ is shown in Fig. 3-3. We see that $x^2(t)$ violates the output constraint at $t = 1$ and remains above the admissible constraint level for the

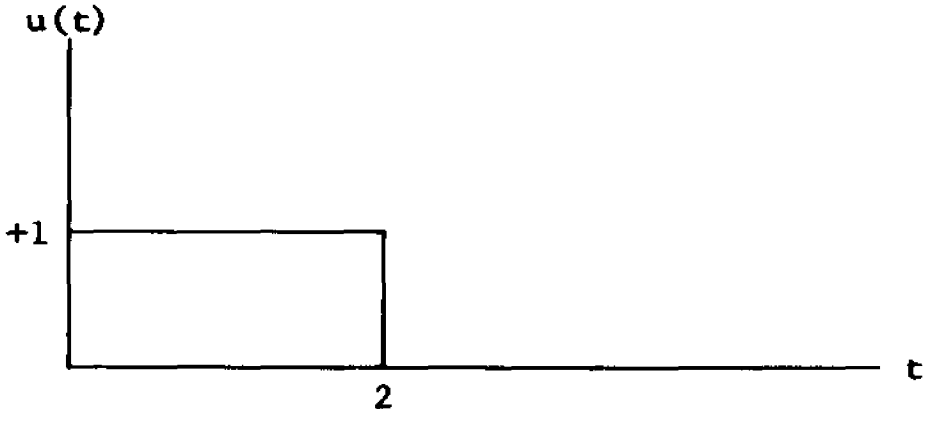


Fig. 3-2 Optimal Control for First Example without Output Constraint

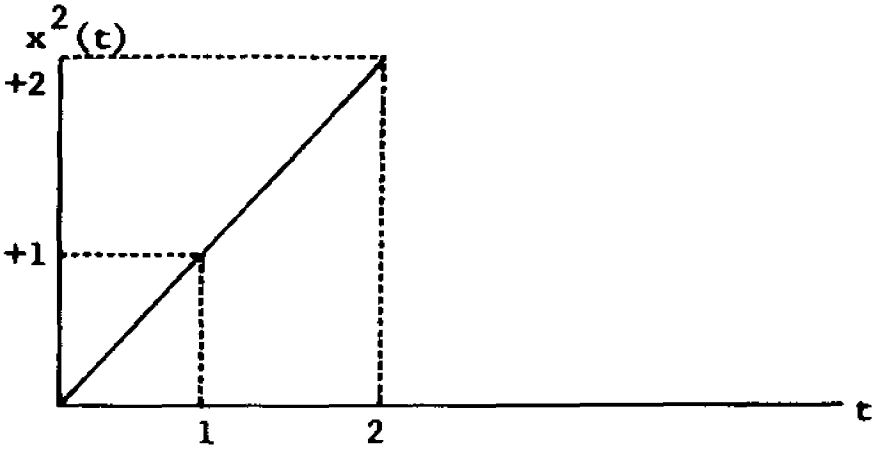


Fig. 3-3 Plot of x^2 for First Example without Output Constraint

duration of the transition period. Using expression (3-36) with $C_1 = q = 1$ we have that

$$|x^2(t)| \leq \int_0^t dt = t \leq 1 \quad \text{for} \quad t \leq 1$$

and therefore it is impossible for $x^2(t)$ to violate the constraint for $t \leq 1$ so that no constraint points are placed in $[0,1]$.

We now develop a 3-point approximation to the problem. From the remarks in the preceding paragraph it is clear that the constraint points should lie in the interval $[1,T_3]$ where T_3 is the optimal time for the 3-point approximation and it is decided to space them equidistantly in this interval as shown in Fig. 3-4, where $a = \frac{1}{7}(T_3 - 1)$ and is the maximum time difference between any point violating the output constraint and a point at which the constraint is satisfied. Making the appropriate substitutions in (3-19) and performing the finite dimensional minimizations we find that the transfer time obtained for this 3-point approximation is $T_3 = 2.36$ at which time $x^1 = 1.995$. The optimal control obtained from Eq.(2-40) is shown in Fig. 3-5, where

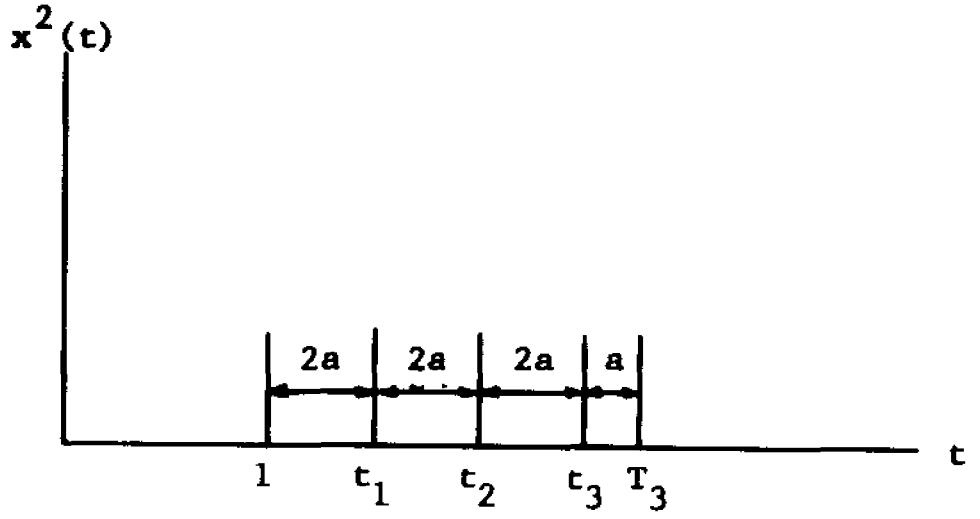


Fig. 3-4 Placement of Points for 3-Point Approximation

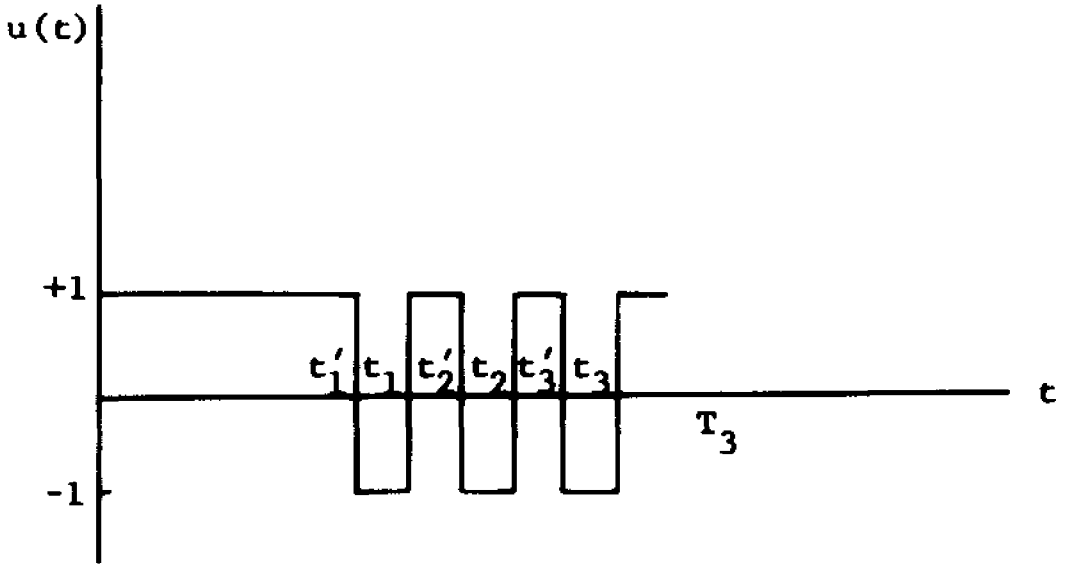


Fig. 3-5 Optimal Control for 3-Point Approximation

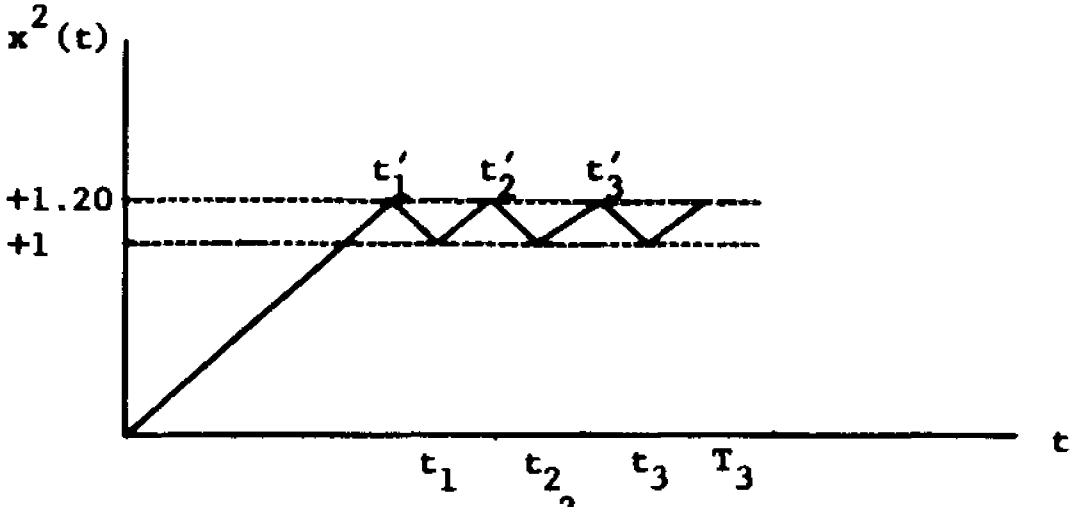


Fig. 3-6 Plot of x^2 for 3-Point Approximation

$$t_1' = 1.195 \qquad t_1 = 1.389$$

$$t_2' = 1.584 \qquad t_2 = 1.776$$

$$t_3' = 1.973 \qquad t_3 = 2.168$$

The corresponding trajectory for $x^2(t)$ is shown in Fig. 3-6 where

$$x^2(t_1') = 1.195$$

$$x^2(t_2') = 1.193$$

$$x^2(t_3') = 1.191$$

$$x^2(T_3) = 1.191$$

Therefore a 3-point approximation yields approximately a 20% violation of the output constraint. If this is an unacceptable violation we must add more constraint points. A five-point approximation with constraint points equidistantly placed as shown in Fig. 3-7, where $a = \frac{1}{11}(T_5 - 1)$ and T_5 is the optimal time for the 5-point approximation yields $T_5 = 2.41$ for which $x^1(2.41) = 2.003$. The optimal control is shown in Fig. 3-8 where

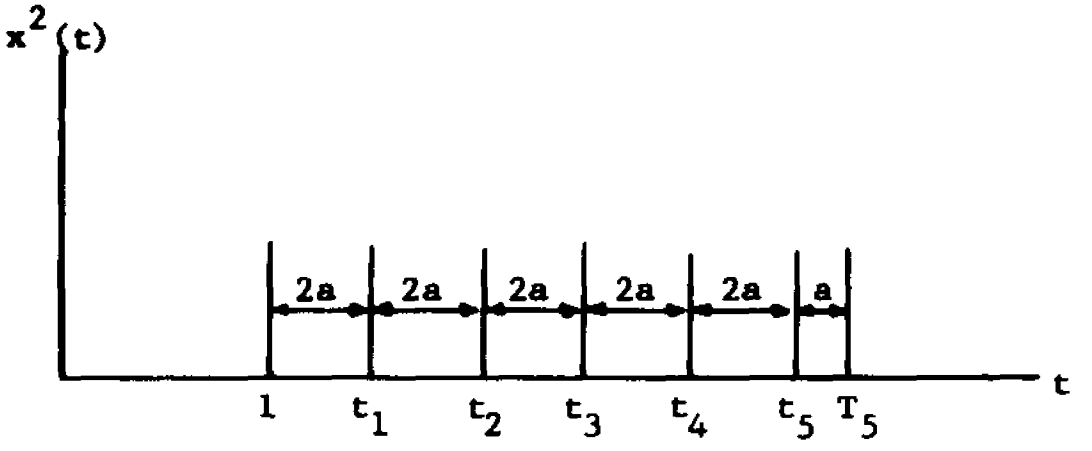


Fig. 3-7 Placement of Points for 5-Point Approximation

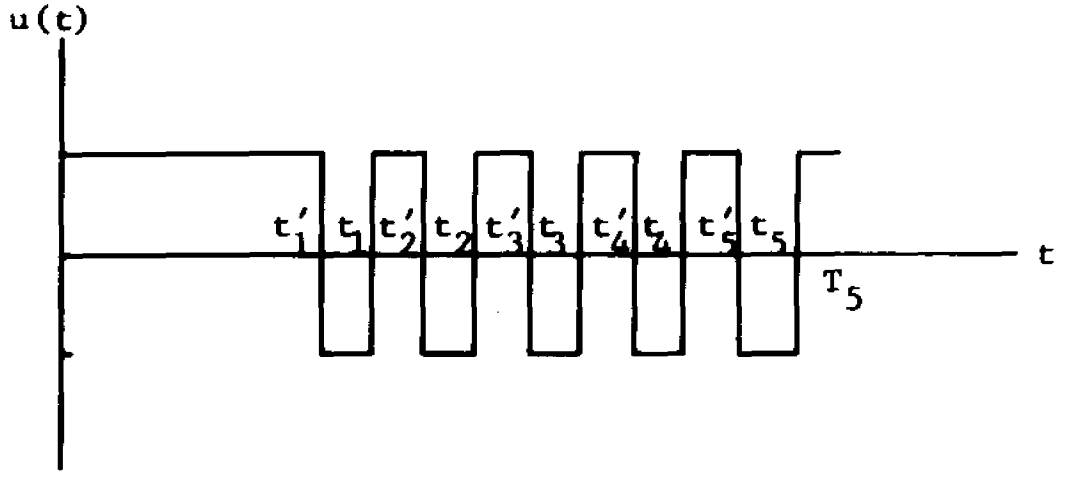


Fig. 3-8 Optimal Control for 5-Point Approximation

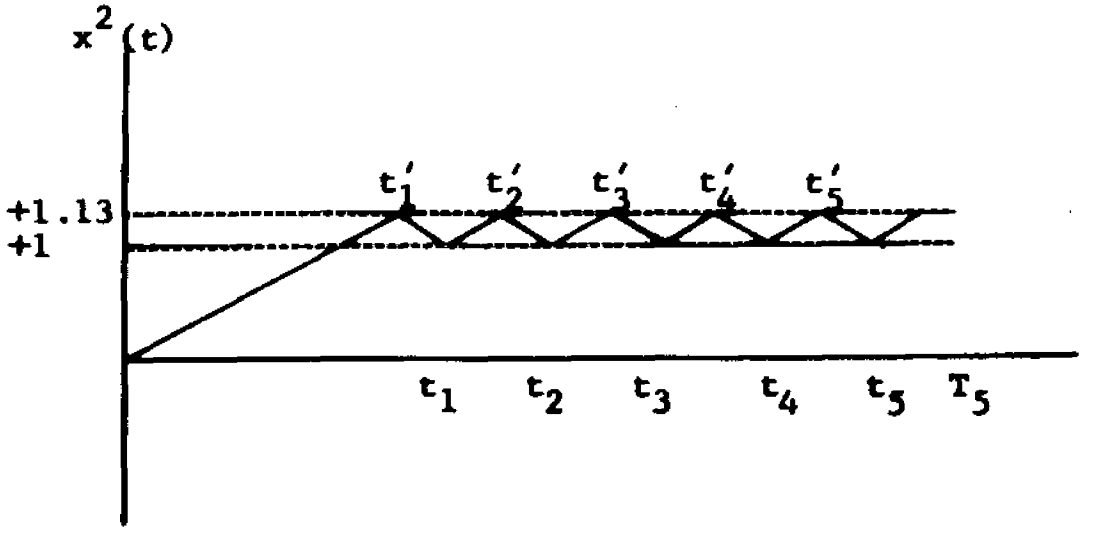


Fig. 3-9 Plot of x^2 for 5-Point Approximation

$t_1' = 1.128$ $\quad$ $t_1 = 1.257$

$t_2' = 1.386$ $\quad$ $t_2 = 1.514$

$t_3' = 1.642$ $\quad$ $t_3 = 1.770$

$t_4' = 1.899$ $\quad$ $t_4 = 2.027$

$t_5' = 2.156$ $\quad$ $t_5 = 2.284$

The corresponding trajectory for $x^2(t)$ is shown in Fig. 3-9 where

$x^2(t_1') = 1.128$ $\quad$ $x^2(t_4') = 1.129$

$x^2(t_2') = 1.128$ $\quad$ $x^2(t_5') = 1.127$

$x^2(t_3') = 1.128$ $\quad$ $x^2(T_5) = 1.127$

We have therefore reduced the output constraint violation to approximately 13% using a 5-point approximation.

If we now use a 5-point approximation for which the output is constrained to be equal or less than 0.8 at the five discrete points shown in Fig. 3-10, where $a = \frac{1}{11}(T_5 - 0.8)$, we obtain a transfer time (T_5) of 2.70 at which $x^1(2.70) = 2.003$. The optimal control and trajectory are similar to those shown in Figs. 3-8 and 3-9, respectively, except that in the present case we have no output constraint violation. Since $T = 2.41$ is a lower bound

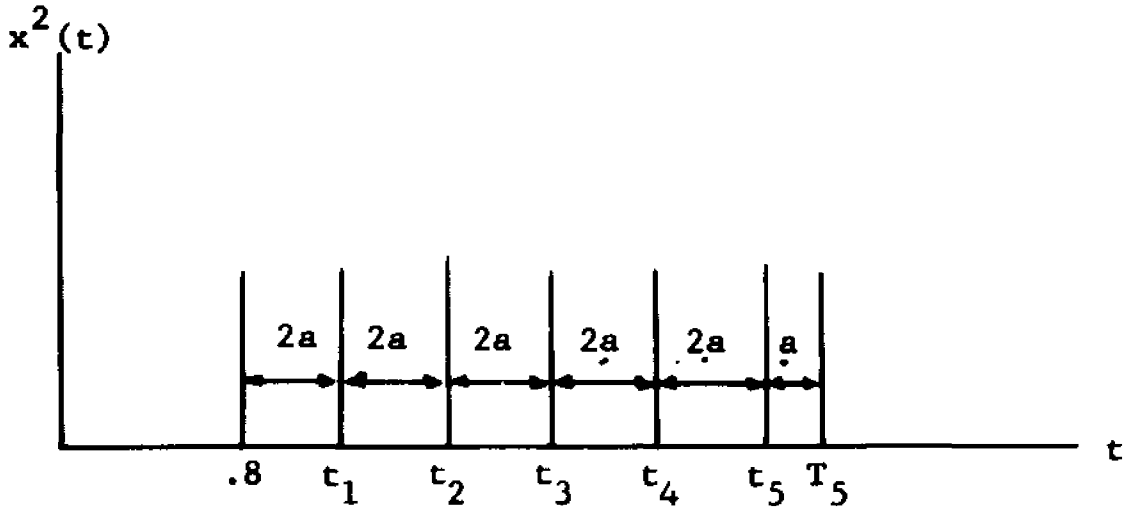


Fig. 3-10 Placement of Points for 5-Point Approximation when Constraint Is Enforced at 0.8 Level

on the optimal time of the exact solution, we see that the only penalty for using this approximation is a maximum increase of about 12% in the transfer time. It is emphasized that this information is obtained without any knowledge of the exact optimal solution.

For the next example consider the linear system specified by the dynamical equations

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -0.5x_2 + 0.5x_3$$

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

$$x^1 = x_1 + 2x_2 - 1.5x_3$$

$$x^2 = x_3$$

We again consider the problem of driving x^1 to the value 2 in the least time subject to the initial conditions $x_1(0) = x_2(0) = x_3(0) = 0$ and constraints as specified in the preceding example. The input output relations are

$$x^1(t) = \int_0^t \left[t - \tau + 2\left(1 - e^{-0.5(t-\tau)}\right) - 1.5 \right] u(\tau)\, d\tau$$

$$x^2(t) = \int_0^t u(\tau)\, d\tau$$

Obtaining the solution to the same problem without output constraint we find that the optimal time is $T = 2.17$, the optimal control is as shown in Fig. 3-11 and the corresponding trajectory for $x^2(t)$ is shown in Fig. 3-12.

A two-point approximation to the problem with the constraint level set at 0.85 and constraint points at $t_1 = \frac{1.1}{2.3} T$, $t_2 = \frac{1.4}{2.3} T$ yields a transfer time of 2.29 for which $x^1(2.29) = 1.994$. The optimal control is shown in Fig. 3-13, where

$$t_1' = 0.973 \qquad t_1 = 1.094$$

$$t_2' = 1.243 \qquad t_2 = 1.392$$

$$t_3' = 1.463$$

The corresponding trajectory for $x^2(t)$ is shown in Fig. 3-14, where

$$x^2(t_1') = 0.973 \qquad x^2(t_3') = 0.919$$

$$x^2(t_2') = 0.997 \qquad x^2(2.29) = 0.153$$

It is again noticed that there is no violation of the 1.00 constraint level on $x^2(t)$. If we run the same problem with the discrete points constrained at a 1.00 level, the transfer time obtained is 2.22 so that by using the

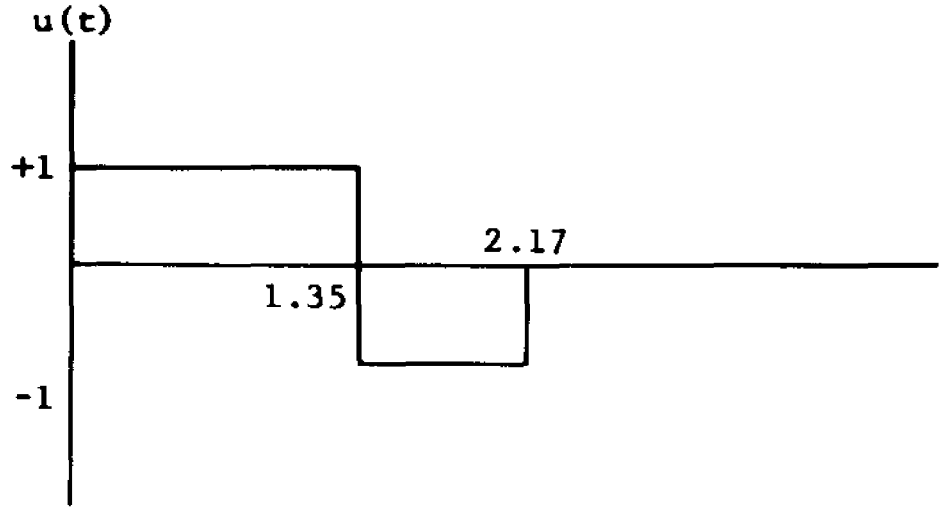


Fig. 3-11 Optimal Control for Second Example without Output Constraint

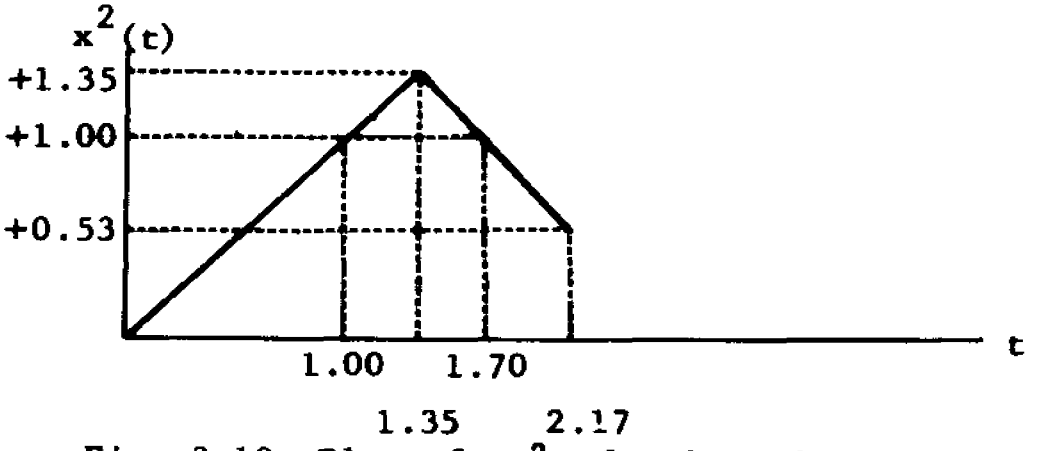


Fig. 3-12 Plot of x^2 for Second Example without Output Constraint

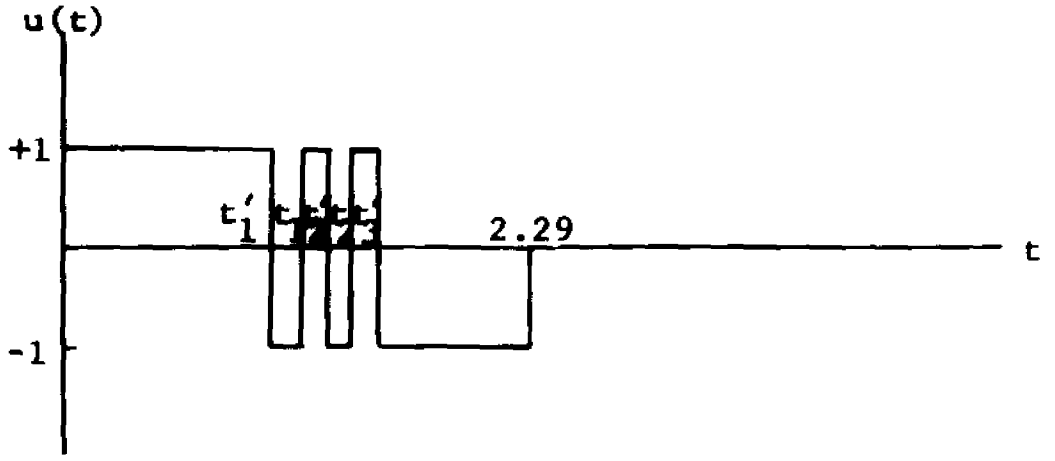


Fig. 3-13 Optimal Control for 2-Point Approximation

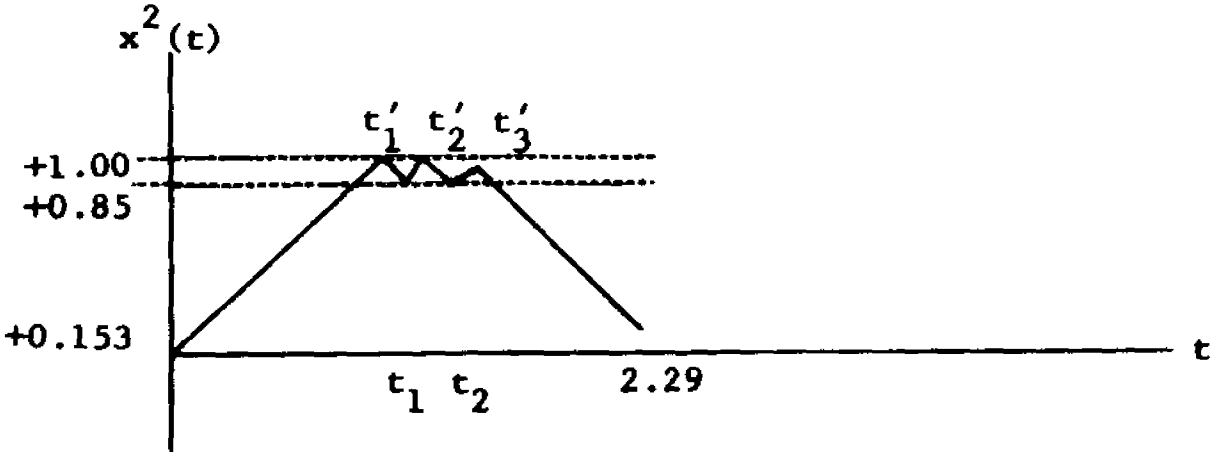


Fig. 3-14 Plot of x^2 for 2-Point Approximation

approximation with points constrained at the 0.85 level the only penalty is a maximum increase of about 3% in the transfer time.

3.6 General Case

In this section we treat an m-output plant with terminal constraints on r_1 of its outputs and amplitude constraints on r_2 of its outputs where r_1 and r_2 are both equal or less than m. Although a single input plant is again considered for convenience, multi-input plants can be handled using the multi-norm procedure of Sarachik and Kranc.[29]

The input output relations of the plant are

$$\underline{x}(t) = \underline{x}_o(t) + \int_0^t H(t,\tau)u(\tau)d\tau$$

where $\underline{x}(t)$ is the m-component output vector, $\underline{x}_o(t)$ is an m-vector expressing the effect on the output of initial conditions in the plant at $t = 0$ and is assumed to be known and continuous, and $H(t,\tau)$ is an n-vector whose components are assumed to be bounded piecewise continuous functions of their arguments.

The problem is to find the smallest T such that

$$\begin{aligned} x^{i_1}(T) &= x_d^{i_1} \\ &\vdots \\ x^{i_{r_1}}(T) &= x_d^{i_{r_1}} \end{aligned} \qquad 1 \le i_1 < i_2 < \ldots < i_{r_1} \le m \qquad (3\text{-}37)$$

where $x_d^{i_1}, \ldots x_d^{i_{r_1}}$ are desired final values for the outputs, $x^{i_1}(t), \ldots x^{i_{r_1}}(t)$ respectively. There is a constraint on the control which is given by

$$\|u\|_p = \left[\int_0^T |u(t)|^p dt\right]^{1/p} \le C_1 \qquad 1 < p \le \infty \qquad C_1 > 0 \quad (3\text{-}38)$$

and amplitude constraints at every instant of time on r_2 of the outputs

$$\begin{aligned} |x^{j_1}(t)| &\le C_{j_1} \\ &\vdots \\ |x^{j_{r_2}}(t)| &\le C'_{j_{r_2}} \end{aligned} \qquad 0 \le t \le T \quad 1 \le j_1 < j_2 < \ldots < j_{r_2} \le m \quad (3\text{-}39)$$

$$C'_{j_1}, \ldots, C'_{j_{r_2}} > 0$$

As in Sections 3.1 and 3.2 we enforce the output constraint only for some countable dense set of times $\{t_k\}$ in

the transition period, tentatively fix T and ask for the control function $u(t)$ which makes

$$\max \left\{ \frac{1}{C_1} \|u\|_p, \frac{1}{C_{j_1}} \sup_k |x^{j_1}(t_k)|, \ldots, \right. \tag{3-40}$$

$$\left. \frac{1}{C_{j_{r_2}}} \sup_k |x^{j_{r_2}}(t_k)| \right\} = \text{minimum}$$

while maintaining

$$\begin{aligned} x_o^{i_1}(T) + \int_0^T h_{i_1}(T,\tau)u(\tau)d\tau &= x_d^{i_1} \\ &\vdots \\ x_o^{i_{r_1}}(T) + \int_0^T h_{i_{r_1}}(T,\tau)u(\tau)d\tau &= x_d^{i_{r_1}} \end{aligned} \tag{3-41}$$

where $h_k(T,\tau)$ is the k^{th} component of $H(T,\tau)$. The smallest T for which the minimum in (3-40) is equal or less than unity is the optimal time.

The product space $\bar{L}_1$ now consists of composite vectors $\bar{y}$ of the form

$$\bar{y} = \left\{ y(t), \underline{a}^1, \underline{a}^2, \ldots, \underline{a}^{r_2} \right\} \tag{3-42}$$

where $y(t)$ is a function in $L_q[0,T]$ and $\underline{a}^j = [a_1^j, a_2^j, \ldots]$ $j = 1, 2, \ldots, r_2$, are vectors in ℓ_1 with norm

$$\|\underline{a}^j\|_{\ell_1} = \sum_{i=1}^{\infty} |a_i^j|$$

The norm of a composite vector in $\bar{L}_1$ is defined as

$$\|\bar{y}\|_{\bar{L}_1} = C_1\|y\|_q + \sum_{k=1}^{r_2} C'_{j_k} \|\underline{a}^k\|_{\ell_1} \tag{3-43}$$

and with this norm $\bar{L}_1$ is a Banach space.

Any bounded linear functional on $\bar{L}_1$ has the representation

$$f(\bar{y}) = \int_0^T y(t)u(t)dt + \sum_{i=1}^{\infty} a_i^1 b_i^1 + \ldots + \sum_{i=1}^{\infty} a_i^{r_2} b_i^{r_2} \tag{3-44}$$

with $u(t)$ in $L_p[0,T]$ and $\underline{b}^k = [b_1^k, b_2^k, \ldots]$, $k = 1, \ldots, r_2$, a vector with bounded components.

The norm of the functional defined by (2-13) is

$$\|f\| = \max\left\{\frac{1}{C_1}\|u(t)\|_p, \frac{1}{C'_{j_1}} \sup_i |b_i^1|, \ldots, \frac{1}{C'_{j_{r_2}}} \sup_i |b_i^{r_2}|\right\} \tag{3-45}$$

If we construct the following vectors in $\bar{L}_1$

$$\begin{aligned}
\bar{\varphi}_{i_1} &= \left\{h_{i_1}(T,\tau),\ \underline{0},\ \ldots,\ \underline{0}\right\} \\
&\vdots \qquad\qquad\qquad\qquad 0 = [0,0,\ldots] \\
\bar{\varphi}_{i_{r_1}} &= \left\{h_{i_{r_1}}(T,\tau),\ \underline{0},\ \ldots,\ 0\right\} \qquad\qquad (3\text{-}46) \\
\bar{\theta}_{k_{j_1}} &= \left\{h_{j_1}(t_k,\tau) - \underline{e}_k,\ \underline{0},\ \ldots,\ \underline{0}\right\} \\
&\vdots \qquad\qquad\qquad\qquad \underline{e}_k = [0,0,\ldots,1,\ldots,0] \\
\bar{\theta}_{k_{j_{r_2}}} &= \left\{h_{j_{r_2}}(t_k,\tau)\underline{0},\ \ldots,\ \underline{0} - \underline{e}_k\right\} \quad k = 1,\ 2,\ \ldots
\end{aligned}$$

where the one occurring in $\underline{e}_k$ is in the k^{th} column and the vector $-\underline{e}_k$ occurs in the $p+1^{st}$ column of $\bar{\theta}_{k_p}$, $p = 1,\ldots,r_2$, and require that a functional f of the form (3-44) satisfy

$$\begin{aligned}
f(\bar{\varphi}_{i_1}) &= x_d^{i_1} - x_o^{i_1}(T) \\
&\vdots \\
f(\bar{\varphi}_{i_{r_1}}) &= x_d^{i_{r_1}} - x_o^{i_{r_1}}(T) \\
f(\bar{\theta}_{k_{j_1}}) &= -x_o^{j_1}(t_k) \\
&\vdots \qquad\qquad\qquad\qquad k = 1,\ 2,\ \ldots \qquad (3\text{-}47) \\
f(\bar{\theta}_{k_{j_{r_2}}}) &= -x_o^{j_2}(t_k)
\end{aligned}$$

then the norm of f (the u(t) component of f satisfies the terminal conditions) as given by (3-45) becomes

$$\|f\| = \left\{ \frac{1}{C_1} \|u(t)\|_p, \frac{1}{C'_{j_1}} \sup_k |x^{j_1}(t_k)|, \ldots, \right. \tag{3-48}$$

$$\left. \frac{1}{C'_{j_{r_2}}} \sup_k |x^{j_{r_2}}(t_k)| \right\}$$

which is just the left side of (3-40).

A straightforward application of the methods of Chapter 2 yields that the functional of minimum norm satisfying (3-47) (which has as its u(t) component the control satisfying (3-40) and (3-41)) has its norm equal to

$$\|f\| = \frac{1}{\inf_{\substack{\lambda_1,\ldots,\lambda_{r_1} \\ \gamma^1_{k_1},\ldots,\gamma^{r_2}_{k_{r_2}} \\ n_1,\ldots n_{r_2}}} \left\{ C_1 \left[\int_0^T |\lambda_1 h_{i_1}(T,\tau) + \ldots \lambda_{r_1} h_{i_{r_1}}(T,\tau) + B|^q d\tau \right]^{1/q} + D \right.} \tag{3-49}$$

subject to

$$\sum_{p=1}^{r_1} \lambda_p c_p + \sum_{k_1=1}^{n_1} \gamma^1_{k_1} d_{k_1} + \ldots \sum_{k_{r_2}=1}^{n_{r_2}} \gamma^{r_2}_{k_{r_2}} d_{k_{r_2}} = 1$$

where

$$B = \sum_{k_1=1}^{n_1} \gamma_{k_1}^{1} h_{j_1}(t_{k_1},\tau) + \ldots \sum_{k_{r_2}=1}^{n_{r_2}} \gamma_{k_{r_2}}^{r_2} h_{j_{r_2}}(t_{k_{r_2}},\tau)$$

$$D = C'_{j_1} \sum_{k_1=1}^{n_1} |\gamma_{k_1}^{1}| + \ldots + C'_{j_{r_2}} \sum_{k_{r_2}=1}^{n_{r_2}} |\gamma_{k_{r_2}}^{r_2}|$$

$$c_p = x_d^{i_p} - x_o^{i_p}(T) \qquad p = 1, \ldots, r_1$$

$$d_{k_1} = - x^{j_1}(t_{k_1})$$

$$\vdots$$

$$d_{k_{r_2}} = - x^{j_{r_2}}(t_{k_{r_2}})$$

The first value of T for which the right side of (3-49) is equal or less than unity is the optimal time.

As an approximation to the infinite dimensional minimization required in (3-49) we pick fixed values of $n_1, \ldots, n_{r_2}$, and n_1 fixed values of $t_{k_1} = f_{k_1}^{1} T, \ldots n_{r_2}$ fixed values of $t_{k_{r_2}} = f_{k_{r_2}}^{r_2} T$ where the $f_{k_p}^{j}$ are real numbers in $[0,1]$ and then assume all $\gamma_{k_p}^{j}$ equal 0 except for k_p corresponding to the chosen values of t_{k_p}.

The minimization is then performed over $\lambda_1, \ldots, \lambda_{r_1}$ and the appropriate $\gamma^j_{k_p}$ to obtain

$$\min_{\substack{\lambda_1,\ldots,\lambda_{r_1} \\ \gamma^1_{k_1},\ldots,\gamma^{r_2}_{k_{r_2}}}} \left\{ C_1 \left[\int_0^T |\lambda_1 h_{i_1}(T,\tau) + \lambda_{r_1} h_{i_{r_1}}(T,\tau) + F|^q d\tau \right]^{1/q} + G \right\} \quad (3\text{-}50)$$

subject to

$$\sum_{p=1}^{r_1} \lambda_p c_p + \sum_{k_1=1}^{n_1} \gamma^1_{k_1} d_{k_1} + \ldots \sum_{k_{r_2}=1}^{n_{r_2}} \gamma^{r_2}_{k_{r_2}} d_{k_{r_2}} = 1$$

where

$$F = \sum_{k_1=1}^{n_1} \gamma^1_{k_1} h_{j_1}(f^1_{k_1} T,\tau) + \ldots \sum_{k_{r_2}=1}^{n_{r_2}} \gamma^{r_2}_{k_{r_2}} h_{j_{r_2}}(f^{r_2}_{k_{r_2}} T,\tau)$$

$$G = C'_{j_1} \sum_{k_1=1}^{n_1} |\gamma_{k_1}| + \ldots C'_{j_{r_2}} \sum_{k_{r_2}=1}^{n_{r_2}} |\gamma_{k_{r_2}}|$$

and we then assume the finite dimensional minimization (3-50) equals the denominator of the right side of (3-49).

This approximation yields the solution to the problem of finding the smallest T and the corresponding control for which the terminal and control constraints and output constraints at the instants of time $t_{k_p} = f^j_{k_p} T$ are satisfied. The solution to this problem is then used as the approximate solution for the original problem.

CHAPTER 4. A SEQUENTIAL APPROXIMATION METHOD

In the previous chapter we saw that each additional point used in a discrete point approximation adds another dimension to the finite dimensional minimization which must be performed to obtain the approximate solution. This leads to the drawback that for some systems we may be forced to perform a minimization of large dimension in order to keep the output violation within reasonable limits. In this chapter we consider the terminal control problem and show in Section 4.4 how it may be applied in a sequential scheme for time invariant systems in order to circumvent the problem of large dimensional minimization.

4.1 Problem Statement — Terminal Control Problem

We consider a double output single input plant for convenience and again the results extend easily to multi-input, multi-output plants as shown in Section 4.5. The input output relations of the plant are

$$x^1(t) = x_o^1(t) + \int_0^t h_1(t,\tau)u(\tau)d\tau$$

$$x^2(t) = x_o^2(t) + \int_0^t h_2(t,\tau)u(\tau)d\tau \qquad (4\text{-}1)$$

where $x^1(t)$, $x^2(t)$ are outputs of the plant, $x_o^1(t)$, $x_o^2(t)$ express the effect of initial conditions in the plant at $t = 0$ and are assumed known, and $u(t)$ is the control. The functions $h_1(t,\tau)$ and $h_2(t,\tau)$ are assumed to be piecewise continuous functions of τ on the interval $[0,t]$ and equal 0 for $\tau > t$ for all finite values of t.

The problem considered is that of finding the control on the fixed time interval $[0,T]$ which minimizes

$$|x^1(T) - x_d^1| \tag{4-2}$$

where x_d^1 is some desired final value, subject to a constraint on the control of the form

$$\|u\|_p = \left(\int_0^T |u(t)|^p dt\right)^{1/p} \leq C_1 \qquad 1 < p \leq \infty \tag{4-3}$$

and an amplitude constraint on the output $x^2(t)$ at n discrete instants of time $t_i = r_i T$, $r_i \in [0,1]$, $i = 1,\ldots,n$,

$$|x^2(t_i)| \leq C_2 \qquad i = 1, \ldots, n \tag{4-4}$$

where C_1 and C_2 are positive constants.

This problem can be reformulated by choosing some fixed value δ and asking for the control function $u(t)$ which makes

$$\max\left\{\frac{1}{C_1}\|u\|_p,\ \frac{1}{C_2}\max|x^2(t_i)|,\ \frac{1}{\delta}|x^1(T) - x_1^d|\right\} = \text{minimum} \qquad (4\text{-}5)$$

The smallest δ for which the minimum in (4-5) is unity then solves the original problem with constraints (4-3) and (4-4). As in Chapter 2, we construct a suitable Banach space so that the variational problem (4-5) may be formulated as an L-problem in the theory of moments.

4.2 L-Problem Formulation

Define the parameter q conjugate to p by

$$\frac{1}{p} + \frac{1}{q} = 1$$

and for functions $y(t)$ in $L_q[0,T]$, n-tuples $\underline{a} = [a_1, \ldots a_n]$ in $\ell_1^{(n)}$ and scalars b, define a composite vector $\bar{y}$ by

$$\bar{y} = \left\{y(t),\ a,\ b\right\} \qquad (4\text{-}6)$$

If the norm of a function in $L_q[0,T]$ is defined as

$$\|y\|_q = \left(\int_0^T |y(t)|^q dt\right)^{1/q} \qquad (4\text{-}7)$$

and the norm of a vector in $\ell_1^{(n)}$ by

$$\|\underline{a}\|_{\ell_1^{(n)}} = \sum_{i=1}^{n} |a_i| \qquad (4\text{-}8)$$

a norm $\left(\| \ \|_{\bar{L}_1}\right)$ can then be introduced for composite vectors $\bar{y}$ of the form (4-6) by defining

$$\|\bar{y}\|_{\bar{L}_1} = C_1\|y\|_q + C_2\|\underline{a}\|_{\ell_1^{(n)}} + \delta|b| \tag{4-9}$$

where $|b|$ is the ordinary absolute value of the scalar b. With the norm defined by (4-9) the composite vectors $\bar{y}$ form a Banach space which we denote by $\bar{L}_1$.

The following can now be shown:

1. Any bounded linear functional f on $\bar{L}_1$ has a representation

$$f(\bar{y}) = \int_0^T y(t)u(t)dt + \sum_{i=1}^{n} a_i c_i + bd \tag{4-10}$$

with $u(t)$ in $L_p[0,T]$, $c = [c_1, \ldots, c_n]$ an n-tuple with finite components and d an ordinary scalar,

2. The norm of the functional f of (4-10) defined as

$$\|f\| = \sup_{\bar{y} \neq 0} \frac{f(\bar{y})}{\|\bar{y}\|_{\bar{L}_1}} \tag{4-11}$$

is given by

$$\|f\| = \max\left\{\frac{1}{C_1}\|u\|_p \ , \ \frac{1}{C_2}\max_i |c_i| \ , \ \frac{1}{\delta}|d|\right\} \tag{4-12}$$

To apply these developments to our problem (4-5) we recognize that the functions $h_1(T,\tau)$ and $h_2(t_i,\tau)$ being piecewise continuous in $[0,T]$ are in $L_q[0,T]$, so that we may construct from them composite vectors of the form (4-6)

$$\bar{\varphi} = \left\{h_1(T,\tau),\ \underline{0},\ -1\right\} \qquad \underline{0} = [0,\ldots,0]$$

$$\bar{\theta}_k = \left\{h_2(t_k,\tau),\ -\underline{e}_k, 0\right\} \qquad \underline{e}_k = [0,\ldots,1,\ldots,0] \quad k = 1,\ \ldots,n \tag{4-13}$$

where the one occurring in $\underline{e}_k$ is in the k^{th} column. If now we require that a functional f of the form (4-10) satisfy

$$f(\bar{\varphi}) = x_d^1 - x_o^1(T) \tag{4-14}$$

then

$$x_d^1 - x_o^1(T) = \int_0^T h_1(T,\tau)u(\tau)d\tau - d$$

$$d = x_o^1(T) + \int_0^T h_1(T,\tau)u(\tau)d\tau - x_d^1 \tag{4-15}$$

$$d = x^1(T) - x_d^1$$

or the d component of the functional f is equal to $x^1(T) - x_d^1$ where $x^1(T)$ is the terminal value of x^1 when the u(t) component of f is applied as the system input. If further we require that

$$f(\bar{\theta}_k) = - x_o^2(t_k) \qquad k = 1, \ldots, n \tag{4-16}$$

we are requiring that the c_k component of f be given by

$$b_k = x_o^2(t_k) + \int_0^T h_2(t_k,\tau)u(\tau)d\tau = x^2(t_k) \qquad k = 1, \ldots, n \tag{4-17}$$

where the last of the equalities in (4-17) follows from (4-1) and the fact that $h_2(t_k,\tau) = 0$ for $\tau > t_k$. The norm of f, as given by (4-12) then becomes

$$\|f\| = \left\{\frac{1}{C_1}\|u\|_p \ , \ \frac{1}{C_2}\max_k|x^2(t_k)| \ , \ \frac{1}{\delta}|x^1(T) - x_d^1|\right\} \tag{4-18}$$

which is the left side of (4-5).

The abstract problem of finding the bounded linear functional of minimum norm on $\bar{L}_1$ which maps the given elements $\bar{\varphi}$ and $\bar{\theta}_k$, k = 1, ..., n of the space into the given fixed scalars specified by (4-14), (4-16) is equivalent to solving the variational problem (4-5) and the least value of δ for which this functional has its norm equal to unity is the smallest distance between $x^1(T)$ and x_d^1

that can be attained by a control satisfying the problem constraints (4-3) and (4-4).

On physical grounds, it seems that the stated problem should always be well posed and that this is indeed the case follows from the fact that the vectors $\bar{\varphi}$ and $\bar{\theta}_k$, $k = 1, \ldots, n$ are linearly independent. Therefore they span an $n + 1$ dimensional linear space contained in $\bar{L}_1$ and we can define a linear functional f_1 on this space satisfying (4-14), (4-16) by

$$f_1\left(\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k \bar{\theta}_k\right) = \alpha f_1(\bar{\varphi}) + \sum_{k=1}^{n} \gamma_k f_1(\bar{\theta}_k) = \alpha c + \sum_{k=1}^{n} \gamma_k d_k \qquad (4\text{-}19)$$

where

$$c = x_d^1 - x_o^1(T) \qquad d_k = -x_o^2(t_k) \qquad (4\text{-}20)$$

The norm of f_1 on this $n + 1$ dimensional linear space is found from (4-11)

$$\|f\| = \sup_{\substack{\alpha,\gamma_k \\ \text{not all} = 0}} \frac{|f_1(\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k\bar{\theta}_k)|}{\|\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k\bar{\theta}_k\|_{\bar{L}_1}}$$

$$= \sup_{\substack{\alpha,\gamma_k \\ \text{not all} = 0}} \frac{|\alpha c + \sum_{k=1}^{n} \gamma_k d_k|}{\|\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k\bar{\theta}_k\|_{\bar{L}_1}} \qquad (4\text{-}21)$$

$$= \frac{1}{\inf_{\alpha,\gamma_k} \|\alpha\bar{\varphi} + \sum_{k=1}^{n} \gamma_k\bar{\theta}_k\|_{\bar{L}_1}}$$
$$\alpha c + \sum_{k=1}^{n} \gamma_k d_k$$

or using the explicit form of the norm in $\bar{L}_1$ given by (4-7) through (4-9)

$$\|f_1\| = \frac{1}{\inf_{\alpha,\gamma_k} \left\{C_1\left[\int_0^T |B(\tau)|^q d\tau\right]^{1/q} + C_2 \sum_{k=1}^{n} |\gamma_k| + \delta|\alpha|\right\}} \qquad (4\text{-}22)$$
$$\alpha c + \sum_{k=1}^{n} \gamma_k d_k$$

where

$$B(\tau) = \alpha h_1(T,\tau) + \sum_{k=1}^{n} \gamma_k h_2(t_k,\tau)$$

and the infimum is evaluated over all real values of α and γ_k, $k = 1, \ldots, n$. The Hahn-Banach theorem asserts that the linear functional f_1 defined by (4-19) can be extended to a linear functional f defined over L_1 such that $\|f\| = \|f_1\|$. Since any extension of f_1 over $\bar{L}_1$ must have its norm equal or greater than that of f_1 we therefore infer the existence of a functional of minimum norm satisfying (4-14), (4-16) which has its norm equal to (4-22). The smallest value of δ for which (4-22) equals unity is the smallest admissible value of (4-2) hereafter called δ_o.

4.3 Form of the Control

The value of δ_o has just been found and we now turn to a derivation of the form of the control which transfers $x^1(T)$ to within δ_o of x_d^1. Let the infimum of (4-22) ($\delta = \delta_o$) be attained at the values $\alpha = \alpha^*$, $\gamma_k = \gamma_k^*$, $k = 1, \ldots, n$, with $\alpha^* c + \sum_{k=1}^{n} \gamma_k^* d_k = 1$. Also, let f be any functional of minimum norm satisfying (4-14), (4-16). From the preceding section, there exists at least one, its norm equals unity and the $u(t)$ component of the functional

is the desired control. Following the same reasoning as in (2-25) through (2-27) we obtain

$$|f(\alpha^{*}\bar{\varphi} + \sum_{k=1}^{n} \gamma_k^{*}\bar{\theta}_k)| = \|f\|\|\alpha^{*}\bar{\varphi} + \sum_{k=1}^{n} \gamma_k^{*}\bar{\theta}_k\|_{\bar{L}_1} \tag{4-23}$$

or more explicitly using (4-7) through (4-12) and letting

$$\alpha^{*}h_1(T_o,\tau) + \sum_{k=1}^{n} \gamma_k^{*}h_2(t_k,\tau) = \sigma^{*}(\tau) \tag{4-24}$$

we obtain

$$\left|\int_0^{T_0} \sigma^{*}(\tau)u(\tau)d\tau + \sum_{k=1}^{n} \gamma_k^{*}c_k + \alpha^{*}d\right| = \tag{4-25}$$

$$\max\left\{\frac{1}{C_1}\|u\|_p,\ \frac{1}{C_2}\max_k|c_k|,\ \frac{1}{\delta_o}|d|\right\}\left\{C_1\|\sigma^{*}\|_q + C_2\sum_{k=1}^{n}|\gamma_k^{*}| + \delta_o|\alpha^{*}|\right\}$$

Repeating the arguments used in (2-29) through (2-40) we arrive at the form of the control

$$u(t) = C_1\|\sigma^{*}\|_q|\sigma^{*}(t)|^{q-1}\,\mathrm{sgn}[\sigma^{*}(t)] \tag{4-26}$$

(4-26) is the control on $[0,T]$ which transfers $x^1(T)$ to within δ_o of x_d^1.

4.4 A Sequential Approximation Method for Time Invariant Systems

A limiting factor in the use of the discrete point approximation of Chapter 3 is the dimensionality of the minimization which must be performed to obtain its solution. Large dimensionality requires large computer storage capacity, increases computational time, etc. Therefore, in most practical cases, there is an upper limit on the dimensionality of the minimization; let us denote this limit by n. In those cases where, in order to obtain a satisfactory discrete point approximation for the time optimal problem, it is necessary to constrain the output at more than n discrete points it may still be possible, by making use of the terminal control problem, to obtain an acceptable approximation without exceeding the limits on the dimensionality of the minimization. This may be done by decomposing the problem into two parts, first fixing the terminal time and using the terminal control problem to find the control which takes us as close as possible to the desired terminal point and then solving the time optimal problem from this intermediate point. If the allowable output constraint violation is denoted by ϵ, the terminal time is fixed at a value for which we can guarantee that the output violation on this interval is no larger than ϵ when we constrain the output at n

discrete points. This time can be found from an examination of the error bounds in Section 3.3 but in general is simply found from a trial and error procedure. In many cases it is possible to transfer the system from the intermediate point to the terminal point using a discrete point approximation containing at most n points which satisfies the allowable output constraint violation on the transition interval. In these cases, since the system is time invariant, we can combine the solutions to both problems to obtain an approximation which satisfies the allowable output constraint violation and yet circumvent the problem of having to perform any minimizations of dimension larger than n.

It is noted that even though there is an output constraint violation it cannot be established, as in the previous chapter, that the transfer time for this approximation is less than or at most equal to the transfer time of the exact solution. However, since the terminal time of the first problem is less than or at most equal to the true optimal time, the maximum increase in transfer time using this method of approximation is the transfer time of the second problem. Then, although for some problems a decomposition technique of this type may not give good results, if this latter transfer time is small we are assured that our approximation is a good one.

As shown in the example below, even when a discrete point approximation can be generated it may be possible to use the above approach to obtain an even better approximation. This occurs because we are able to constrain the output at twice as many points as in the discrete point approximation.

It is a straightforward matter to extend this approach to the case where the decomposition consists of more than two problems, i.e., where we have a number of intermediate points. However, the larger the number of problems into which the original one is decomposed the less knowledge we have of the discrepancy between the transfer time of the approximate solution and the true optimal time.

We also remark that the terminal control problem can be considered to be a generalization of the time optimal problem in the sense that the first value of T for which the minimum value of (4-2) equals zero is the optimal time.

4.5 Generalizations

In this section we consider the generalization of the work in the preceding sections to include the case of multi-output plants with more than one terminal and one output constraint. We consider a single input, m-output plant with input output relations given by (3-36). The transition period is a fixed interval [0,T], the terminal conditions

are given by (3-37) and the control constraint by (3-28). There are amplitude constraints at n discrete instants of time, t_i, $i = 1, \ldots, n$ of the form

$$\begin{aligned} |x^{j_1}(t_i)| &\leq C'_{j_1} \qquad i = 1, \ldots, n \\ &\vdots \qquad\qquad 1 \leq j_1 < j_2 < \cdots < j_{r_2} \leq m \\ |x^{j_{r_2}}(t_i)| &\leq C'_{j_{r_2}} \qquad C'_{j_1}, \ldots, C'_{j_{r_2}} > 0 \end{aligned} \tag{4-27}$$

It is desired to find the control on the given interval $[0,T]$ which, subject to the given constraints, minimizes the quantity

$$\left[\sum_{k=1}^{r_1} |x^{i_k}(T) - x_d^{i_k}|^{\omega} \right]^{1/\omega} \qquad \omega \geq 1 \tag{4-28}$$

For different values of ω, we obtain different meanings of the closeness of the actual terminal state and the desired terminal state. For example, for $\omega = \infty$ we minimize the maximum deviation of any component of the terminal state from the desired value of that component while for $\omega = 2$ we minimize the sum of the squared deviations of the components of the terminal state from their desired final values.

This problem may be reformulated by choosing some fixed value of δ and asking for the control function which makes

$$\max\left\{\frac{1}{C_1}\|u\|_p,\ \frac{1}{C_{j_1}}\max_i|x^{j_1}(t_i)|,\ \ldots\right.$$

$$\left.\frac{1}{C_{j_{r_2}}}\max_i|x^{j_{r_2}}(t_i)|,\ \frac{1}{\delta}\left[\sum_{k=1}^{r_1}|x^{i_x}(T)-x_d^{i_k}|^{\omega}\right]^{1/\omega}\right\}=\text{minimum} \tag{4-29}$$

The smallest δ for which the minimum in (4-29) equals unity then solves the original problem with constraints (3-37) and (3-38).

The L-problem formulation of this variational problem is straightforward and its solution proceeds along the same lines as in the earlier sections of this chapter.

4.6 Example

Let us consider the double integrater problem of Section 3.5 with the additional restriction that we are not allowed to perform any minimizations of dimension larger than three.

Also, let us look for an approximation which yields no output constraint violation. If we use the discrete point approximation technique of Chapter 3 and constrain x^2 to be equal or less than 0.7 at the discrete points

$$t_1 = 0.7 + 2a \qquad t_2 = 0.7 + 4a \qquad t_3 = 0.7 + 6a$$

where $a = \frac{1}{7}(T - 0.7)$ then we obtain an approximation which satisfies all problem constraints but yields an increase of approximately 15% in the transfer time of this approximation over the true optimal time. We can obtain a considerable improvement in this respect if we use the sequential approximation technique developed in the preceding sections.

From examination of the error bounds in Section 3.3 it can be seen that by fixing the terminal time at $T = 1.75$ and constraining x^2 to be equal or less than 0.85 at the discrete points

$$t_1 = 1.15 \qquad t_2 = 1.45 \qquad t_3 = 1.75 ,$$

that no violation of the 1.00 level for x^2 can occur for $0 \leq t \leq 1.75$ when we minimize

$$|x^1(1.75) - 2.00| \tag{4-30}$$

After performing the minimization (4-22) with the appropriate substitutions, we find that the minimum value of (4-30) is 0.81 and $x^1(1.75) = 1.19$. The corresponding control and trajectory for $x^2(t)$ are shown in Figs. 4-1 and 4-2, respectively. We now proceed by finding the control which steers the plant from $x^1 = 1.19$, $x^2 = 0.85$ to $x^1 = 2.00$ in minimum time T subject to the control constraint

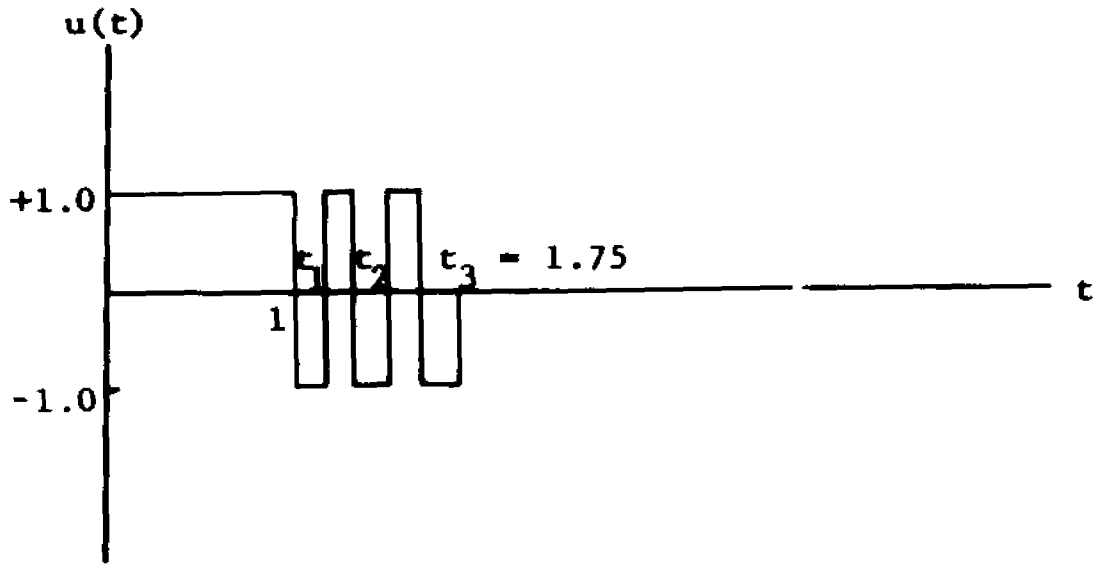


Fig. 4-1 Optimal Control for First Problem in Decomposition

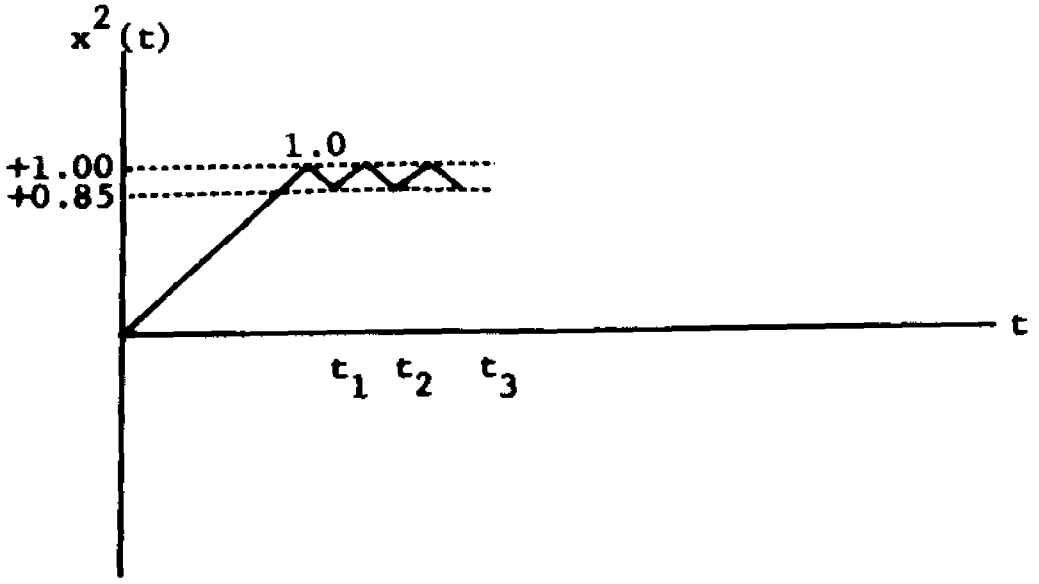


Fig. 4-2 Plot of x^2 for First Problem in Decomposition

$$|u(t)| \leq 1$$

and an output constraint on x^2 which we take as

$$|x^2(t_i)| \leq 0.85 \qquad i = 1, 2, 3$$

where

$$t_1 = \frac{2}{7} T \qquad t_2 = \frac{4}{7} T \qquad t_3 = \frac{6}{7} T$$

Performing the minimization (3-19) with the appropriate substitutions we obtain $T = 0.87$ at which $x^1(0.87) = 2.00$. The corresponding control and trajectory for $x^2(t)$ for this problem are shown in Figs. 4-3 and 4-4, respectively. Note that $x^2(t)$ does not violate the 1.00 level on this transition interval.

Combining both solutions to the above problems (the system is time invariant) we obtain a transfer time of $1.75 + 0.87 = 2.62$ which is less than 5% greater than the true optimal time of 2.5. The control and trajectory for $x^2(t)$ for this sequential approximation technique are shown in Figs. 4-5 and 4-6, respectively.

As a somewhat contrived example of a case in which the sequential approximation technique can generate a solution satisfying the problem constraints whereas the discrete point approximation method fails to do so, consider the same problem as above where now we are not permitted to perform any

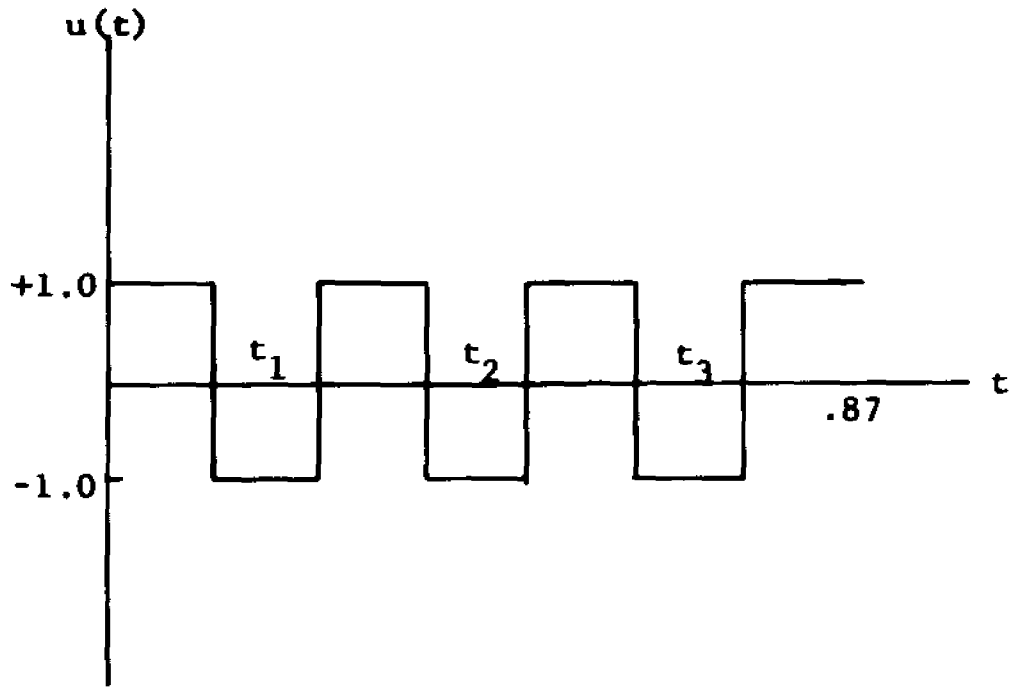


Fig. 4-3 Optimal Control for Second Problem in Decomposition

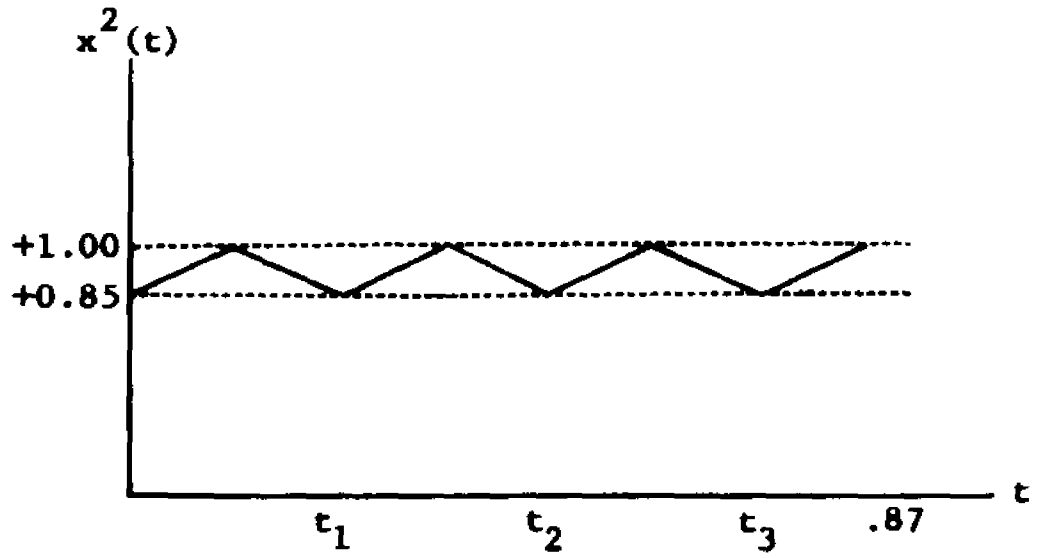


Fig. 4-4 Plot of x^2 for Second Problem in Decomposition

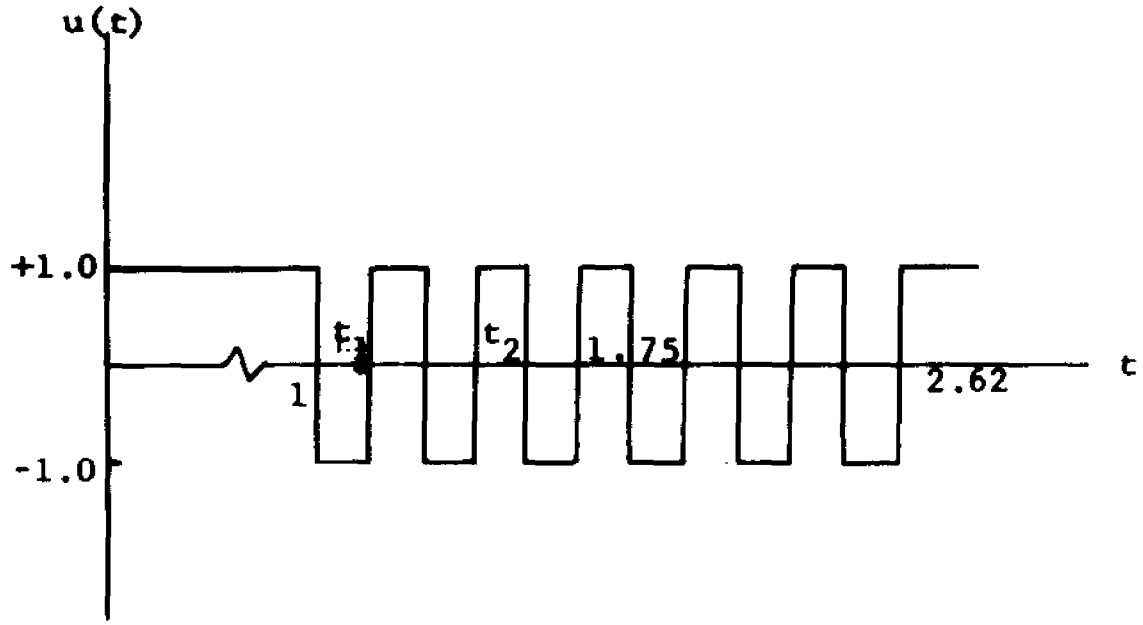


Fig. 4-5 Approximate Optimal Control Obtained from Sequential Approximation Method

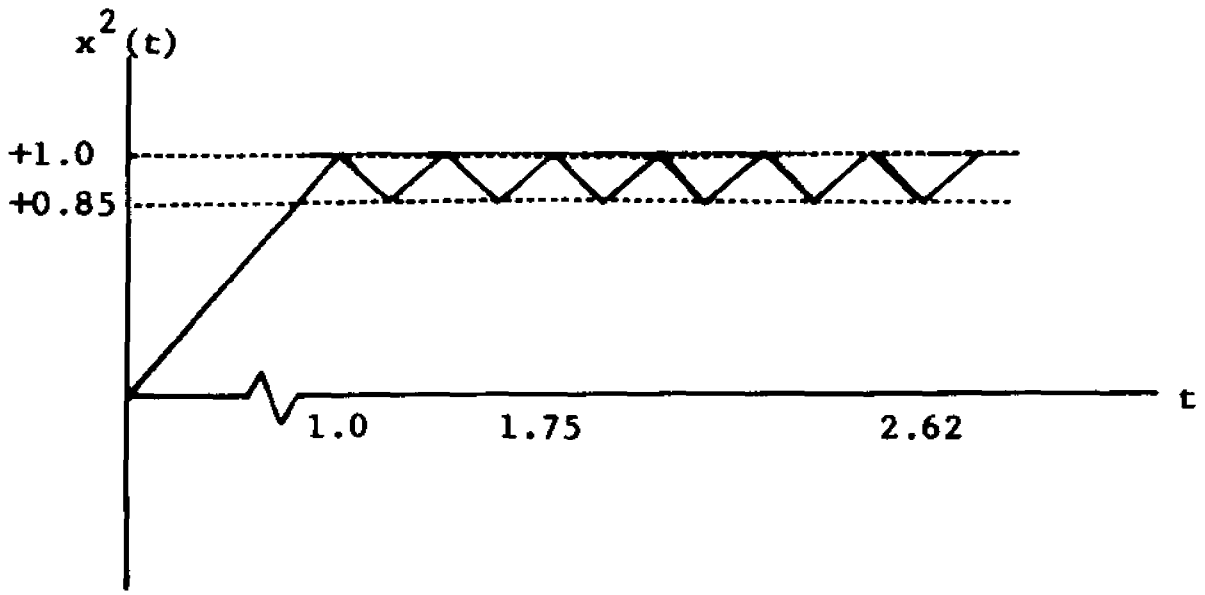


Fig. 4-6 Plot of x^2 for Sequential Approximation Method

minimizations of dimension larger than one. No discrete point approximation using one point exists which satisfies the terminal constraint yet yields no output constraint violation. However, by proceeding as in the above example it is a straightforward matter to generate a sequential approximation which satisfies all problem constraints.

CHAPTER 5. AN "APPROXIMATION TO THE CONTROL" SCHEME

In the preceding two chapters we discussed an approximation method for continuously bounded output problems whereby the problem itself was approximated, the approximate problem solved exactly and the control obtained then used as the approximate optimal control for the original problem. In this chapter we develop a method for directly obtaining an approximation to the optimal control by deriving a necessary condition for optimality and then trying to satisfy this condition as closely as possible in some given set of parameters.

5.1 Initially Quiescent Plants

In this section we restrict ourselves for ease of presentation to consideration of a double output-single input plant which is initially in a quiescent state. The form of solution when the plant has nonzero initial conditions is given in Section 5.5 and the extension to multi-input plants with multiple terminal and output constraints using the more general product spaces developed in Section 3.6 is straightforward.

The precise problem considered is as follows: the plant is described by the input output relations

$$x^1(t) = \int_0^t h_1(t,\tau)u(\tau)d\tau$$

$$x^2(t) = \int_0^t h_2(t,\tau)u(\tau)d\tau$$

where $x^1(t)$, $x^2(t)$ are outputs of the plant which are to be controlled by the single variable - $u(t)$. In order to ensure that all integrals are well defined, we assume that $h_1(t,\tau)$ and $h_2(t,\tau)$ are bounded piecewise continuous and continuous functions of their arguments, respectively. The results can be shown to hold when $h_2(t,\tau)$ has a simple discontinuity at $t = \tau$. We wish to find the smallest T such that

$$x^1(T) = x_d^1$$

where x_d^1 is some desired final value. There is an amplitude constraint on the control which can be expressed as

$$\|u\|_\infty \leq C_1$$

and a constraint on the output $x^2(t)$ at every instant of time

$$|x_2(t)| \leq C_2 \qquad 0 \leq t \leq T$$

As in Section 3.1, we formulate the equivalent problem wherein we only enforce the output constraint for some countable dense set of times in $[0,T]$ which in this case is taken to be the set $\left\{t_i = \frac{m}{n} T\right\}$ where m and n are arbitrary positive integers and $m \leq n$.

We consider the product space $\bar{L}_1$ defined in Section 3.2 and the elements $\bar{\varphi}$ and $\bar{\theta}_k$, $k = 1, 2, \ldots$ of $\bar{L}_1$ defined by (3-12). Again by requiring a linear functional f to satisfy

$$\begin{aligned} f(\bar{\varphi}) &= x_d^1 \\ f(\bar{\theta}_k) &= 0 \qquad k = 1, 2, \ldots \end{aligned} \tag{5-1}$$

it follows that the norm of f (the $u(t)$ component of f satisfies the terminal condition) is given by

$$f = \left\{\frac{1}{C_1} \|u\|_\infty \,,\ \frac{1}{C_2} \sup_k |x_2(t_k)|\right\}$$

The functional on $\bar{L}_1$ having minimum norm of those satisfying (5-1) has its norm given by (3-15) which for the present case $\left(x_o^1(t) = x_o^2(t) \equiv 0,\ p = \infty\right)$ is

$$\|f\| \quad \frac{1}{\inf_{\gamma_{m,n}} \left\{C_1 \left| \int_0^T \left| \frac{1}{x_d^1} h_1(T,\tau) + \sum_{m=1}^{n} \gamma_m h_2\left(\frac{m}{n} T, \tau\right) \right| d\tau \right| + C_2 \sum_{m=1}^{n} |\gamma_m| \right\}} \tag{5-2}$$

where the infimum is over n as well as the variables γ_m. The first value of T for which (5-2) is equal or less than unity is the optimal time, which we designate by T_o. We now fix $T = T_o$ and find a sequence of controls which converges in some sense to the optimal control.

A word is necessary to explain the circuitous method of deriving the optimal control. In the discrete point approximation, the form of the optimal control was derived in Section 2.3 by taking the values of α^* and γ_k^*, $k = 1,2,\ldots,N$ at which the infimum is attained in (2-24) and finding for the particular element

$$\bar{\sigma}^* = \alpha^* \bar{\varphi} + \sum_{k=1}^{n} \gamma_k^* \bar{\theta}_k \tag{5-3}$$

that

$$|f(\sigma^*)| = \|f\| \|\bar{\sigma}^*\|_{L_1} \tag{5-4}$$

As a consequence of (5-4), inequalities (2-29) through (2-34) all had to be satisfied by equality and we could then utilize the equality conditions for Holder's inequality to find the form of the optimal control.

The reason why we cannot proceed in the same way and find values γ_m^* such that the infimum in (5-2) is attained at $\gamma_m = \gamma_m^*$ is that such values of γ_m might not exist.

Whereas it can be shown that the infimum in (2-24) is attained by some set of $n + 1$ finite real numbers α^*, γ_k^* $k = 1, 2, \ldots, n$, it is very possible that we can only find values of γ_m for which the infimum in (5-2) is approached arbitrarily closely no matter how large we take n. Of course, if there exist values which minimize the denominator of (5-2) we can then proceed as before to obtain the form of the optimal control.

5.2 Form of the Optimal Control

At $T = T_o$, the minimum norm of any functional f on $\bar{L}_1$ satisfying (5-1) is given by (5-2) which becomes

$$\frac{1}{\inf_{\gamma_{k,n}} \left\{ C_1 \left| \int_0^{T_o} |B(\tau)|\, d\tau \right| + C_2 \sum_{m=1}^{n} |\gamma_m| \right\}} = A \leq 1 \qquad (5\text{-}5)$$

where

$$B(\tau) = \frac{1}{x_d^1} h_1(T_o, \tau) + \sum_{m=1}^{n} \gamma_m h_2\left(\frac{m}{n} T_o, \tau\right)$$

Let $n = 2, 4, 8, \ldots$ and for each n let $\gamma_{m_n}^*$ be those γ_m which minimize the denominator of (5-5), then

$$\frac{1}{C_1\left|\int_0^{T_o} |C(\tau)|\,d\tau\right| + C_2 \sum_{m_n=1}^{n} |\gamma^*_{m_n}|} = A_n \qquad (5\text{-}6)$$

where

$$C(\tau) = \frac{1}{x_d^1} h_1(T_o,\tau) + \sum_{m_n=1}^{n} \gamma^*_{m_n} h_2\left(\frac{m_n T_o}{n},\ \tau\right)$$

and A_n is a monotone increasing sequence approaching A. This follows because the denominator of (5-6) is a monotone decreasing function of n for $n = 2, 4, 8, \ldots$ which approaches the denominator of (5-5) for large enough values of n.

If we now solve the sequence of problems in which, for each $n = 2, 4, 8, \ldots,$ we wish to find the control which makes

$$\max\left\{\frac{1}{C_1}\|u\|_\infty,\ \frac{1}{C_2}\max_{m_n=1,\ldots,n} |x^2\left(\frac{m_n}{n} T_o\right)|\right\} = \text{minimum} \qquad (5\text{-}7)$$

while keeping

$$\int_0^{T_o} h_1(T_o,\tau)u(\tau)\,d\tau = x_d^1 \qquad (5\text{-}8)$$

then by reproducing the analysis given in Chapter 2 we obtain that the minimum value of the quantity in (5-7) is the value A_n defined in (5-6). Furthermore, the control which imparts the value A_n to (5-7) is given by

$$u_n(t) = A_n C_1 \operatorname{sgn} \sigma_n^*(t) \tag{5-9}$$

where

$$\sigma_n^*(\tau) = \frac{1}{x_d^1} h_1(T_o,\tau) + \sum_{m_n=1}^{n} \gamma_{m_n}^* h_2\left(\frac{m_n}{n} T_o,\tau\right) \tag{5-10}$$

By examining (5-6) through (5-8), it is clear that we have created a sequence of controls $u_n(t)$ defined by (5-9) and (5-10) such that each control satisfies the control and terminal constraints of the problem and as n increases, satisfies the output constraint at an increasing number of points. We intend to find a limiting control of some subsequence of the above controls and show that it is an optimal control.

Let $\psi_n(t)$ be a step function with $\psi_n(0) = 0$ and discontinuities of magnitude $\gamma_{m_n}^*$ at times $t = \frac{m_n}{n} T_o$. The variation of each step function (see Section 2.4) is then

$$\overset{T_o}{\underset{0}{V}}(\psi_n) = \sum_{m_n=1}^{n} |\gamma_{m_n}^*| \tag{5-11}$$

Since A_n is a monotone increasing sequence we obtain from (5-6) that, for all n,

$$\overset{T_o}{\underset{0}{V}}(\psi_n) \leq \frac{1}{C_2 A_n} \leq \frac{1}{C_2 A_1} \tag{5-12}$$

Therefore, the step functions $\psi_n(t)$ have a uniform bound on their variations and the sequence satisfies the requirements of Theorem 1, Section 2.4 so that we may select a subsequence which converges pointwise on $[0,T_o]$ to a function $\psi^*(t)$ of finite variation. Utilizing Lemma 1 of the same section we can ensure that the limit of the variations of the subsequence, still denoted by $\psi_n(t)$, exists and satisfies

$$\lim \overset{T_o}{\underset{0}{V}}(\psi_n) \geq \overset{T_o}{\underset{0}{V}}(\psi^*) \tag{5-13}$$

In (5-10), $\sigma_n^*(\tau)$ can now be rewritten

$$\sigma_n^*(\tau) = \frac{1}{x_d^1} h_1(T_o,\tau) + \int_0^{T_o} h_2(t,\tau)\, d\psi_n(t) \tag{5-14}$$

because for any τ, the Stieltjes integral over a point of discontinuity $\bar{t}$ of $\psi_n(t)$ yields

$$h_2(\bar{t},\tau)\,[\psi_n(\bar{t}+0) - \psi_n(\bar{t}-0)\,] \tag{5-15}$$

Since the discontinuity points $\bar{t}$ of $\psi_n(t)$ occur at $t = \frac{m_n}{n} T_o$ and

$$\left| \psi_n \left(\frac{m_n}{n} T_o + 0 \right) - \psi_n \left(\frac{m_n}{n} T_o - 0 \right) \right| = \gamma^*_{m_n} \qquad (5\text{-}16)$$

(5-10) and (5-14) are equivalent.

These controls define a sequence of functionals $\{g_n\}$ on $L_1[0,T_o]$

$$g_n\big(x(\tau)\big) = \int_0^{T_o} x(\tau) u_n(\tau)\, d\tau \qquad x(\tau) = L_1[0,T_o] \qquad (5\text{-}17)$$

where

$$\|g_n\| = \|u_n\|_\infty = A_n C_1 \qquad (5\text{-}18)$$

We now introduce the concepts of weak convergence and weak compactness which are essential for deriving the form of the optimal control. (Note: the following definitions are taken from Ref. 29. These concepts are usually called weak* convergence and weak* compactness, respectively, in the American literature.)

Definition: Let E be a normed linear space. A sequence $\{g_n\}$ of functionals from the dual space $\bar{E}$ (the space of bounded linear functionals on E) is called weakly convergent

to the functional $g_o \in \bar{E}$, if the sequence $g_n(x) \to g_o(x)$ for every $x \in E$.

Definition: A set G of functionals from the dual space $\bar{E}$ is said to be weakly compact if for every sequence $\{g_n\}$ from G, we can choose a subsequence which converges weakly to some functional $g_o \in G$.

From (5-18), since $A_n \leq A \leq 1$ for all n, it follows that $\|g_n\| \leq AC_1$ for all n and consequently the functionals g_n lie in the sphere of radius AC_1 in the dual space of $L_1[0,T_o]$ $\left(\text{i.e., } L_\infty[0,T_o]\right)$. But from Theorem 3, p. 117 of Liusternik and Sobolev,[29] this sphere is weakly compact so that it is possible to pick a subsequence, again denoted by g_n, such that

$$\lim g_n\big(x(t)\big) = g\big(x(t)\big) \quad \text{for all} \quad x(t) \in L_1[0,T_o] \tag{5-19}$$

where g is a linear functional on $L_1[0,T_o]$ with

$$\|g\| \leq AC_1 \tag{5-20}$$

The Riesz representation theorem[33] states that the bounded linear functional g has the representation

$$g\big(x(t)\big) = \int_0^{T_o} x(t)u(t)\,dt \qquad u(t) \in L_\infty[0,T_o] \tag{5-21}$$

with $\|g\| = \|u\|_\infty$. It will now be shown that $u(t)$ is an optimal control.

First, since $h_1(t,\tau)$, $h_2(t,\tau)$ are bounded, they are elements of $L_1[0,T_o]$ for any value of t. (τ is considered as a variable, t as a parameter.) Next, from (5-8), (5-17)

$$g_n[h_1(T_o,\tau)] = \int_0^{T_o} h_1(T_o,\tau)u_n(\tau)d\tau = x_1^d \qquad (5\text{-}22)$$

for all n. Since

$$\lim_{n \to \infty} g_n[h_1(T_o,\tau)] = g[h_1(T_o,\tau)] = \int_0^{T_o} h_1(T_o,\tau)u(\tau)d\tau \qquad (5\text{-}23)$$

it follows that

$$\int_0^{T_o} h_1(T_o,\tau)u(\tau)d\tau = x_1^d \qquad (5\text{-}24)$$

and $u(t)$ satisfies the terminal constraint. Since $\|u\|_\infty = \|g\| \leq AC_1$ $u(t)$ also satisfies the control constraint. It remains to show that $u(t)$ satisfies the output constraint.

Let us take any instant of time $t = \frac{m}{n_1} T_o$ where n_1 is either 2, 4, 8, ... etc. and m is any positive integer equal or less than n_1. Because of the method of construction of the sequence $u_n(t)$ we have <u>for all</u> $n \geq n_1$ (from (5-6) through (5-8))

$$|x_n^2 \left(\frac{m}{n_1} T_o\right)| = \left|\int_0^{T_o} h_2 \left(\frac{m}{n_1} T_o, \tau\right) u_n(\tau) d\tau\right|$$

$$= |g_n\left[h_2 \left(\frac{m}{n_1} T_o, \tau\right)\right]| \leq C_2 \qquad (5\text{-}25)$$

where $x_n^2(t)$ is the plant output when control $u_n(t)$ is the plant input. It follows from $g_n \overset{\text{weakly}}{\rightarrow} g$ that

$$C_2 \geq \lim_{n \to \infty} |g_n\left[h_2\left(\frac{m}{n_1} T_o, \tau\right)\right]| = |g\left[h_2\left(\frac{m}{n_1} T_o, \tau\right)\right]|$$

$$= \left|\int_0^{T_o} h_2\left(\frac{m}{n_1} T_o, \tau\right) u(\tau) d\tau\right| \qquad (5\text{-}26)$$

$$= |x^2\left(\frac{m}{n_1} T_o\right)|$$

Since the set of times $\left\{t_i = \frac{m}{n_1} T\right\}$ where $n_1 = 2, 4, 8, \ldots$ and m is any positive integer equal or less than n_1 is

dense in $[0,T_o]$, the output constraint is satisfied at all instants of time in $[0,T_o]$ when the control $u(t)$ is applied and therefore $u(t)$ is an optimal control.

It is possible to find a more explicit characterization of the optimal control. For any element $x(\tau)$ in $L_1[0,T_o]$

$$\int_0^{T_o} x(\tau)u(\tau)d\tau = \lim_{n \to \infty} \int_0^{T_o} x(\tau)u_n(\tau)d\tau \tag{5-27}$$

$$= \lim_{n \to \infty} \int_0^{T_o} x(\tau)A_nC_1 \operatorname{sgn}[\sigma_n^*(\tau)]d\tau$$

From (5-14), the fact that $\psi_n(t)$ converges pointwise to $\psi^*(t)$ and making use of Theorem 2 of Section 2.4

$$\sigma_n^*(\tau) = \frac{1}{x_d^1} h_1(T_o,\tau) + \int_0^{T_o} h_2(t,\tau)d\psi_n(t) \to \frac{1}{x_d^1} h_1(T_o,\tau) \tag{5-28}$$

$$+ \int_0^{T_o} h_2(t,\tau)d\psi^*(t) = \sigma^*(t)$$

If $\sigma^*(t) = 0$ only on a set of measure zero, then by using the dominated convergence theorem[32]

$$\lim_{n \to \infty} \int_0^{T_o} x(\tau) A_n C_1 \operatorname{sgn}[\sigma_n^*(\tau)]\, d\tau =$$

$$\int_0^{T_o} x(\tau) A C_1 \operatorname{sgn}[\sigma^*(\tau)]\, d\tau \tag{5-29}$$

and by comparing (5-27) and (5-29), we see that

$$u(\tau) = A C_1 \operatorname{sgn} \sigma^*(\tau)$$

$$= A C_1 \operatorname{sgn}\left[\frac{1}{x_d} h_1(T_o, \tau) + \int_0^{T_o} h_2(t, \tau)\, d\psi^*(t) \right] \tag{5-30}$$

is an optimal control. It also follows from the additivity of the integral that (5-30) holds even if the argument of the sgn function equals zero on some set W of positive measure although, of course, (5-30) does not yield the actual value of $u(t)$ on W.

It is now shown that the function $\psi^*(t)$ minimizes

$$C_1 \left[\int_0^{T_o} \left| \frac{1}{x_d} h_1(T_o, \tau) + \int_0^{T_o} h_2(t, \tau)\, d\psi(t) \right| d\tau \right] + C_2 \underset{0}{\overset{T_o}{V}} (\psi) \tag{5-31}$$

in the class of measurable functions. It is clear that we need only consider functions with variations equal or less than

$$C_1 \int_0^{T_o} \left| \frac{1}{x_d^1} h_1(T_o, \tau) \right| d\tau \tag{5-32}$$

since the function $\psi(t) \equiv 0$ yields the above value for expression (5-31). From the proof of Lemma 2 in Section 2.4, it follows that the infimum of (5-31) over the class of functions of finite variation equals its infimum over the class of step functions with discontinuities at points $\frac{m}{n} T$, $m \leq n$. Now the functions $\psi_n(t)$ defined previously form a sequence of functions such that

$$\lim_{n \to \infty} \left\{ C_1 \left[\int_0^{T_o} \left| \frac{1}{x_d^1} h_1(T_o, \tau) + \int_0^{T_o} h_2(t, \tau) d\psi_n(t) \right| d\tau \right] + C_2 \bigvee_0^{T_o} (\psi_n) \right\} = B \tag{5-33}$$

where B is the infimum of (5-31) over the aforementioned class of step functions and thus over the class of functions of finite variation. From (5-13), (5-28), and (5-33)

$$\left\{C_1\left[\int_0^{T_o}\left|\frac{1}{x_d^1}\,h_1(T_o,\tau)+\int_0^{T_o}h_2(t,\tau)\,d\psi^*(t)\right|d\tau\right]\right.$$
$$\left.+\,C_2\,\underset{0}{\overset{T_o}{V}}\,(\psi^*)\right\}\le B \qquad (5\text{-}34)$$

But the definition of B implies that inequality (5-34) is satisfied by equality and therefore $\psi^*(t)$ minimizes (5-31) over the class of functions of finite variation.

We have now shown that at the optimal time $T = T_o$, there exists a function of finite variation $\psi^*(t)$ which minimizes (5-31). Moreover, the optimal control is given by

$$u(\tau) = AC_1\ \mathrm{sgn}\left[\frac{1}{x_d^1}\,h_1(T_o,\tau)+\int_0^{T_o}h_2(t,\tau)\,d\psi^*(t)\right] \qquad (5\text{-}35)$$

if the argument of the sgn function in (5-35) equals zero over some finite subinterval in $[0,T_o]$, then the control is undefined over the same subinterval. In some cases the control does not operate on the boundary of its constraint because such action would cause a violation of the output constraint. Specifically, the control might be determined by the fact that certain portions of the optimal trajectory

lie on the boundary of the output constraint and it is worthwhile to investigate if this is the case for those subintervals for which the argument equals zero.

Since $h_2(t,\tau) = 0$ for $t < \tau$, expressions (5-31) and (5-35) may be written as

$$C_1\left[\int_0^{T_o} \left|\frac{1}{x_d^1} h_1(T_o,\tau) + \int_\tau^{T_o} h_2(t,\tau)\,d\psi(t)\right| d\tau\right] + C_2 \overset{T_o}{V}_0(\psi) \tag{5-36}$$

$$u(\tau) = AC_1\ \mathrm{sgn}\left[\frac{1}{x_d^1} h_1(T_o,\tau) + \int_\tau^{T_o} h_2(t,\tau)\,d\psi^*(t)\right]$$

respectively. The above expressions are the ones that must be used if $h_2(t,\tau)$ has a simple discontinuity at $t = \tau$ with

$$h_2(t,\tau)\Big|_{t=\tau} = \lim_{\substack{t \to \tau \\ t > \tau}} h_2(t,\tau)$$

in order that the Stieltjes integrals appearing in these expressions be well defined.

5.3 Approximation Method

In the preceding section, we established a relationship between the optimal control and a function $\psi^*(t)$ of finite

variation minimizing (5-31). Since, in general, one cannot minimize (5-31) over the class of functions of finite variation to find $\psi^*(t)$, we assume $\psi^*(t)$ lies in some smaller class of functions, for example, those functions having a parametric representation of the form

$$\psi^*(t) = s_n(t) + \sum_{i=1}^{n} a_i t^i \tag{5-37}$$

$$s_n(t) = \sum_{i=1}^{k_n} \gamma_i \delta^1(t - t_i) \qquad \delta^1(t) = 1 \quad t \geq 0, \ \delta^1(t) = 0 \quad t < 0$$

This representation is chosen because it is known that any function of finite variation can be represented as the sum of a saltus function and a continuous function, where a saltus function is a step function that may possess a denumerable number of continuities. The function $s_n(t)$ can then be considered as an approximation to the saltus function while the polynomial expansion is an approximation for the continuous part. If $\psi^*(t)$ has the representation (5-37), we can then find it by minimizing (5-31) over the class of functions with this representation. Substituting the right side of (5-37) into (5-31) we obtain

$$C_1\left[\int_0^{T_o} \left|\frac{1}{x_d^1} h_1(T_o,\tau) + \sum_{i=1}^{k_n} \gamma_i h_2(t_i,\tau)\right.\right.$$

$$\left.\left. + \int_0^{T_o} h_2(t,\tau)\left[\sum_{i=1}^{n} i a_i t^{(i-1)}\right] dt\right| d\tau\right] \qquad (5\text{-}38)$$

$$+ C_2\left[\sum_{i=1}^{k_n} |\gamma_i| + \int_0^{T_o} \left|\sum_{i=1}^{n} i a_i t^{(i-1)}\right| d\tau\right]$$

and perform a finite dimensional minimization over the parameters γ_i, $i = 1, \ldots, k_n$ and a_i, $i = 1, \ldots, n$. It is noted that in (5-38) the Stieltjes integration has been replaced by a finite summation and a Riemann integration (see Natanson[32] for justification).

A priori, it is not known how large the values of k_n and n should be chosen. In addition, the constraint of computational feasibility places a limit on how large they can be chosen. Even after the value of k_n is selected, one must determine where to place the points t_i which are the locations of the discontinuities in the step function $s_n(t)$. These points are discussed below.

If in (5-37) we set $n = 0$ and $k_n = p$, then (5-38) reduces to (3-19) and therefore the present method of approximation can be considered a generalization of the scheme proposed in Chapter 3. It seems likely that even if $n \neq 0$ the parameters γ_i associated with the points t_i will mainly be associated with satisfying the output constraint at time t_i. Therefore we should choose the number and location of the points t_i to correspond with those places (in time) at which it is essential for the output constraint to be satisfied. If there are no such points we can set $k_n = 0$ and then choose a larger value of n in the polynomial expansion to obtain a greater measure of constraint dispersed over every instant of time in the transition period. In either case one should choose values of k_n and n such that the finite dimensional minimization of

$$C_1\left[\int_0^T \left|\frac{1}{x_d} h_1(T,\tau) + \sum_{i=1}^{k_n} \gamma_i h_2(t_i,\tau)\right. \right.$$

$$\left. + \int_0^T h_2(t,\tau)\left[\sum_{i=1}^{n} i a_i t^{(i-1)}\right] dt\right| d\tau \Bigg] \qquad (5\text{-}39)$$

$$+ C_2\left[\sum_{i=1}^{k_n} |\gamma_i| + \int_0^T \left|\sum_{i=1}^{n} i a_i t^{(i-1)}\right| dt\right]$$

which is of order $k_n + n$, can be performed in a reasonable computation time. The actual values depend on the computing facilities available, the form of the impulse response functions $h_1(t,\tau)$, $h_2(t,\tau)$ and the method of minimization. Some comments on the types of minimization schemes which may be employed appear in Chapter 7.

The first value of T for which the minimum of (5-39) is equal to or greater than unity is taken as the optimal time T_o. The approximation to the optimal control is given by

$$u(\tau) = A_1 C_1 \operatorname{sgn}\left[\frac{1}{x_d} h_1(T_o,\tau) + \sum_{i=1}^{k_n} \gamma_i^* h_2(t_i,\tau) + \int_0^{T_o} h_2(t,\tau)\left[\sum_{i=1}^{n} i a_i^* t^{(i-1)}\right] dt\right] \tag{5-40}$$

where A_1 is the minimum of expression (5-39) at $T = T_o$ and γ_i^*, $i = 1, \ldots, k_n$, a_i^*, $i = 1, \ldots, n$ are those values of γ_i, a_i at which the minimum is attained. The effectiveness of the approximation is ascertained by applying the control given by (5-40) to the plant and examining how well the constraints are satisfied.

We remark that the same approximation scheme and the same expressions ((5-39) and (5-40)) are used even if $h_2(t,\tau)$ has a simple discontinuity at $t = \tau$. The integrals in these expressions are Riemann integrals and therefore are still well defined.

It is possible to see, by examining the preceding derivation, that the function $\psi^*(t)$ is constant over those intervals for which $x^2(t)$ cannot lie on the boundary of its constraint. Since $\psi^*(0) = 0$, if we can determine some interval of time $[0,T_1]$ such that $x^2(t)$ cannot exceed its constraint, then $\psi^*(t) \equiv 0$ on this same interval. We make use of this fact in the example.

5.4 Example

In order to illustrate the above procedure we consider the double integrator problem treated in Chapter 3. We first look at the solution to the problem with unconstrained outputs and note that the given output constraint on $x^2(t)$ is violated for $t > 1$ second. As pointed out in Section 3.5 it is impossible for the output constraint to be violated for $t \leq 1$ second. The statement in the preceding paragraph then implies that $\psi^*(t) \equiv 0$ for $0 \leq t \leq 1$ seconds. Instead of choosing an approximation for $\psi^*(t)$ of the form (5-37), we choose

$$\pi_n(t) = s_n(t) + \sum_{i=1}^{n} a_i (t-1)^i \delta^1(t-1)$$

$$s_n(t) = \sum_{i=1}^{k_n} \gamma_i \delta^1(t - t_i)$$

To be specific, let $n = 2$, $k_n = 1$, $t_1 = \frac{2}{3} T$. As stated previously, this will place heavy emphasis on the output constraint being satisfied at $t_1 = \frac{2}{3} T$ where T is the optimal time. Expression (5-39) becomes

$$\left| \int_0^T \left| \tfrac{1}{2}(T - \tau) + \gamma_1 \delta^1\left(\tfrac{2}{3} T - \tau\right) + \int_1^T \delta^1(t - \tau)\,[a_1 + 2a_2(t-1)]\,dt \right| d\tau \right| + |\gamma_1| \quad (5\text{-}41)$$

$$+ \int_1^T |a_1 + 2a_2(t-1)|\,dt \Bigg|$$

For $\tau \leq 1$

$$\int_1^T \delta^1(t - \tau)\,[a_1 + 2a_2(t-1)]\,dt = a_1(T-1) + a_2(T^2 - T) \quad (5\text{-}42)$$

For $\tau \geq 1$

$$\int_1^T \delta^1(t-\tau)\,[a_1 + 2a_2(t-1)]\,dt = a_1(T-\tau) + a_2[(T^2-\tau^2) - (T-\tau)] \quad (5\text{-}43$$

Using (5-42) and (5-43) we find that the first value of T for which the minimum of (5-41) with respect to γ_1, a_1, a_2 is equal or greater than unity is $T = 2.5$ for which the minimum equals unity. The values of the parameters which yield this minimum are

$$\gamma_1 = 0\,, \quad a_1 = -\frac{1}{2}\,, \quad a_2 = 0$$

The control specified by (5-40) is

$$\begin{aligned} u(t) &= \operatorname{sgn}\left[\tfrac{1}{2}(1-t)\right] = 1 \qquad 0 \leq t \leq 1 \\ u(t) &= \operatorname{sgn}[0] = ? \qquad\quad 1 < t \leq 2.5 \end{aligned} \tag{5-44}$$

At $t = 1$, $x^2(1) = 1$ which places x^2 on the boundary of its constraint. In accordance with the discussion following (5-35) we investigate the possibility of $x^2(t)$ remaining on the boundary for the interval for which the argument of the signum function is zero. This requires

$$u(t) = 0 \qquad 1 < t \leq 2.5 \tag{5-45}$$

so that the control is now completely determined.

It happens that in this case the approximate optimal control given by (5-44) and (5-45) turns out to be the exact optimal control. The plant outputs and optimal control are shown in Fig. 5-1.

5.5 Plants with Nonzero Initial Conditions

If the plant is in a nonquiescent state at the initial time then the modification of expression (5-2) to take account of initial conditions is

$$\frac{1}{\inf_{\alpha,\gamma_{m,n}} \left\{ C_1 \left| \int_0^T |D(\tau)|\,d\tau \right| + C_2 \sum_{m=1}^{n} |\gamma_m| \right\}} \qquad (5\text{-}46)$$

$$\alpha c + \sum_{m=1}^{n} \gamma_m d_m = 1$$

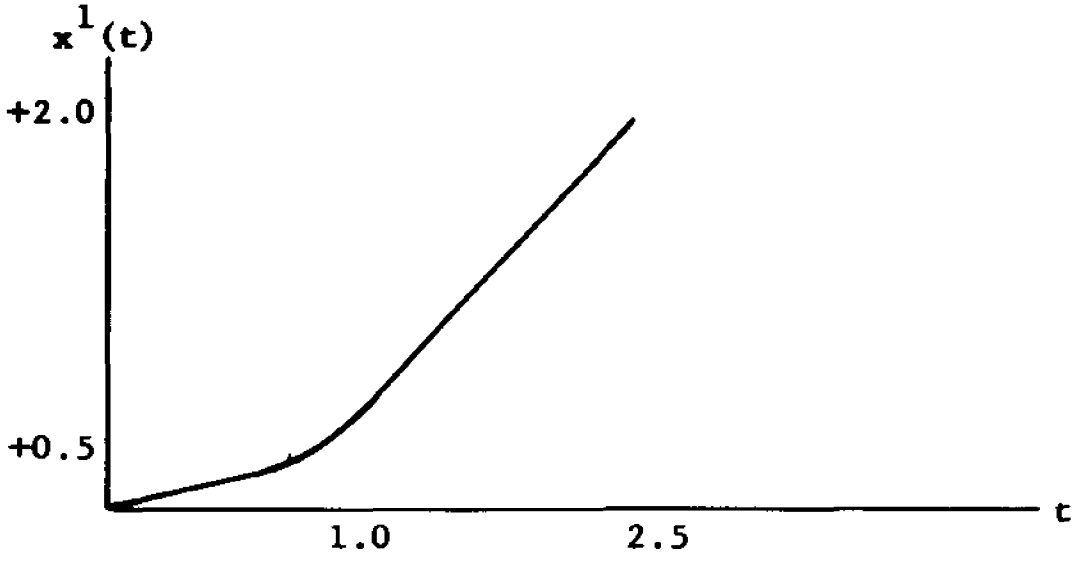
where

$$D(\tau) = \alpha h_1(T,\tau) + \sum_{m=1}^{n} \gamma_m h_2\left(\frac{m}{n}T,\tau\right)$$

$$c = x_d^1 - x_o^1(T)$$

$$d_m = -x_o^2\left(\frac{m}{n}T\right)$$

and $x_o^i(t)$, $i = 1, 2,$ expresses the effect of nonzero initial conditions on the i^{th} plant output at time t. By repeating the steps in Section 5.2 it is found that at the



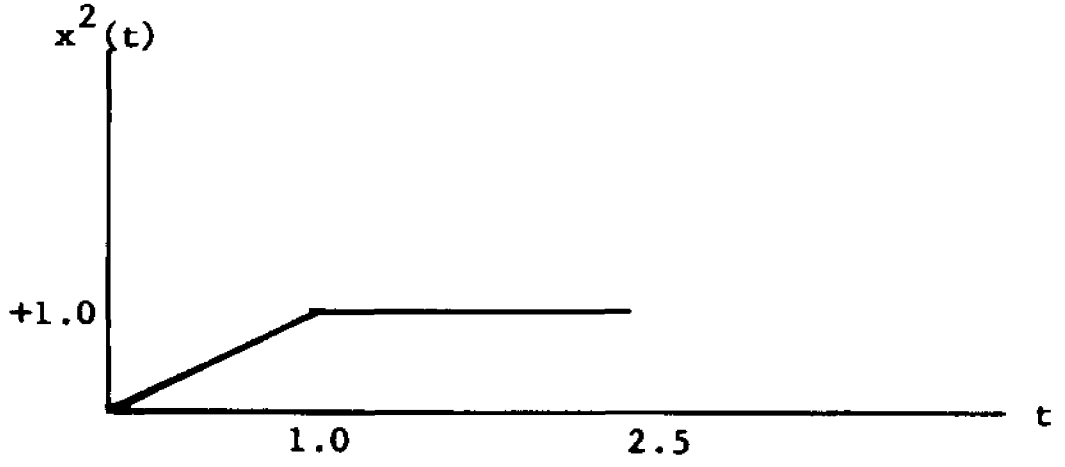


Fig. 5-1 Plant Outputs and Optimal Control for Example

optimal time T_o there exists some constant α^* and a function of finite variation $\psi^*(t)$ which minimizes

$$C_1\left[\int_0^{T_o} \left|\alpha h_1(T_o,\tau) + \int_0^{T_o} h_2(t,\tau)\,d\psi(t)\right| d\tau\right] + C_2 \underset{0}{\overset{T_o}{V}}(\psi) \qquad (5\text{-}47)$$

subject to the constraint

$$\alpha\left(x_d^1 - x_o^1(T_o)\right) - \int_0^T x_o^2(t)\,d\psi(t) = 1$$

The optimal control is given by

$$u(\tau) = AC_1\ \mathrm{sgn}\left[\alpha^* h_1(T_o,\tau) + \int_0^{T_o} h_2(t,\tau)\,d\psi^*(t)\right] \qquad (5\text{-}48)$$

where A is the inverse of the minimum of expression (5-47).

Applying the method of Section 5.3 to approximate the optimal control (5-48), we first minimize

$$C_1\left[\int_0^T |\alpha h_1(T,\tau) + \sum_{i=1}^{k_n} \gamma_i h_2(t_i,\tau) + \int_0^T h_2(t,\tau)\left[\sum_{i=1}^{n} i a_i t^{(i-1)}\right]dt\right|d\tau\Bigg] + C_2\left[\sum_{i=1}^{k_n} |\gamma_i| + \int_0^T \left|\sum_{i=1}^{n} i a_i t^{(i-1)}\right|dt\right] \tag{5-49}$$

where k_n and n are fixed, subject to the constraint

$$\alpha\left(x_d^1 - x_o^1(T)\right) + \sum_{i=1}^{k_n} \gamma_i\left(-x_o^2(t_i)\right) - \int_0^T x_o^2(t)\left[\sum_{i=1}^{n} i a_i t^{(i-1)}\right]dt = 1 \tag{5-50}$$

The first value of T for which the minimum of (5-49) (denoted by A_1) is equal or greater than unity is taken as the optimal time T_o.

The approximation to the optimal control is given by

$$u(\tau) = A_1 C_1 \operatorname{sgn}\left[\alpha^* h_1(T_o,\tau) + \sum_{i=1}^{k_n} \gamma_i^* h_2(t_1,\tau) + \int_0^{T_o} h_2(t,\tau)\left[\sum_{i=1}^{n} i a_i^* t^{(i-1)}\right] dt\right]$$

where α^*, γ_i^*, $i = 1, \ldots, k_n$ and a_i^*, $i = 1, \ldots, n$ are the parameters which minimize (5-49) subject to (5-50) at time T_o.

CHAPTER 6. TIME OPTIMAL CONTROL OF DISCRETE SYSTEMS WITH BOUNDED OUTPUTS

In this chapter we consider discrete systems with bounds on the outputs as well as the inputs. Difficulties arise from the fact that for systems with control amplitude constraints the value of the control at the i^{th} step is proportional to $\text{sgn}[f(i)]$ where $f(i)$ is some known function of i. In many cases $f(i)$ is equal to zero for several values of i and at these steps the control is undefined. A procedure is given for finding the value of the control at these steps. We remark that in this chapter the product space approach yields the exact optimal solutions.

6.1 Problem Statement

Consider a double output-single input plant whose dynamic behavior is described by the discrete operator equation

$$\underline{y}(N) = \underline{y}_o(N) + \sum_{i=0}^{N} H(N,i)u(i) \qquad (6\text{-}1)$$

where $\underline{y}(N)$ is a two dimensional vector representing the output of the plant at the N^{th} step, $u(i)$ is the scalar input at the i^{th} step and the 2×1 matrix $H(N,i)$ is

an "impulse response" matrix whose element $h_{\ell 1}(N,i)$, $\ell = 1,2$ represents the ℓ^{th} output response observed at step N when all initial conditions and inputs are zero except for $u(i) = 1$.

The vector $y_o(N)$ represents the contribution to the output vector at time N due to nonzero initial conditions in the plant at time zero. This is assumed to be known for all $N \geq 0$.

Equation (6-1) may be obtained when a discrete plant is considered which is described by the linear difference equations

$$\underline{x}(N+1) = \varphi(N)\underline{x}(N) + \Delta(N)u(N)$$

$$\underline{y}(N) = M(N)\underline{x}(N) + P(N)u(N)$$

For the particular case when $\varphi(N) = \varphi$ is constant[35]

$$\underline{y}_o(N) = M(N)\varphi^N\underline{x}(0)$$

$$H(N,i) = P(N) \qquad N = i$$

$$H(N,i) = M(N)\varphi^{N-1-i}\Delta(i) \qquad N > i$$

The problem considered is that of finding the smallest N and the corresponding control such that

$$y_1(N) = y_1^d \tag{6-2}$$

where y_1 is the first component of the output vector $\underline{y}$ and y_1^d is some desired terminal value. We have a constraint on the input sequence $u(i)$ of the form

$$\|u(i)\|_p = \left[\sum_{i=0}^{N} |u(i)|^p \right]^{1/p} \leq C_1 \qquad p \geq 1 \tag{6-3}$$

where C_1 is a positive constant. For $p = 1, 2, \infty$ we have area, energy and amplitude constraints on the input sequence, respectively.

We also have a constraint on the second output component $y_2(i)$ of the form

$$\|y_2(i) - y_2^d(i)\|_{p'} = \left[\sum_{i=0}^{N} |y_2(i) - y_2^d(i)|^{p'} \right]^{1/p} \leq C_2 \qquad p' \geq 1 \tag{6-4}$$

where C_2 is some positive constant and $y_2^d(i)$ is some desired trajectory. For $p = \infty$, we constrain the maximum deviation of the actual trajectory from the desired one, if $y_2^d(i) = 0$ this becomes an amplitude constraint on $y_2(i)$. For $p = 2$, we constrain the squared error of the deviation.

It is remarked that the following development directly extends to multi-input systems with multiple control, terminal and output constraints and that it is also possible, by

using the multi-norm procedure of Sarachik and Kranc,[28] to handle different types of constraints (e.g., amplitude and energy) on distinct components of the input and/or output vector.

This problem may be reformulated by tentatively fixing N and asking for the control which makes

$$\max \left\{ \frac{1}{C_1} \|u(i)\|_p \ , \ \frac{1}{C_2} \|y_2(i) - y_2^d(i)\|_p \right\} = \min \qquad (6\text{-}5)$$

while maintaining

$$\sum_{i=0}^{N} H_1(N,i)u(i) = y_1^d - y_{o1}(N) \qquad (6\text{-}6)$$

where $H_1(N,i)$ is the first row of the impulse response matrix $H(N,i)$ and $y_{o1}(N)$ is the first component of the initial condition vector. The first value of N for which the minimum is (6-5) is equal or less than unity then solves the original problem specified by (6-1) through (6-4).

6.2 Formulation as an L-Problem in the Theory of Moments

Using the method employed by Katz and Kranc, suitable Banach spaces are constructed so that the problem specified by (6-5) and (6-6) may be reformulated as an L-problem in the theory of moments.[30]

Consider the space of $N + 1$ dimensional sequences $a(i)$, $i = 0, 1, .. N$ where $a(i)$ is a real number for each i, having

$$\left[\sum_{i=0}^{N} |a(i)|^r \right]^{1/r} < \infty \qquad r \geq 1 \tag{6-7}$$

Defining an addition operation between two sequences $a_1(i)$ and $a_2(i)$ by

$$(a_1 + a_2)(i) = a_1(i) + a_2(i) \qquad i = 0, 1, \ldots, N_j \tag{6-8}$$

an operation of multiplication between sequences $a(i)$ and real scalars α by

$$(\alpha a)(i) = \alpha a(i) \qquad i = 0, 1, \ldots, N \tag{6-9}$$

and designating (6-7) as the norm of any element $a(i)$ in this space, it is clear that the space of such sequences is a Banach space which we denote by $\ell_r^{(N+1)}$.

Define the indices q, q' conjugate to p, p', respectively by

$$\frac{1}{p} + \frac{1}{q} = 1 , \quad \frac{1}{p'} + \frac{1}{q'} = 1 \tag{6-10}$$

and consider the product space $\bar{L}_1$ consisting of composite vectors $\bar{a}(i)$ formed by pairing sequences $b(i)$ in $\ell_q^{(N+1)}$ with sequences $c(i)$ in $\ell_{q'}^{(N+1)}$

$$\bar{a}(i) = \left\{ b(i) \ , \quad b(i) \in \ell_q^{(N+1)} \ ; \quad c(i) \ , \quad c(i) \in \ell_{q'}^{(N+1)} \right\} \tag{6-11}$$

If we define addition of 2 composite vectors $\bar{a}_1(i) = \left\{ b_1(i), c_1(i) \right\}$ and $\bar{a}_2(i) = \left\{ b_2(i), c_2(i) \right\}$ by

$$(\bar{a}_1 + \bar{a}_2)(i) = \left\{ b_1(i) + b_2(i) \ , \quad c_1(i) + c_2(i) \right\} \tag{6-12}$$

multiplication between composite vectors $\bar{a}(i)$ and real scalars α by

$$(\alpha\bar{a})(i) = \left\{ \alpha b(i) \ , \quad \alpha c(i) \right\} \tag{6-13}$$

and the norm of a composite vector $\bar{a}(i)$ as

$$\|\bar{a}(i)\|_{\bar{L}_1} = C_1 \|b(i)\|_q + C_2 \|c(i)\|_{q'}, \qquad C_1, \ C_2 > 0 \tag{6-14}$$

where

$$\|b(i)\|_q = \left[\sum_{i=0}^{N} |b(i)|^q \right]^{1/q}, \quad \|c(i)\|_{q'} = \left[\sum_{i=0}^{N} |c(i)|^{q'} \right]^{1/q} \tag{6-15}$$

then it easily follows that L_1 is a Banach space.

a) Any bounded linear functional f on $\bar{L}_1$ has a representation

$$f\left(\bar{a}(i)\right) = \sum_{i=0}^{N} b(i)u(i) + \sum_{i=0}^{N} c(i)v(i) \tag{6-16}$$

with $u(i)$ in $\ell_p^{(N+1)}$ and $v(i)$ in $\ell_{p'}^{(N+1)}$.

b) The norm of the functional f of (6-16) defined as

$$\|f\| = \sup_{\bar{a}(i) \neq \bar{0}} \frac{|f[\bar{a}(i)]|}{\|\bar{a}(i)\|_{\bar{L}_1}} \tag{6-17}$$

is given by

$$\|f\| = \max \left\{ \frac{1}{C_1} \|u(i)\|_p \, , \; \frac{1}{C_2} \|v(i)\|_{p'} \right\} \tag{6-18}$$

where

$$\|u(i)\|_p = \left[\sum_{i=0}^{N} |u(i)|^p \right]^{1/p} , \; \|v(i)\|_{p'} = \left[\sum_{i=0}^{N} |v(i)|^{p'} \right]^{1/p'} \tag{6-19}$$

To apply these developments to the problem specified by (6-5) and (6-6) we assume that the elements of $H(N,i)$ are finite for all N and i. We may then construct from them composite vector of the form (6-11)

$$\bar{\varphi}(i) = \left\{ H_1(N,i); \; 0(i) \right\} \quad 0(i) \equiv 0 \quad i = 0, 1, \ldots, N$$

$$\bar{\theta}_k(i) = \left\{ H_2(k,i); \; -e_k(i) \right\} \quad e_k(i) = 1, \quad i = k, \; e_k(i) = 0, \quad i \neq k, \quad k = 0, 1, \ldots, N \tag{6-20}$$

where H_1 and H_2 are the first and second rows of the impulse response matrix respectively. From the definition of the impulse response matrix $H_2(k,i) = 0$ for $i > k$.

By requiring that a functional f of the form (6-16) satisfy

$$f\left(\bar{\varphi}(i)\right) = y_1^d - y_{o1}(N) \tag{6-21}$$

we are requiring that the sequence $u(i)$ satisfy (6-6). If we further require that f satisfy

$$f\left(\bar{\theta}_k(i)\right) = y_2^d(k) - y_{o2}(k) \qquad k = 0, 1, \ldots, N \tag{6-22}$$

we are simply requiring that

$$\begin{aligned} v(k) &= y_{o2}(k) + \sum_{i=0}^{N} H_2(k,i)u(i) - y_2^d(k) \\ &= y_2(k) - y_2^d(k) \qquad k = 0, 1, \ldots, N \end{aligned} \tag{6-23}$$

where the last equality in (6-23) follows from (6-1) and the fact that $H_2(k,i) = 0$ for $i > k$. The norm of f, as given by (6-18) then becomes

$$\|f\| = \left\{ \frac{1}{C_1} \|u(i)\|_p \ , \ \frac{1}{C_2} \|y_2(i) - y_2^d(i)\|_p \right\} \tag{6-24}$$

which is the left side of expression (6-5).

It is now clear that the abstract problem of finding the bounded linear functional of minimum norm on $\bar{L}_1$ which maps the given elements $\bar{\varphi}(i)$ and $\bar{\theta}_k(i)$, $k = 0, 1, \ldots, N$ of this space into the given fixed scalars specified by

(6-21) and (6-22) is equivalent to solving the variational problem (6-5) and (6-6) and the first value of N for which this functional has its norm equal or less than unity is the optimal step number. This abstract problem is what is known as an L-problem in the theory of moments

6.3 Existence of an Optimal Solution

If the scalars in (6-21) and (6-22) are to be arbitrary, then the vectors $\bar{\varphi}(i)$ and $\bar{\theta}_k(i)$, $k = 0, 1, \ldots, N$, must be linearly independent and although this is obviously true in the present case for any nontrivial $H_1(N,i)$, we must impose suitable restrictions (such as complete output controllability)[35] on the plant when there is more than one terminal constraint to ensure this condition.

The vectors $\bar{\varphi}(i)$ and $\bar{\theta}_k(i)$, $k = 0, 1, \ldots, N$ being linearly independent, span an $N + 2$ dimensional space and a linear functional f_1 satisfying (6-21) and (6-22) can be defined on this space by

$$f_1\Big(\alpha\bar{\varphi} + \sum_{k=0}^{N} \gamma_k \bar{\theta}_k\Big) = \alpha f_1(\varphi) + \sum_{k=0}^{N} \gamma_k f_1(\bar{\theta}_k) = \alpha c + \sum_{k=i}^{N} \gamma_k d_k \qquad (6\text{-}25)$$

where

$$c = y_1^d - y_{o1}(N) \ , \quad d_k = y_2^d(k) - y_{o2}(k) \qquad (6\text{-}26)$$

and $\bar{\varphi}(i)$, $\bar{\theta}_k(i)$ have been replaced by $\bar{\varphi}$, $\bar{\theta}_k$ for convenience. The norm of f_1 on this $N+2$ dimensional space is found from (6-17)

$$\|f_1\| = \sup_{\substack{\alpha,\gamma_k \\ \text{not all} = 0}} \frac{|f_1(\alpha\bar{\varphi} + \sum_{k=0}^{N} \gamma_k\bar{\theta}_k)|}{\|\alpha\bar{\varphi} + \sum_{k=0}^{N} \gamma_k\bar{\theta}_k\|_{\bar{L}_1}} \tag{6-27}$$

$$= \sup_{\substack{\alpha,\gamma_k \\ \text{not all} = 0}} \frac{|\alpha c + \sum_{k=0}^{N} \gamma_k d_k|}{\|\alpha\bar{\varphi} + \sum_{k=0}^{N} \gamma_k\bar{\theta}_k\|_{\bar{L}_1}}$$

or since $\alpha c + \sum_{k=0}^{N} \gamma_k d_k = 0$ clearly does not give the largest value of this ratio,

$$\|f_1\| = \frac{1}{\inf_{\substack{\alpha,\gamma_k \\ \alpha c + \sum_{k=0}^{N} \gamma_k d_k = 1}} \|\alpha\bar{\varphi} + \sum_{k=0}^{N} \gamma_k\bar{\theta}_k\|_{\bar{L}_1}} \tag{6-28}$$

Using the explicit form of the norm in $\bar{L}_1$, given by (6-14) and (6-15)

$$\|f_1\| \quad \frac{1}{\inf_{\substack{\alpha,\gamma_k \\ \alpha c \sum_{k=0}^{N} \gamma_k d_k = 1}} \left\{ C_1 \left[\sum_{i=0}^{N} |B(i)|^q \right]^{1/q} + C_2 \left[\sum_{k=0}^{N} |\gamma_k|^{q'} \right]^{1/q'} \right.} \tag{6-29}$$

where

$$B(i) = \alpha H_1(N,i) + \sum_{k=0}^{N} \gamma_k H_2(k,i)$$

The Hahn-Banach theorem asserts that the functional defined by (6-25) can be extended to a linear functional f over $\bar{L}_1$, without increase in norm. Since any extension of f_1 over $\bar{L}_1$ must have its norm equal or greater than that of f_1, we therefore infer that the functional of minimum norm on $\bar{L}_1$ satisfying (6-21) and (6-22) has its norm equal to (6-29). The first value of N for which (6-29) is equal or less than unity is the optimal step number hereafter called N_o.

6.4 Form of the Optimal Solution

The existence of an optimal control for the original problem specified by (6-1) through (6-4) at step N_o has

just been established and we now turn to a derivation of the actual form of an optimal input sequence.

It can be shown that the infimum in expression (6-29) ($N = N_o$) is attained at some values of α and γ_k, $k = 0, 1, \ldots, N$ and let $\alpha = \alpha^*$, $\gamma_k = \gamma_k^*$, $k = 0, 1, \ldots, N$ with $\alpha^* c + \sum_{k=0}^{N} \gamma_k^* d_k = 1$ be such values. Also let f be any functional of minimum norm satisfying (6-21) and (6-22). From what has just preceded, the norm of f is equal or less than unity and the $u(i)$ component of f specifies an optimal input sequence. Let the norm of f equal A where $A \leq 1$. From the definition of the norm of a functional, (6-17)

$$|f(\bar{a}(i))| \leq \|f\| \|\bar{a}(i)\|_{L_1} \tag{6-30}$$

But

$$1 = |f(\alpha^* \varphi + \sum_{k=0}^{N} \gamma_k^* \theta_k)| \leq \|f\| \|\alpha^* \varphi + \sum_{k=0}^{N} \gamma_k^* \theta_k\|_{L_1} = 1 \tag{6-31}$$

where the last equality is obtained from (6-29) using the fact that the infimum is attained at the values $\alpha = \alpha^*$, $\gamma_k = \gamma_k^*$, $k = 0, 1, \ldots, N$ and that $\|f_1\| = \|f\|$. Therefore

$$|f(\alpha^* \bar{\varphi} + \sum_{k=0}^{N} \gamma_k^* \bar{\theta}_k)| = \|f\| \|\alpha^* \bar{\varphi} + \sum_{k=0}^{N} \gamma_k^* \bar{\theta}_k\|_{L_1} \tag{6-32}$$

or, more explicitly, using (6-14) through (6-19) and letting

$$\alpha^* H_1(N_o,i) + \sum_{k=0}^{N_o} \gamma_k^* H_2(k,i) = \sigma^*(i),\ g_2(k) = y_2(k) - y_2^d(k),$$

we have

$$\left| \sum_{i=0}^{N_o} \sigma^*(i)u(i) + \sum_{k=0}^{N_o} \gamma_k^* g_2(k) \right| = \tag{6-33}$$

$$\max\left\{\frac{1}{C_1}\|u(i)\|_p,\ \frac{1}{C_2}\|g_2(i)\|_p,\right\}\left\{C_1\|\sigma^*(i)\|_q + C_2\|\gamma^*(i)\|_q,\right\}$$

where

$$\gamma^*(i) = \gamma_i^* \qquad i = 0,\ 1,\ \ldots,\ N_o$$

Since $|a + b| \leq |a| + |b|$

$$\left| \sum_{i=0}^{N_o} \sigma^*(i)u(i) + \sum_{k=0}^{N_o} \gamma^*(k)g_2(k) \right| \leq \left| \sum_{i=0}^{N_o} \sigma^*(i)u(i) \right| \tag{6-34}$$

$$+ \left| \sum_{k=0}^{N_o} \gamma^*(k)g_2(k) \right|$$

From Holder's inequality

$$\left| \sum_{i=0}^{N_o} \sigma^*(i)u(i) \right| \leq \|\sigma^*(i)\|_q \|u(i)\|_p = C_1\|\sigma^*(i)\|_q \left[\frac{1}{C_1}\|u(i)\|_p\right] \tag{6-35}$$

$$\left| \sum_{k=0}^{N_o} \gamma^*(k) g_2(k) \right| \leq \|\gamma^*(i)\|_q, \|g_2(i)\|_p, = C_2 \|\gamma^*(i)\|_q \left[\frac{1}{C_2} \|g_2(i)\|_p, \right] \quad (6\text{-}36$$

Adding the above two inequalities

$$\left| \sum_{i=0}^{N_o} \sigma^*(i) u(i) \right| + \left| \sum_{k=0}^{N_o} \gamma_k^* g_2(k) \right| \leq C_1 \|\sigma^*(i)\|_q \left[\frac{1}{C_1} \|u(i)\|_p \right] \quad (6\text{-}37)$$

$$+ C_2 \|\gamma^*(i)\|_q, \left[\frac{1}{C_2} \|g_2(i)\|_p, \right]$$

From

$$ab + cd \leq \left\{ \max \left\{ b, d \right\} \right\} [a + c] \qquad a, b, c, d > 0 \quad (6\text{-}38)$$

it follows that

$$C_1 \|\sigma^*(i)\|_q \left[\frac{1}{C_1} \|u(i)\|_p \right] + C_2 \|\gamma^*(i)\|_q, \left[\frac{1}{C_2} \|g_2(i)\|_p, \right] \quad (6\text{-}39)$$

$$\max \left\{ \frac{1}{C_1} \|u(i)\|_p, \frac{1}{C_2} \|g_2(i)\|_p, \right\} \left[C_1 \|\sigma^*(i)\|_q + C_2 \|\gamma^*(i)\|_q, \right]$$

Comparing (6-33) and (6-34) through (6-39) one finds that all inequalities in (6-34) through (6-39) must be satisfied by equality. A necessary condition for equality in (6-33) is

$$u(i) = K |\sigma^*(i)|^{q-1} \operatorname{sgn} \sigma^*(i) \quad (6\text{-}40)$$

where $\mathrm{sgn}[a] = 1$ if $a > 0$, $\mathrm{sgn}[a] = -1$ if $a < 0$ and $\mathrm{sgn}[a]$ is arbitrary with its magnitude less than 1 when $a = 0$. K is an arbitrary nonzero constant.

A necessary condition for equality in (6-38) is $b = d$ so for equality in (6-39)

$$\frac{1}{C_1}\|u(i)\|_p = \frac{1}{C_2}\|g_2(i)\|_p, \tag{6-41}$$

(6-41) is true provided $\|\gamma^*(i)\|_q, > 0$ and this condition holds whenever the output constraints are active, i.e., the constrained output solution differs from the unconstrained output solution. Therefore

$$\frac{1}{C_1}\|u(i)\|_p = \max\left\{\frac{1}{C_1}\|u(i)\|_p, \frac{1}{C_2}\|g_2(i)\|_p,\right\} = \|f\| = A \tag{6-42}$$

At this point we see that, if $A < 1$, we are driving the system to the desired terminal condition while satisfying the constraints,

$$\|u(i)\|_p \leq AC_2 , \quad \|y_2(i) - y_2^d(i)\|_p, \leq AC_2$$

From (6-40) and (6-42)

$$\|u(i)\|_p = K\left[\sum_{i=0}^{N_o}\left(|\sigma^*(i)|^{q-1}\right)^p\right]^{1/p} = K\left[\sum_{i=0}^{N_o}|\sigma^*(i)|^q\right]^{1/p} \tag{6-43}$$

$$= K\|\sigma^*(i)\|_p^{q-1} = AC_1$$

or

$$K = AC_1 \| \sigma^*(i) \|_q^{1-q} \tag{6-44}$$

and

$$u(i) = AC_1 \| \sigma^*(i) \|_q^{1-q} | \sigma^*(i) |^{q-1} \operatorname{sgn}[\sigma^*(i)] \tag{6-45}$$

The above expression is the explicit form of the optimal input sequence.

6.5 Special Considerations for Amplitude Constraints

In the case when there is an amplitude constraint on the control sequence $(p = \infty, q = 1)$, the control sequence is not defined whenever $\sigma^*(i) = 0$.

However, the existence of an optimal control at $N = N_o$ has already been established and the following procedure can be used to obtain the control at those steps where $\sigma^*(i) = 0$ provided the number of such steps is not excessively large. The control is determined (and equals $\pm AC_1$) when $\sigma^*(i) \neq 0$ and for those points at which $\sigma^*(i) = 0$ we form a suitable grid for the interval $[-AC_1, +AC_1]$. A search is performed over the grid to find the values of the control which satisfy both the terminal and output constraints by finding the outputs of the system when the determined part of the input sequence and the chosen grid point are

applied as the system input. Each step at which $\sigma^*(i) = 0$ adds another dimension to the search and consequently if there are many such steps, the suggested procedure may become too time-consuming.

For the case when there is a limitation on the maximum deviation of the output from some desired trajectory it is frequently possible to reduce the search time. Suppose that at the j^{th} step $\gamma^*(j) \neq 0$. The fact that $\gamma^*(j) \neq 0$ implies that the output is on the boundary of its constraint at the j^{th} step, i.e.,

$$|y_2(j) - y_2^d(j)| = AC_2 \qquad (6\text{-}46)$$

Therefore, for all $i \leq j$ for which $\sigma^*(i) = 0$ we need only consider values of the input which satisfy (6-46). If for a particular selection of input values at the steps where $\sigma^*(i) = 0$, $i \leq j$, we reach the j^{th} step and find that (6-46) is not satisfied, then the chosen input values are not part of an optimal input sequence. It is clear that the reduction in search time is obtained by eliminating input sequences which merely satisfy the output constraint at the j^{th} step and examining only those input sequences which force the output to the boundary of its constraint. In some cases (as shown in the example below) the values of the control at certain steps of the input sequence are determined directly from (6-46).

6.6 Special Considerations for Area Constraints

In the case when the input sequence has a constraint on its area ($p = 1$, $q = \infty$) expression (6-45) must be given a proper interpretation. This can be accomplished by using a limiting process as done by Sarachik and Kranc[35] in their paper considering discrete systems without output constraints, but here we revert directly back to the Holder equality conditions. Inequality (6-35) still must be satisfied by equality and if $i_1, i_2, \ldots, i_p$, are those values of i for which $\sigma^*(i)$ attains its maximum magnitude, i.e.,

$$|\sigma^*(i_j)| = \max_{0 \leq i \leq N_o} |\sigma^*(i)| \qquad j = 1, \ldots, p \tag{6-47}$$

then any assignment of values to $u(i)$ such that

$$\begin{aligned} u(i) &= 0 && i \neq i_1, \ldots, i_p \\ u(i_j) &= a_j \operatorname{sgn}[\sigma^*(i_j)] && j = 1, \ldots, p \qquad a_j \geq 0 \end{aligned} \tag{6-48}$$

meets this condition. Since (6-42) remains true the only remaining requirement on the input sequence is that

$$\sum_{j=1}^{p} |a_j| = AC_1 \tag{6-49}$$

Therefore any input sequence satisfying (6-48) and (6-49) is an optimal input sequence.

6.7 Example

As an illustration of the above procedure, consider the system described by the following difference equations

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \\ x_3(n+1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{4} \\ \frac{1}{2} \end{bmatrix} u(n) \qquad (6\text{-}50)$$

$$\begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} \qquad (6\text{-}51)$$

Where $u(n)$ is a scalar input. Let $x_1(0) = \frac{1}{3}$, $x_2(0) = 0$, $x_3(0) = 0$ and it is required that y_1 return to the origin in the least number of steps with the constraint that in the transition process the following inequalities remain valid:

$$|u(n)| \leq \frac{2}{3}, \quad y_2(n) \leq \frac{1}{3} \qquad n = 0, 1, \ldots$$

The input-output relations for this system are

$$\begin{bmatrix} y_1(N) \\ y_2(N) \end{bmatrix} = \begin{bmatrix} x_1(0) \\ 2^{-N}x_2(0) \end{bmatrix} + \sum_{i=0}^{N-1} \begin{bmatrix} \frac{1}{4}[2 - 2^{-N+i+2}] \\ \frac{1}{2}\, 2^{-N+i+1} \end{bmatrix} u(i)$$

The first value of N for which expression (6-29) $\left(y_2^d(i) \equiv 0\right)$ is equal or less than unity is $N = 3$ for which it equals unity. One set of minimizing parameters are $\alpha = -3$, $\gamma_1 = \frac{1}{2}$, $\gamma_2 = \frac{3}{2}$, $\gamma_3 = 0$. When these parameters are substituted in (6-45) we find that $u(0) = -\frac{2}{3}$ and $\sigma^*(1) = \sigma^*(2) = 0$ so that this formula does not determine $u(1)$ and $u(2)$. However, γ_1, γ_2 are not zero so that

$$y_2(1) = \pm \frac{1}{3} \qquad y_2(2) = \pm \frac{1}{3}$$

By examining (6-50) and (6-51) we see that the input at the i^{th} step, $u(i)$ has no effect on $y_1(i)$ or $y_2(i)$. Application of $u(0) = -\frac{2}{3}$ yields $y_2(1) = -\frac{1}{3}$. Since

$$y_2(2) = \frac{1}{2}\, y_2(1) + \frac{1}{2}\, u(1) = \pm \frac{1}{3}$$

we obtain

$$u(1) = 2\left(\frac{1}{3} - \frac{1}{2}\left(-\frac{1}{3}\right)\right) = 1$$

or

$$u(1) = 2\left(-\frac{1}{3} - \frac{1}{2}\left(-\frac{1}{3}\right)\right) = -\frac{1}{3}$$

Since the first value violates the control constraint, we must have $u(1) = -\frac{1}{3}$. The value of $u(2)$ is determined by a search to be equal to zero. The value of $u(3)$ need

not be determined since it does not affect the output at the third step. The output sequence $y_1(i)$ is given by

$$y_1(0) = \frac{1}{3}, \qquad y_1(1) = \frac{1}{3}, \qquad y_1(2) = \frac{1}{6}, \qquad y_1(3) = 0$$

while the output sequence $y_2(i)$ is

$$y_2(0) = 0, \qquad y_2(1) = -\frac{1}{3}, \qquad y_2(2) = -\frac{1}{3}, \qquad y_2(3) = -\frac{1}{6}$$

CHAPTER 7. NUMERICAL CONSIDERATIONS

We now discuss the numerical procedures used in the preceding chapters. First, some theoretical properties of the functions we are trying to minimize are derived and then we examine some of the practical aspects of performing the minimization. No new algorithms are presented.

7.1 Convexity

For the discrete point approximation method of Chapter 3, the approximate solution is obtained from the minimization of expression (3-19) which is repeated below for convenience

$$\min_{\alpha,\gamma_i} \left\{ C_1 \left[\int_0^T \left| \alpha h_1(T,\tau) + \sum_{i=1}^{p} \gamma_i h_2(t_i,\tau) \right|^q d\tau \right]^{1/q} + C_2 \sum_{i=1}^{p} |\gamma_i| \right\}$$

$$\alpha c + \sum_{i=1}^{p} \gamma_i d_i = 1$$

This is a constrained minimization problem however, since there is only a single equality constraint, we can eliminate the variable α (assuming $c \neq 0$) to reduce the above expression to

$$\min_{\gamma_i}\left\{C_1\left[\int_0^T \left|\frac{1}{c}\, h_1(T,\tau) + \sum_{i=1}^{p}\gamma_i\left[h_2(t_k,\tau) - \frac{d_i}{c}\, h_1(T,\tau)\right]\right|^q d\tau\right]^{1/q} + C_2 \sum_{i=1}^{p}|\gamma_i|\right\} \tag{7-1}$$

which is an unconstrained problem. Therefore, in the following, we restrict our discussion to minimization problems of the type (7-1). The basic property of the term in the braces of (7-1) is that it is a convex function of its arguments as we prove in the next theorem.

Theorem 7.1

Consider the function of the variables $\gamma_1, \ldots \gamma_p$ defined by

$$C_1\left[\int_0^T \left|f(\tau) + \sum_{i=1}^{p}\gamma_i g_i(\tau)\right|^q d\tau\right]^{1/q} + C_2 \sum_{i=1}^{p}|\gamma_i| \qquad q \geq 1 \tag{7-2}$$

where the functions $f(\tau)$, $g_1(\tau)$, ..., $g_p(\tau)$ are bounded piecewise continuous functions of their arguments. Then (7-2) is a convex function of the variables $\gamma_1, \ldots, \gamma_p$.

Proof: A function $y(\gamma_1, \ldots, \gamma_p)$ is convex if it is never underestimated by a linear interpolation between any

two points, that is, if $(\bar{\gamma}_1, \ldots, \bar{\gamma}_p)$ and $(\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ represent any two points and if for every number α satisfying

$$0 < \alpha < 1$$

then it is true that

$$y\langle \alpha\bar{\gamma}_1 + (1 - \alpha)\hat{\gamma}_1, \ldots, \alpha\bar{\gamma}_p + (1 - \alpha)\hat{\gamma}_p \rangle \leq \alpha y\langle \bar{\gamma}_1, \ldots, \bar{\gamma}_p \rangle + (1 - \alpha) y\langle \hat{\gamma}_1, \ldots, \hat{\gamma}_p \rangle$$

From the definition of the norm in $L_q[0,T]$

$$\left[\int_0^T |a(\tau) + b(\tau)|^q d\tau\right]^{1/q} \leq \left[\int_0^T |a(\tau)|^q d\tau\right]^{1/q} + \left[\int_0^T |b(\tau)|^q d\tau\right]^{1/q} \tag{7-3}$$

Therefore

$$C_1\left[\int_0^T |f(\tau) + \sum_{i=1}^{p}\left[\alpha\bar{\gamma}_i + (1-\alpha)\hat{\gamma}_i\right]g_i(\tau)|^q d\tau\right]^{1/q}$$

$$= C_1\left[\int_0^T |\alpha f(\tau) + \sum_{i=1}^{p}\alpha\bar{\gamma}_i g_i(\tau) + (1-\alpha)f(\tau) + \sum_{i=1}^{p}(1-\alpha)\hat{\gamma}_i g_i(\tau)|^q d\tau\right]^{1/q} \qquad (7\text{-}4)$$

$$\leq \alpha C_1\left[\int_0^T |f(\tau) + \sum_{i=1}^{p}\bar{\gamma}_i g(\tau)|^q d\tau\right]^{1/q} + (1-\alpha)\left[\int_0^T |f(\tau) + \sum_{i=1}^{p}\hat{\gamma}_i g(\tau)|^q d\tau\right]^{1/q}$$

Also

$$C_2 \sum_{i=1}^{p} |\alpha\bar{\gamma}_i + (1-\alpha)\hat{\gamma}_i| \leq \alpha C_2 \sum_{i=1}^{p} |\bar{\gamma}_i| + (1-\alpha)C_2 \sum_{i=1}^{p} |\hat{\gamma}_i| \qquad (7\text{-}5)$$

Adding inequalities (7-4) and (7-5) we obtain the statement of the theorem.

The importance of the convexity of (7-2) is that it assures us that any point which is a local minimum of (7-2) is also a global minimum.[37] Since practically all of the minimization schemes in present use converge only to a local minimum, the convexity property yields that this local minimum is indeed a global minimum and therefore may be used to generate the optimal control by (2-40).

7.2 Minimization Schemes Requiring Defined Gradients

There are many techniques one can use to perform the minimization required in (7-1); the two that were used in this dissertation were the methods of Fletcher and Powell[38] and Fletcher and Reeves.[39] It is convenient to group functions into two main classes according to whether the gradient vector is defined analytically at each point or must be estimated from the differences of values of the function. The two schemes mentioned above require an analytically defined gradient and we now show with the aid of the proposition below that under suitable restrictions these schemes can be applied to the minimization (7-1). The proof of the proposition is essentially one given by Neustadt.[40]

Proposition 7.1: Consider the function of the variables $\gamma_1, \ldots, \gamma_p$ defined by

$$z(\gamma_1, \ldots, \gamma_p) = \int_0^T \left| f(\tau) + \sum_{i=1}^{p} \gamma_i g_i(\tau) \right| d\tau \qquad (7\text{-}6)$$

where the functions $f(\tau)$, $g_1(\tau)$, ..., $g_p(\tau)$ are bounded piecewise continuous functions of their arguments. If, for any values of the variables γ_1, ..., γ_p,

$$f(\tau) + \sum_{i=1}^{p} \gamma_i g_i(\tau) = 0 \qquad (7\text{-}7)$$

only on a set of measure 0 in $[0,T]$, then the function $z(\gamma_1, \ldots, \gamma_p)$ has continuous partial derivations with respect to γ_1, ..., γ_p

Proof: Let

$$\sigma(\gamma_1, \ldots, \gamma_p) = f(\tau) + \sum_{i=1}^{p} \gamma_i g_i(\tau)$$

then

$$\Delta_j z(\gamma_1, \ldots, \gamma_p) = z(\gamma_1, \ldots, \gamma_j + \Delta\gamma_j, \ldots, \gamma_p) - z(\gamma_1, \ldots, \gamma_p)$$

$$= \int_0^T \Big\{ \left| \sigma(\gamma_1, \ldots, \gamma_p, \tau) + \Delta\gamma_j g_j(\tau) \right| \qquad (7\text{-}8)$$

$$- \left| \sigma(\gamma_1, \ldots, \gamma_p, \tau) \right| \Big\} \, d\tau$$

From assumption (7-7) the set $\{\tau: \sigma(\gamma_1, \ldots, \gamma_p, \tau) = 0\}$ has measure zero for any $\gamma_1, \ldots, \gamma_p$. Hence, for any $\epsilon > 0$, there exists a $\delta > 0$ such that the set $A_o = \{\tau: |\sigma(\gamma_1, \ldots, \gamma_p, \tau)| < \delta\}$ has measure less than ϵ. Let

$$A_+ = \left\{\tau: \sigma(\gamma_1, \ldots, \gamma_p, \tau) \geq \delta\right\}$$

$$A_- = \left\{\tau: \sigma(\gamma_1, \ldots, \gamma_p, \tau) \leq -\delta\right\}$$

Also, let

$$M = \max_{0 \leq \tau \leq T} |g_j(\tau)|$$

If we choose $\Delta\gamma_j$ such that $|\Delta\gamma_j| < \frac{\delta}{M}$, then

$$|\Delta\gamma_j g_j(\tau)| < \delta$$

and we obtain on A_+ that

$$|\sigma(\gamma_1, \ldots, \gamma_p, \tau) + \Delta\gamma_j g_j(\tau)| - |\sigma(\gamma_1, \ldots, \gamma_p, \tau)| =$$

$$+ \Delta\gamma_j g_j(\tau)$$

and on A_- that

$$|\sigma(\gamma_1, \ldots, \gamma_p, \tau) + \Delta\gamma_j g_j(\tau)| - |\sigma(\gamma_1, \ldots, \gamma_p, \tau)| =$$

$$- \Delta\gamma_j g_j(\tau)$$

Rewriting (7-8), we have

$$\Delta_j z(\gamma_1, \ldots, \gamma_p) = \Delta\gamma_j \int_{A_+ \cup A_-} g_j(\tau)\,\mathrm{sgn}[\sigma(\gamma_1, \ldots, \gamma_p, \tau)]d\tau$$

$$+ \int_{A_o} \Big\{ |\sigma(\gamma_1, \ldots, \gamma_p, \tau) + \Delta\gamma_j g_j(\tau)|$$

$$- |\sigma(\gamma_1, \ldots, \gamma_p, \tau)| \Big\}\, d\tau$$

The second integral is less in absolute value than $|\Delta\gamma_j| M\epsilon$. Hence

$$\left| \frac{\Delta_j z(\gamma_1, \ldots, \gamma_p)}{\Delta\gamma_j} - \int_0^T g_j(\tau)\,\mathrm{sgn}[\sigma(\gamma_1, \ldots, \gamma_p, \tau)]d\tau \right|$$

$$\leq \int_{A_o} |g_j(\tau)|\,d\tau + M\epsilon \leq 2M\epsilon$$

Letting $\Delta\gamma_j \to 0$

$$\frac{\partial z(\gamma_1, \ldots, \gamma_p)}{\partial\gamma_j} = \int_0^T g_j(\tau)\,\mathrm{sgn}[\sigma(\gamma_1, \ldots, \gamma_p, \tau)]d\tau \qquad (7\text{-}9)$$

for $j = 1, \ldots, p$. The continuity of (7-9) follows from assumption (7-7). Similar arguments suffice to prove the continuity of the other partial derivatives and the proposition is obtained

This proposition shows that for a magnitude constraint on the control $(q = 1)$, the expression

$$C_1\left[\int_0^T \left|\frac{1}{c}\, h_1(T,\tau) + \sum_{i=1}^{p} \gamma_i\left[h_2(t_i,\tau) - \frac{d_i}{c}\, h_1(T,\tau)\right]\right|^q d\tau\right]^{1/q} \qquad (7\text{-}10)$$

has continuous partial derivatives with respect to $\gamma_1, \ldots, \gamma_p$. This is also true for an energy constraint on the control $(q = 2)$. We restrict ourselves to these two cases. Since

$$\frac{d}{da}|a| = \operatorname{sgn}[a] \qquad a \neq 0$$

and the derivative is undefined when $a = 0$, it follows that (7-1) has an analytically defined gradient everywhere except for those points where at least one γ_i equals zero. If we define

$$\left.\frac{d}{da}|a|\right|_{q=0} = 1$$

then (7-1) has an everywhere defined although discontinuous gradient. The minimization schemes of Fletcher and Powell

and Fletcher and Reeves both converged to the global minimum of (7-1) for all examples of Chapters 3 and 4. Even for some test problems for which some of the constraints were not active, and hence $\gamma_i = 0$ for some i, convergence was obtained although the computational time was somewhat greater.

When solving time optimal bounded output problems for discrete systems with magnitude constraints on the control the expression analogous to (7-1) from which the solution is obtained is found from (6-29) to be

$$\min_{\gamma_i} \left\{ C_1 \left| \sum_{i=0}^{N} \left| \frac{1}{c} H_1(N,i) + \sum_{k=0}^{N} \gamma_k \left[H_2(k,i) - \frac{d_i}{c} H_1(N,i) \right] \right| \right| + C_2 \left[\sum_{k=0}^{N} |\gamma_k|^{q'} \right]^{1/q'} \right\} \tag{7-11}$$

In contrast to the situation for continuous systems, the first part of this expression

$$C_1 \left| \sum_{i=0}^{N} \left| \frac{1}{c} H_1(N,i) + \sum_{k=0}^{N} \gamma_k \left[H_2(k,i) - \frac{d_i}{c} H_1(N,i) \right] \right| \right| \tag{7-12}$$

does not possess continuous derivatives. Specifically (7-12) does not have its derivative defined at those values of $\gamma_1, \ldots, \gamma_N$ for which

$$\frac{1}{c} H_1(N,i) + \sum_{k=0}^{N} \gamma_k \left[H_2(k,i) - \frac{d_i}{c} H_1(N,i) \right] = 0 \quad \text{for some} \quad i \qquad (7\text{-}13)$$

However, the expression in braces in (7-11) is still a convex function of the variables $\gamma_1, \ldots, \gamma_N$ and it seems that a finite difference method such as the one proposed by Powell[41] should converge reasonably well for problems of this type.

7.3 Practical Aspects of the Search

The methods of Fletcher and Powell and Fletcher and Reeves are both available in the IBM supplied Scientific Subroutine Package so that the programming required for the search procedure is relatively simple. The user must only supply a subroutine which yields function and gradient information for given values of the arguments.

One point to be noted is that the integrands appearing in the function and gradient evaluations are discontinuous. Since most of the standard integration procedures are computationally accurate only when the integrand is continuous, we should first locate the integrand discontinuities, then subdivide the interval of integration such that the integrand is continuous on each subdivision and then sum the results of the integrations on the subintervals to obtain the total integral. The method used for locating discon-

tinuities was an initial rough search followed by three refinement searches. The initial search is used to isolate the discontinuities and the number of subintervals into which the original interval is divided should be as small as possible consistent with this objective. For the examples worked out in this thesis, two hundred subintervals were found to be sufficient. Once a discontinuity was located in a certain subinterval, a series of three searches of five increments each was employed to narrow the interval of uncertainty.

The optimal time was also located by a search procedure. A value of T was chosen, the minimization (7-1) was performed and if its value was less than one T was increased, if its value was greater than one T was decreased. Once the optimal time was bracketed by values of T for which (7-1) was less than and greater than one respectively, a bisection method was used to find the optimal time to the desired degree of accuracy. This method was used because the initial guess for T was taken as slightly greater than the optimal time for the problem without output constraints. This latter time provides a lower bound on the optimal time for the constrained output problem.

The computational time for the examples in this thesis which were worked out on the computer was approximately two minutes.

REFERENCES

1. Kreindler, E., "Contributions to the Theory of Time Optimal Control," J. Franklin Institute, Vol. 275, April 1963, pp. 314-344.
2. Chang, S. L., "Optimal Control in Bounded Phase Space," Automatica, Vol. 1, No. 1, Jan.-Mar. 1963, pp. 55-67.
3. Coddington, E. A. and N. Levinson, "Theory of Ordinary Differential Equations," McGraw-Hill Book Company, Inc., New York, 1955.
4. Berkovitz, L. D., "On Control Problems with Bounded State Variables," Jour. Math. Anal. and Appl., Vol. 5, No. 3, December 1962, pp. 488-498.
5. Bryson, A. E., W. E. Denham, and S. E. Dreyfus, "Optimal Programming Problems with Inequality Constraints I: Necessary Conditions for External Solutions," AIAA Journal, Vol. 1, No. 11, November 1963, pp. 2544-2550.
6. Pontryagin et al., "The Mathematical Theory of Optimal Processes," Interscience Publishing Co., New York, New York, 1962, Chapter 6.
7. Chang, S. L., "An Extension of Ascoli's Theorem and Its Applications to the Theory of Optimal Control," Tech. Report No. 400-51, Dept. Elec. Eng., New York Univ., January 1962.

8. Chang, S. L., "Minimal Time Control with Multiple Saturation Limits," IEEE Trans. Auto. Control, Vol. 8, No. 1, 1963.

9. Dreyfus, S. E., "Variational Problems with Inequality Constraints," Jour. Math. Anal. and Appl., Vol. 4, 1962, pp. 297-308.

10. Speyer, J. L. and A. E. Bryson, "Optimal Programming Problems with a Bounded State Space," Proc. JACC, 1968, pp. 485-491.

11. Lee, E. B. and L. Markus, "Foundations of Optimal Control Theory," John Wiley and Sons, Inc., New York, N.Y., 1967.

12. Denham, W. F. and A. E. Bryson, "Optimal Programming Problems with Inequality Constraints II: Solution by Steepest Ascent," AIAA Journal, Vol. 2, No. 1, January 1964, pp. 25-34.

13. Courant, R., "Calculus of Variations and Supplementary Notes and Exercises 1945-1946," revised and amended by J. Moser, New York University Institute of Math. Sciences, 1956-57.

14. Kelley, H. J., "Method of Gradients" in "Optimization Techniques," edited by G. Leitmann, Academic Press, Inc., New York 1962, Chapter 6.

15. McGill R., "Optimal Control, Inequality State Constraints, and the Generalized Newton-Raphson Algorithm," J. SIAM, Ser. A., Control 3, 1965, pp. 291-298.

16. Kirin, N. W., "An Iterative Method for the Solution of Extremal Problems," Automat. Remote Control, Vol. 27, No. 10, 1966, pp. 5-12.

17. Lee, E. B., "An Approximation to Linear Bounded Phase Coordinate Control Problems," Jour. Math. Anal. and Appl., Vol. 13, March 1966, pp. 550-564.

18. Lasdon, L. S., A. D. Waren, and R. K. Rice, "An Interior Penalty Method for Inequality Constrained Optimal Control Systems," IEEE Trans. Auto. Control, Vol. 12, No. 4, August 1967, pp. 388-394.

19. Russell, O. L., "Penalty Functions and Bounded Phase Coordinate Control," J. Soc. Ind. Appl. Math., Ser. A, Control, Vol. 2, No. 3, 1965, pp. 409-422.

20. Okamur, K., "Some Mathematical Theory of the Penalty Method for Solving Optimum Control Problems," J. Soc. Ind. Appl. Math, Ser. A, Control, Vol. 2, No. 3, 1965, pp. 317-331.

21. Ho, Y. C. and P. B. Brentani, "On Computing Optimal Control with Inequality Constraints," J. Soc. Ind. Appl. Math., Control, Ser. A, Vol. 1, No. 3, 1963, pp. 319-348.

22. Negata, A., S. Kodama, and S. Kumagai, "Time-Optimal Discrete Control Systems with Bounded State Variables," IEEE Trans. Auto. Control, Vol. 9, 1965, pp. 155-164.

23. Fath, A. F., "Approximation to the Time-Optimal Control of Linear State-Constrained Systems," Proc. JACC, 1968, pp. 962-969.

24. Gabasov, R., "On Optimal Processes in Coupled Digital Systems," Automat. Remote Control, Vol. 23, No. 7, 1962, pp. 808-817.

25. Gabasov, R. and F. M. Kirillova, "Optimum Control of Linked Discrete Systems," Automat. Remote Control, Vol. 24, No. 7, 1963, pp. 825-830.

26. Gabasov, R. and F. M. Kirillova, "Optimal Control Problems," Engineering Cybernetics, No. 1, 1964, pp. 111-120.

27. Katz, S. and G. M. Kranc, "On the Least-Time Control Problem with Interior Output Constraints," IEEE Trans. Auto. Control, Vol. 14, 1969, pp. 255-261.

28. Sarachik, P. E. and G. M. Kranc, "On Optimal Control of Systems with Multi-Norm Constraints," Proc. IFAC Constraints," Proc. IFAC Congress, Basel, 423 (1963).

29. Liusternik, L. A. and V. J. Sobolev, "Elements of Functional Analysis," Frederick Ungar Publishing Co., New York, 1961.

30. Akhiezer, N. I. and M. Krein, "Some Questions in the Theory of Moments," Article IV, Amer. Math. Soc. Publ. 1962.

31. Zaanen, A. C., "Linear Analysis," Interscience Publishers, Inc., New York, 1956, pp. 126-127.

32. Natanson, I. P., "Theory of Functions of a Real Variable," Vol. 1, Frederick Ungar Publishing Co., New York, 1955.

33. Riesz, F. and B. Sz-Nagy, "Functional Analysis," Frederick Ungar Publishing Co., New York, 1955.

34. Kranc, G. M. and P. E. Sarachik, "An Application of Functional Analysis to the Optimal Control Problem," Trans. A.S.M.E., Jour. of Basic Eng., Vol. 85, pp. 143-150, 1963.

35. Sarachik, P. E. and G. M. Kranc, "Optimal Control of Discrete Systems with Constrained Inputs," Journ. Franklin Institute, Vol. 277, No. 3, March, 1969, pp. 237-255.

36. Kreindler, E. and P. E. Sarachik, "On the Concepts of Controllability and Observability of Linear Systems," IEEE Trans. Auto. Control, Vol. 9, No. 2, 1964, pp. 129-136.

37. Wilde, D. J. and C. S. Beightler, "Foundations of Optimization," Prentice Hall, Inc., Englewood Cliffs, N. J., 1967.

38. Fletcher, R. and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," British Computer J., Vol. 6, 1963, pp. 163-168.

39. Fletcher, R. and C. M. Reeves, "Function Minimization by Conjugate Gradients," British Computer J., Vol. 7, 1969, pp. 149-154.

40. Neustadt, L. W., "Synthesizing Time Optimal Control Systems," J. Math. Anal. Appl., Vol. 1, pp. 484-493 (1960).

41. Powell, M. J. D., "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives," British Computer J., Vol. 7, (1964) pp. 155-162.

42. Kantorovich, L. V. and G. P. Akilov, "Functional Analysis in Normed Spaces," Pergamon Press, New York, 1964.

APPENDIX I

Given two Banach spaces X and Y, we form the set $X \times Y$ of all pairs (x,y) $(x \in X, y \in Y)$. When linearized in the natural way, the set $X \times Y$ becomes a Banach space on introducing the norm

$$\|(x,y)\| = \|x\| + \|y\| \tag{A1-1}$$

As shown in Kantorovich[42] every linear functional f on $X \times Y$ has a representation

$$f[(x,y)] = f_x(x) + f_y(y) \tag{A1-2}$$

where $f_x(f_y)$ is a linear functional on $X(Y)$.

Let the norm of f be defined as

$$\|f\| = \sup \frac{|f[(x,y)]|}{\|(x,y)\|} \tag{A1-3}$$

Then since

$$|f_x(x)| \leq \|f_x\|\|x\| \qquad |f_y(y)| \leq \|f_y\|\|y\| \tag{A1-4}$$

we obtain

$$\begin{aligned} \|f\| = \sup \frac{|f_x(x) + f_y(y)|}{\|(x,y)\|} &\leq \sup \frac{|f_x(x)| + |f_y(y)|}{\|(x,y)\|} \\ &\leq \sup \frac{\|f_x\|\|x\| + \|f_y\|\|y\|}{\|(x,y)\|} \qquad \text{(A1-5)} \\ \leq \sup \frac{\{\max(\|f_x\|, \|f_y\|)\}\{\|x\| + \|y\|\}}{\|(x,y)\|} &= \max(\|f_x\|, \|f_y\|) \end{aligned}$$

We now assume $\|f_x\| \geq \|f_y\|$. The arguments to follow are easily modified if $\|f_x\| < \|f_y\|$. From the definition of the norm of a functional, for any ϵ there exists an element $x_1 \in X$ such that

$$|f_x(x_1)| > \|f_x\|\|x_1\| - \epsilon \|x_1\|$$

If we now consider the element $(x_1, 0_y)$ (0_y is the zero element of the space Y) then

$$\|f\| = \sup \frac{|f[(x,y)]|}{\|(x,y)\|} \geq \frac{|f[(x_1,0_y)]|}{\|(x_1,0_y)\|} = \frac{|f_x(x_1)|}{\|x_1\|}$$

$$> \|f_x\| - \epsilon = \max \left\{\|f_x\|, \|f_y\|\right\} - \epsilon$$

Since ϵ is arbitrary

$$\|f\| \geq \max \left\{\|f_x\|, \|f_y\|\right\} \qquad \text{(A1-6)}$$

Inequalities (A1-5) and (A1-6) yield that

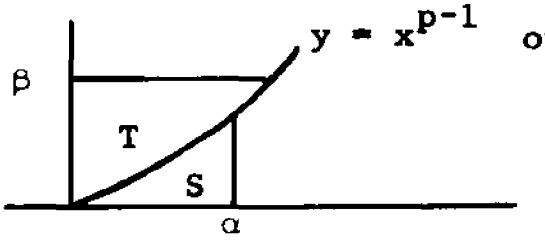
$$\|f\| = \max \left\{\|f_x\|, \|f_y\|\right\}$$

Equations (2-15) and (2-17) follow by associating the space $L_q[0,T]$ with X and $\ell_1^{(n)}$ with Y and making use of the representation of functionals on these particular spaces.[29]

APPENDIX II

In this appendix we derive various forms of the Holder inequality needed in the main body of the thesis. Depending upon the form, this inequality can be considered a generalization of the Cauchy and Schwartz inequalities.

Consider the function $y = x^{p-1}$, $p > 1$



$y = x^{p-1}$ or $x = y^{\frac{1}{p-1}} = y^{q-1}$

Clearly, area S + area T $\geq \alpha\beta$ with equality if and only if $\beta = \alpha^{p-1}$. Now

$$\text{Area } S = \int_0^{\alpha} x^{p-1}dx = \frac{\alpha^p}{p} \tag{A2-1}$$

$$\text{Area } T = \int_0^{\beta} y^{q-1}dy = \frac{\beta^q}{q}$$

Therefore

$$\alpha\beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q} \tag{A2-2}$$

Let

$$\alpha = \frac{|x(t)|}{\|x\|_p}\ ,\ \|x\|_p = \left[\int_0^T |x(t)|^p dt\right]^{1/p}\ ,$$

$$\beta = \frac{|y(t)|}{\|y\|_q}\ ,\ \|y\|_q = \left[\int_0^T |y(t)|^q dt\right]^{1/q} \qquad \text{(A2-3)}$$

Then, substituting (A2-3) into (A2-2) integrating both sides and using $|x(t)y(t)| = |x(t)y(t)|$

$$\frac{\int_0^T |x(t)y(t)|dt}{\|x\|_p\|y\|_q} \le \frac{\int_0^T |x(t)|^p dt}{p\|x\|_p^p} + \frac{\int_0^T |y(t)|^q dt}{q\|y\|_q^q} = \frac{1}{p} + \frac{1}{q} = 1 \qquad \text{(A2-4)}$$

or

$$\int_0^T |x(t)y(t)dt \le \|x\|_p\|y\|_q \qquad \text{(A2-5)}$$

with equality if and only if

$$|y(t)| = \frac{\|y\|_q}{\|x\|_p^{p-1}}\ |x(t)|^{p-1} \qquad \text{(A2-6)}$$

Moreover, if

$$|y(t)| = K|x(t)|^{p-1} \tag{A2-7}$$

for any positive constant K, then y(t) in (A2-7) satisfies condition (A2-6). Since

$$\left|\int_0^T x(t)y(t)dt\right| \leq \int_0^T |x(t)y(t)|dt \tag{A2-8}$$

with equality if and only if

$$\text{sgn } y(t) = \pm \text{ sgn } x(t) \tag{A2-9}$$

one obtains, combining (A2-5), (A2-7), (A2-9), that

$$\left|\int_0^T x(t)y(t)dt\right| \leq \|x\|_p \|y\|_q \tag{A2-10}$$

with equality if and only if

$$y(t) = K|x(t)|^{q-1} \text{ sgn } x(t) \tag{A2-11}$$

where K is a nonzero constant.

Letting

$$\alpha = \frac{|a_i|}{\|\underline{a}\|_p}, \quad \underline{a} = [a_1, a_2, \ldots], \quad \|\underline{a}\|_p = \left(\sum_{i=1}^{\infty} |a_i|^p \right)^{1/p},$$

$$\beta = \frac{|b_i|}{\|\underline{b}\|_q}, \quad \underline{b} = [b_1, b_2, \ldots], \quad \|\underline{b}\|_q = \left(\sum_{i=1}^{\infty} |b_i|^q \right)^{1/q} \tag{A2-12}$$

then one obtains, by going through a similar analysis

$$\left| \sum_{i=1}^{\infty} a_i b_i \right| \leq \|\underline{a}\|_p \|\underline{b}\|_q \tag{A2-13}$$

with equality if and only if

$$b_i = K|a_i|^{q-1} \operatorname{sgn} a_i \quad \text{for all} \quad i \tag{A2-14}$$

with K a nonzero constant.

APPENDIX III

Consider the function $f(T)$ defined by

$$f(T) = \max_{\lambda_1, \alpha_1} \frac{\lambda_1 x^1(T) + \alpha_1 d_1(T)}{\int_0^T |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)| d\tau + |\alpha_1|} \qquad 0 < r_1 < 1$$

where $x^1(T)$ and $d_1(T)$ are continuous functions of T and $h_1(t,\tau)$ and $h_2(t,\tau)$ are continuous functions of their arguments in $[0,A] \times [0,A]$ for all finite values of A except that $h_2(t,\tau)$ considered as a function of τ with t fixed at any finite value may have a simple discontinuity at $\tau = t$, that is

$$\lim_{\tau \to t-0} h_2(t,\tau) = a \,, \qquad \lim_{\tau \to t+0} h_2(t,\tau) = b$$

both exist where a is not necessarily equal to b. Also, assume that for any fixed T and r_1, there do not exist values of λ_1 and α_1 such that

$$\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau) \equiv 0$$

in the interval $[0,T]$. This implies

$$\int_0^T |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)| d\tau > 0$$

for any value of T.

Theorem: Under the above assumptions $f(T)$ is a continuous function of T.

Proof: We fix T at any arbitrary value and show that corresponding to any $\epsilon > 0$ there exists a $\delta' > 0$ such that

$$|f(T + \delta) - f(T)| < \epsilon \quad \text{for all} \quad |\delta| < \delta'$$

We can assume that the values of λ_1 and α_1 (λ_1^* and α_1^*) at which the maximum of $f(T)$ is attained (by dividing both numerator and denominator of $f(T)$ by the same constant multiplier if necessary) satisfies $|\lambda_1^*|^2 + |\alpha_1^*|^2 = 1$.

If $h_2(t,\tau)_{\tau=t}$ is defined to be equal to its left hand limit, i.e.,

$$h_2(t,\tau)|_{\tau=t} = \lim_{\tau \to t-0} h_2(t,\tau)$$

then $h_2(t,\tau)$ can be considered a continuous function of τ on the interval $[0,t]$ and a continuous function of both its arguments on the compact set $[t_1, t_1 + \gamma] \times [0, t_1]$ for any finite t_1, γ. Let $\delta > 0$. Then

$$f_1(T+\delta) = \max_{\lambda_1,\alpha_1} \frac{\lambda_1 x^1(T+\delta) + \alpha_1 d_1(T+\delta)}{\int_0^{T+\delta} |\lambda_1 h_1(T+\delta,\tau) + \alpha_1 h_2\big(r_1(T+\delta),\tau\big)|\,d\tau + |\alpha_1|} \tag{A3-1}$$

$$= \max_{\lambda_1,\alpha_1} \frac{\lambda_1 x^1(T+\delta) + \alpha_1 d_1(T+\delta)}{\int_0^{r_1 T} + \int_{r_1 T}^{r_1(T+\delta)} + \int_{r_1(T+\delta)}^{T} + \int_{T}^{T+\delta} + |\alpha_1|}$$

where the integrated in the above expressions is

$$|\lambda_1 h_1(T + \delta,\tau) + \alpha_1 h_2\big(r_1(T + \delta),\ \tau\big)|d\tau$$

Let A be some positive number and assume $T + \delta < A$. h_1 is continuous on the compact set $[0,A] \times [0,A]$ and therefore is uniformly continuous on the same set. Similarly h_2 is uniformly continuous on $[r_1T,\ A] \times [0,\ r_1T]$. Therefore corresponding to any $\epsilon_1 > 0$, there exists a $\delta_1 > 0$ such that

$$\begin{aligned} &|h_1(T + \delta,\tau) - h_1(T,\tau)| < \epsilon_1 \qquad \tau \in [0,T] \\ &|h_2\big(r_1(T+\delta),\tau - h_2(r_1T,\tau)| < \epsilon_1 \quad \tau \in [0,r_1T] \end{aligned} \tag{A3-2}$$

for all $0 \leq \delta < \delta_1$. It now follows that for all $0 \leq \delta < \delta_1$

$$\int_0^{r_1T} |\lambda_1 h_1(T + \delta,\tau) + \alpha_1 h_2\big(r_1(T + \delta),\tau\big)\ |d\tau \geq$$

$$\int_0^{r_1T} |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1T,\tau)|d\tau - \epsilon_1 r_1 T$$

$$\int_{r_1(T+\delta)}^{T} |\lambda_1 h_1(T+\delta,\tau) + \alpha_1 h_2\big(r_1(T+\delta),\tau\big)|\,d\tau =$$

$$\int_{r_1(T+\delta)}^{T} |\lambda_1 h_1(T+\delta,\tau)|\,d\tau$$

(A3-3)

$$\geq \int_{r_1(T+\delta)}^{T} |\lambda_1 h_1(T,\tau)|\,d\tau - \epsilon_1\big(T(1-r_1)+r_1\delta\big)$$

$$= \int_{r_1(T+\delta)}^{T} |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)|\,d\tau - \epsilon_1\big(T(1-r_1)+r_1\delta\big)$$

Also, it follows from the assumptions on h_1, h_2 that

$$\int_{r_1 T}^{r_1(T+\delta)} |\lambda_1 h_1(T+\delta,\tau) + \alpha_1 h_2\big(r_1(T+\delta),\tau\big)|\,d\tau \leq M_1 r_1 \delta$$

$$\int_{T}^{T+\delta} |\lambda_1 h_1(T+\delta,\tau) + \alpha_1 h_2\big(r_1(T+\delta),\tau\big)|\,d\tau \leq M_2\delta \qquad \text{(A3-4)}$$

$$\int_{r_1 T}^{r_1(T+\delta)} |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)|\,d\tau \leq M_3 r_1 \delta$$

where M_1, M_2, M_3 are positive constants bounding the respective integrands in the above inequalities (remember we need only consider λ_1, α_1 such that $|\lambda_1|^2 + |\alpha_1|^2 = 1$). From (A3-1) through (A3-4) for all $0 \leq \delta < \delta_1$

$$\max_{\lambda_1,\alpha_1} \frac{\lambda_1 x^1(T+\delta) + \alpha_1 d_1(T+\delta)}{\int_0^{r_1 T} + \int_{r_1 T}^{r_1(T+\delta)} + \int_{r_1(T+\delta)}^{T} + B} \geq f(T+\delta) \geq$$

$$\max_{\lambda_1,\alpha_1} \frac{\lambda_1 x^1(T+\delta) + \alpha_1 d_1(T+\delta)}{\int_0^{r_1 T} + \int_{r_1 T}^{r_1(T+\delta)} + \int_{r_1(T+\delta)}^{T} + C} \qquad \text{(A3-5)}$$

where

$$B = -\epsilon_1 r_1 T - M_3 r_1 \delta - \epsilon_1\big(T(1-r_1) + r_1\delta\big) + |\alpha_1|$$

$$C = \epsilon_1 r_1 T + M_1 r_1 \delta + \epsilon\big(T(1-r_1) + r_1\delta\big) + M_2\delta$$

and the integrand in the above expressions is

$$|\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)|\,d\tau$$

From the continuity of $x^1(T)$, $d_1(T)$ we have that for any $\epsilon_2 > 0$ there exists a $\delta_2 > 0$ such that for all $0 \leq \delta < \delta_2$

$$|x^1(T + \delta) - x^1(T)| < \epsilon_2$$

$$|d_1(T + \delta) - d_1(T)| < \epsilon_2 \tag{A3-6}$$

Therefore

$$\lambda_1 x^1(T) + \alpha_1 d_1(T) + |\lambda_1|\epsilon_2 + |\alpha_1|\epsilon_2 \geq \lambda_1 x^1(T+\delta)$$

$$+ \alpha_1 d_1(T+\delta) \geq \lambda_1 x^1(T) + \alpha_1 d_1(T) - |\lambda_1|\epsilon_2 - |\alpha_1|\epsilon_2 \tag{A3-7}$$

Now, given any fixed $\epsilon' > 0$, choose ϵ_1, ϵ_2, δ_3 such that

$$\epsilon_1\left(r_1 T + T(1 - r_1) + r_1(A - T)\right) < \frac{\epsilon'}{2}$$

$$\epsilon_2(|\lambda_1| + |\alpha_1|) < \epsilon'$$

$$\delta_3\left(r_1 \max\left\{M_1, M_3\right\} + M_2\right) < \frac{\epsilon'}{2}$$

find δ_1, δ_2 corresponding to ϵ_1, ϵ_2 and choose

$$\delta' = \min(\delta_1, \delta_2, \delta_3)$$

Then from (A3-4) through (A3-7)

$$\max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{\lambda_1 x^1(T) + \alpha_1 d_1(T) + \epsilon'}{\int_0^T |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)|d\tau + |\alpha_1| - \epsilon'} \geq f(T+\delta)$$

(A3-8)

$$\geq \max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{\lambda_1 x^1(T) + \alpha_1 d_1(T) - \epsilon'}{\int_0^T |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)|d\tau + |\alpha_1| + \epsilon'}$$

for all $0 \leq \delta < \delta'$ since $f(T + \delta)$ is attained at values λ_1, α_1 for which $|\lambda_1|^2 + |\alpha_1|^2 = 1$.

Define

$$\lambda_1 x^1(T) + \alpha_1 d_1(T) = p(\lambda_1,\alpha_1)$$

$$\int_0^T |\lambda_1 h_1(T,\tau) + \alpha_1 h_2(r_1 T,\tau)|d\tau + |\alpha_1| = q(\lambda_1,\alpha_1)$$

$p(\lambda_1,\alpha_1)$ is clearly a continuous function of λ_1 and α_1 and from Proposition 7.1 in Chapter 7 it follows that $q(\lambda_1,\alpha_1)$ is continuous in (λ_1,α_1). Since

$$f(T) = \max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{p(\lambda_1,\alpha_1)}{q(\lambda_1,\alpha_1)}$$

we need only show that if

$$p_1(\lambda_1,\alpha_1) = p(\lambda_1,\alpha_1) + \epsilon' \qquad p_2(\lambda_1,\alpha_1) = p(\lambda_1,\alpha_1) - \epsilon'$$

$$q_1(\lambda_1,\alpha_1) = q(\lambda_1,\alpha_1) - \epsilon' \qquad q_2(\lambda_1,\alpha_1) = q(\lambda_1,\alpha_1) + \epsilon'$$

that corresponding to any $\epsilon > 0$ there exists an $\epsilon' > 0$ such that

$$\max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{p(\lambda_1,\alpha_1)}{q(\lambda_1,\alpha_1)} + \epsilon > \max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{p_1(\lambda_1,\alpha_1)}{q_1(\lambda_1,\alpha_1)} \tag{A3-9}$$

$$\max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{p_2(\lambda_1,\alpha_1)}{q_2(\lambda_1,\alpha_1)} > \max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{p(\lambda_1,\alpha_1)}{q(\lambda_1,\alpha_1)} - \epsilon \tag{A3-10}$$

By assumption $q(\lambda_1,\alpha_1) > 0$ for all λ_1,α_1 and therefore from the compactness of the unit circle q attains a minimum and

$$\min_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} q(\lambda_1,\alpha_1) = C > 0$$

Then since

$$\frac{p_1}{q_1} - \frac{p}{q} = \frac{q(p_1 - p) + p(q - q_1)}{qq_1} < \frac{(p + q)\epsilon'}{C(C - \epsilon')}$$

Again from the compactness of the unit circle $p + q$ attains a maximum on the unit circle and let

$$\max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} p + q = D$$

Choosing ϵ' such that

$$\frac{D\epsilon'}{C(C - \epsilon')} < \epsilon$$

We have

$$\frac{p_1}{q_1} - \frac{p}{q} < \epsilon \quad \text{or} \quad \frac{p_1}{q_1} < \frac{p}{q} + \epsilon$$

and therefore

$$\max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{p_1(\lambda_1,\alpha_1)}{q_1(\lambda_1,\alpha_1)} < \max_{\substack{\lambda_1,\alpha_1 \\ |\lambda_1|^2+|\alpha_1|^2=1}} \frac{p}{q} + \epsilon$$

which is inequality (A3-9). Similar arguments yield inequality (A3-10).

We have now shown that corresponding to any $\epsilon > 0$ there exists a $\delta' > 0$ such that

$$|f(T + \delta) - f(T)| < \epsilon \quad \text{for all} \quad 0 < \delta < \delta'$$

Again similar reasoning allows us to conclude

$$|f(T + \delta) - f(T)| < \epsilon \quad \text{for all} \quad -\delta' < \delta < 0$$

and the theorem is obtained.

The extension of the theorem to prove the continuity of expression (3-19) is straightforward.

AUTOBIOGRAPHICAL STATEMENT

Michael B. Shilman was born in New York City, New York on January 22, 1943. He graduated from the High School of Science (1959). Mr. Shilman received a BEE (1964), an MEE (1967) and a Ph.D. (1970) all from The City College of the City University of New York.

Mr. Shilman was an NDEA fellow during the 1967-1968 academic year and a NASA trainee during the 1968-1969 academic year. He is presently working as a Research Scientist in the Research Department of the Grumman Aerospace Corporation, Bethpage, Long Island, where he has been employed since July 1969.

Mr. Shilman is married to the former Marilyn Leibowitz and they have two children, Perri and Jeffrey.