Errata

p. 26 last line: Delete "s" on "follows".

p. 29 line 4: Insert the word "linear" between "a" and "mapping".

p. 33 line 27: Replace G by ϵ in the bracketed expression.

p. 59 5th line of Sec. 3.8: Insert "of" between "that" and "finding".

p. 63 line 2: Insert "a" between "is" and "case".

p. 66 line 20: $\frac{d\bar{C}(t_1)}{dt}$ should be $\frac{d\bar{C}(t_1)}{dt_1}$

p. 83 Eq. 5.3: Add the condition $p \geq 1$.

p. 165 Footnote: Change "Dranc" to Kranc".

Technical Report No. 164

6137-7-T

MINIMUM PEAK AMPLITUDE CONTROL

by F. M. Waltz

Approved by: 

B. F. Barton

for

COOLEY ELECTRONICS LABORATORY

Department of Electrical Engineering
The University of Michigan
Ann Arbor

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in
The University of Michigan

May 1965

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Cont.)

LIST OF SYMBOLS

General notes: Underlining is used to denote vector quantities. The components of an m-component vector quantity $\underline{v}$ are denoted by $v_1, v_2, \ldots, v_m$. A symbol with a horizontal line above it usually denotes an optimal value of the indicated quantity. A dot above a symbol denotes the time derivative of the indicated quantity.

Symbol	Description	First Reference
$a, \underline{a}$	the given initial state	1
$A, A(t)$	the system matrix	16
$b, \underline{b}, \underline{b}(t_1)$	the desired final output	1
$b_{11}(t), b_{12}(t)$, etc.	elements of the matrix B(t)	42
$B, B(t)$	the input matrix	16
B	a certain Banach space	31
B*	the conjugate of the Banach space B	31
B**	the conjugate of B*	32
$\underline{c}$	a vector in k-dimensional Euclidean space	40
$\bar{\underline{c}}$	an optimal value of $\underline{c}$ for the minimum peak amplitude problem	38
$\underline{c}_p, \underline{c}_n$	optimal value of $\underline{c}$ for the related problem	85
$C, C(\underline{u})$	the cost associated with the control $\underline{u}$	1
$\bar{C}$	the minimum peak amplitude for a given problem	35
C_p, C_n	constants appearing in the solution of the related linear problem	85
C_∞	the limiting value of C_p	90
$\underline{d}$	an arbitrary vector in E^k	114
$D, D(t_1)$	the output matrix	16
E, E_1, E_i	errors (used in the computational algorithms)	113
$E, E(t)$	a certain k x r matrix	21

LIST OF SYMBOLS (Cont.)

Symbol	Description	First Reference
E^n	n-dimensional real Euclidean space	16
$\underline{f}(\underline{x},t)$, $\underline{f}(\underline{x},\underline{u},t)$	nonlinear vector functions characterizing system behavior	101
$F(\underline{c},t_1)$	a certain nonlinear function of $\underline{c}$ and t_1	66
$\underline{g}$	the goal of the control problem	24
$G, G(t)$	the cost weighting matrix	17
$\underline{h}, \underline{h}(\underline{c})$	a point on the surface of the set S_1	113
$H, H(\underline{x},\underline{u},\underline{\psi},t)$	a Hamiltonian-like function	73
I	the identity matrix	23
$J, J(\underline{u})$	a cost functional	102
J_x, J_u	Jacobian matrices	103
k	the number of components in the output	16
K	a constant	30
K(t)	a function of time t in L_1	176
$L, L(\underline{z})$	the mapping taking controls into final outputs	27
$\|L\|$	the norm of the mapping L	28
$L_p^r, L_q^r, L_1^r, L_\infty^r$	Banach spaces	83
m	an arbitrary constant ≥ 1	86
M_i	an open interval on the ith coordinate axis of E^k	33
M	a linear manifold in state space	37
n	the order of the controlled system	4
n	an arbitrary constant ≥ 1	86
N_g, N_z, N_ϵ, etc.	a neighborhood of a point in some topological space	29
N	a positive constant	30
$N(\epsilon)$	a positive integer	174
p	a constant ≥ 1 used in the related problem	80
q	the index conjugate to p	84

Symbol	Description	First Reference
$Q(\underline{g})$	the set of bounded measurable $\underline{z}(t)$ which yield the point $\underline{g}$	25
Q_i	a certain open set in Z; the inverse image of M_i	33
Q	the terminal constraint function	103
r	the number of input or control variables	4
R_1	the closed unit sphere (ball) in B* or Z	27
R	the closed sphere of radius $\overline{C}$ in B* or Z	34
s	a dummy variable of integration	23
$\underline{s}, \underline{s}_1, \underline{s}_2$	points in S_1 or S	29
S_1	the image of R_1 under the mapping L	27
S	the image of R under the mapping L	35
t	time	1
t_0	the initial time	1
t_1	the final time	1
T	as a set: the set of points in the interval $[t_0, t_1]$	1
T	as an algebraic quantity: $t_1 - t_0$	80
T	as a superscript: indicates transposition	16
$T_1, T_1(\underline{c})$	the subset of T on which $\lvert\underline{v}(t)\rvert$ or $\lvert\mathbf{V}^T\underline{c}\rvert \neq 0$	37
T_2	the subset of T on which $\lvert\underline{z}(t)\rvert = \lVert\underline{z}\rVert$	176
$u, \underline{u}, \underline{u}(t)$	the input or control variable or vector	1
$\overline{u}, \overline{\underline{u}}(t)$	an optimal control for the minimum peak amplitude problem	26
$\underline{u}_j(t)$	the elements of a sequence of vector-valued controls	9
$U(\underline{a}, \underline{b}, T)$	the set of admissible controls causing the boundary conditions to be satisfied	17
$\underline{v}$	an arbitrary (row) vector	28
$v_{ij}(t_1, t)$	an element of the matrix $V(t_1, t)$	28
$v^{**}(\underline{z})$, $v_i^{**}(\underline{z})$	bounded linear functionals in B**	32

LIST OF SYMBOLS (Cont.)

Symbol	Description	First Reference
$V, V(t_1, t)$	the k x r matrix $\mathbf{DX}(t_1, t)\ B(t)\ G^{-1}(t)$	24
$\|V\|$	the maximum gain of V over all $\underline{c}$ in E^k and all t in T	89
W	a matrix appearing in the complete controllability condition	30
$W(\underline{c})$	a nonlinear matrix function of $\underline{c}$	126
$x, \underline{x}, \underline{x}(t)$	the state variable or state vector	1
$X, X(t, s)$	the fundamental matrix or state transition matrix	23
$y, \underline{y}, \underline{y}(t_1)$	the output variable or output vector	16
$\underline{z}, \underline{z}(t)$	the vector $G(t)\ \underline{u}(t)$	24
$\bar{\underline{z}}, \bar{\underline{z}}(t)$	a $\underline{z}(t)$ having minimum peak amplitude	25
$z(\underline{v})$	a bounded linear functional defined over all $\underline{v}(t)$ in B	31
$\underline{z}_m, \underline{z}_n, \underline{z}_p$	the optimal control for the related problem of index m, n, or p	84
Z	the Banach space of equivalence classes of bounded measurable controls	27
$\delta, \delta(t_1)$	a small positive number	66
ϵ	set inclusion symbol	1
ϵ	a small positive number	66
η	the minimum eigenvalue of the matrix W	89
θ	a number between zero and one	29
$\underline{\theta}$	a constant vector in E^k	102
θ	the angle between two vectors in E^k	113
ψ_0	a Lagrange multiplier	73
$\underline{\psi}, \underline{\psi}(t)$	a Lagrange multiplier vector	11

LIST OF SYMBOLS (Cont.)

Symbol	Description	First Reference
Special Symbols:		
$\lvert\underline{v}\rvert$, $\lvert\underline{x}\rvert$, etc.	the Euclidean norm of the vectors $\underline{v}$, $\underline{x}$, etc.	1
$\lVert\underline{v}\rVert$, $\lVert\underline{v}\rVert_\infty$	the essential supremum of the Euclidean norm of the vector $\underline{v}(t)$	24
$\lVert\underline{v}\rVert_1$	the L_1-norm of the Euclidean norm of the vector $\underline{v}(t)$	31
$\lVert\underline{v}\rVert_p$	the L_p-norm of the Euclidean norm of the vector $\underline{v}(t)$	80
∂S_1, ∂R, etc.	the boundaries of the sets S_1, R , etc.	114

LIST OF APPENDICES

ABSTRACT

MINIMUM PEAK AMPLITUDE CONTROL

by Frederick Marshall Waltz

The purpose of this study is to develop methods of solution for minimum peak amplitude control problems; i. e., optimal control problems with specified initial and final conditions on the state variables and in which the cost functional is given by

$$\underset{t \in [t_0, t_1]}{\text{supremum}} \left| G(t)\, \underline{u}(t) \right|$$

where t_0 is the initial time, t_1 is the final time, $\underline{u}(t)$ is a measurable vector-valued control variable, G(t) is a matrix of weighting functions, and the symbols $|\ \ |$ denote the Euclidean norm of the enclosed vector. This problem can be identified with various practical problems, such as the problem of determining the input signal to a given electrical device (e. g., a filter) which produces the desired output while requiring minimum peak input power, and the problem of determining the smallest thrust-limited steerable rocket engine capable of performing a specified task.

The problem is initially formulated in terms of linear time-varying systems which can be modelled by differential equations of the form

$$\dot{\underline{x}} = A(t)\underline{x} + B(t)\underline{u}$$

$$\underline{y} = D\underline{x}$$

where $\underline{x}$ is the state vector, $\underline{y}$ is the output vector, the dot indicates differentiation with respect to time, and where A(t), B(t), and D are given matrices satisfying a certain complete controllability assumption. A theorem guaranteeing the existence of an optimal control (similar to the one given by Neustadt) is proven for this class of systems using a functional analysis approach, and a theorem about the form of optimal controls is obtained. It is shown by

means of examples that the optimal control is not unique, in general. A suitably restricted class of systems (with wide engineering applicability) is described for which the optimal control is unique, and takes on the simple form

$$\underline{u}(t) = C\, G^{-1}(t)\, \frac{G^{-T}(t)\, B^{T}(t)\, X^{T}(t_1,t)\, D^{T}\underline{c}}{\left| G^{-T}(t)\, B^{T}(t)\, X^{T}(t_1,t)\, D^{T}\underline{c} \right|} \quad \text{a. e. on } [t_0,t_1]$$

where $X(t_1,t)$ is the state transition matrix for the given system, C is a nonnegative constant, $\underline{c}$ is a constant unit vector, and the superscripted T denotes transposition. The original problem of finding an optimal control function $\underline{u}(t)$ has thus been reduced to the much simpler problem of finding constants C and $\underline{c}$ such that the $\underline{u}(t)$ determined by the above expression causes the desired boundary conditions on the output to be satisfied. The implications of the uniqueness or lack of uniqueness of optimal controls for these problems are discussed in terms of the rotundity and smoothness of the reachable set (the set of all points in state space that can be reached in the given interval, starting from the origin and using controls with costs not exceeding some specified amount). Examples are presented to illustrate the various kinds of behavior that can be obtained.

The relationship of the original class of problems to a class of minimum time problems in which a bound is placed on the quantity $\left| G(t)\, \underline{u}(t) \right|$ is investigated, and it is shown that while all such minimum time problems correspond to minimum peak amplitude problems, the converse is not true.

As a possible approach to nonlinear minimum peak amplitude problems, to which functional analysis techniques and classical variational methods are not readily applicable, a so-called related problem (involving a generalization of the L_p-space norm as a cost functional, and solvable by classical variational techniques) is defined, and it is shown that for linear systems the limiting form of the solution of this problem is an optimal solution of the original problem. Corresponding results for nonlinear systems are not proven, but the technique is applied successfully to two nonlinear examples.

Finally, computational algorithms for the solution of linear minimum peak amplitude problems up to tenth order are presented and discussed, along with additional numerical examples.

CHAPTER I

INTRODUCTION

1.1 Brief Statement of the Problem

In its simplest form, the problem to be considered here can be stated as follows: From the class of bounded measurable vector-valued input or control functions $\underline{u}(t)$ defined on the interval $T = [t_0, t_1]$ which cause a given linear system described at each moment of time by a state[1] vector $\underline{x}(t)$ to transfer from the given initial state $\underline{x}(t_0) = \underline{a}$ at the given initial time t_0 to the desired final state $\underline{x}(t_1) = \underline{b}$ at the given final time t_1, choose one for which the cost functional

$$C(\underline{u}) = \sup_{t \in T} |\underline{u}(t)|$$

is a minimum. Here $|\underline{u}(t)|$ is the Euclidean norm (or, synonymously, the amplitude) of the vector $\underline{u}(t)$ defined at each moment of time as the square root of the sum of the squares of the components of the vector $\underline{u}(t)$ at that moment.

The following generalizations and modifications of this problem are also treated here:

a) The final value of the state is required to lie in a specified linear manifold in state space, rather than at a fixed point $\underline{b}$.

b) The square root of a symmetric positive definite quadratic form in $\underline{u}(t)$ is used in place of the Euclidean norm of $\underline{u}(t)$ in the cost functional.

c) The final time t_1 is left unspecified initially, and must be chosen along with $\underline{u}(t)$ so as to produce the lowest possible cost.

---

[1] A discussion of state variables and the formulation of problems in terms of state variables is given by Tou, Ref. 74.

d) Optimal control laws[2] for certain of the problems considered here are discussed, and the general form required for these control laws is described.

e) A generalization is made to certain nonlinear problems.

Certain other generalizations and variations of the basic problem are also discussed briefly. We shall refer to all the problems noted above as minimum peak amplitude control problems.

This problem formulation encompasses practical problems in various fields. An example from electrical engineering is as follows: The given system is the amplifier-filter shown in Fig. 1.1. It is assumed that the vacuum tube interelectrode capacitances can be ignored and that the tube is operating in an essentially linear region and draws no grid current over the range of inputs involved. The state variables x_1, x_2, x_3 are chosen as the voltages across the capacitors, and the input voltage is denoted by u(t). The input u(t) is to be chosen so as to cause the final state $\underline{x}(t_1) = [x_1(t_1), x_2(t_1), x_3(t_1)]$ to lie in some specified manifold in state space and so that the cost $C(u) = \sup_{t \in T} |u(t)|$ is a minimum. The instantaneous input power to this circuit is simply $\frac{u^2(t)}{R}$, so that minimization of $C(u) = \sup_{t \in T} |u(t)|$ is equivalent in this case to minimization of the peak instantaneous input power to the circuit.[3]

As another example, let the system in question be a rocket vehicle propelled by a single steerable rocket engine, the maximum thrust of which is to be specified. The thrust produced by the engine is regarded as a vector in three-dimensional Euclidean space which can vary in both magnitude and direction as a function of time. The components of this thrust vector along the three coordinate axes of the space are considered as the components of the input vector $\underline{u}(t)$. The Euclidean norm of $\underline{u}(t)$ is thus simply the magnitude of the thrust at a given moment of time. A typical minimum peak amplitude problem in this

---

[2] A control law is a rule or formula which states the value of the control that is to be used at each moment of time as a function of the state at that moment and the desired final state. An optimal control law is one which results in a control which is optimal—i.e., a control which causes the cost to be minimum. This control law (which specifies the operating characteristics of the optimal feedback controller) is to be contrasted with the original formulation of the problem, which requires the determination of $\underline{u}$ as a function of time—a result which is often termed open-loop or programmed control.

[3] This follows from the fact that if we minimize the peak value of a function, we also minimize the peak value of its square and of any constant times its square.

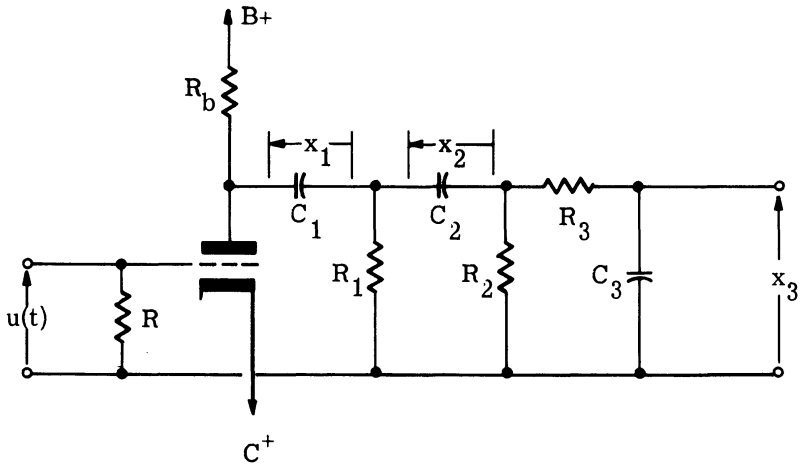

Fig. 1.1. An example in which the cost functional is related to peak instantaneous input power.

case is the problem of determining the smallest thrust-limited engine[4] that is capable of performing a specified task (e.g., putting a certain satellite into a certain orbit), and of determining the steering pattern to be used.

From these two examples, it can be seen that the techniques to be developed here for the solution of such problems are essentially system synthesis techniques; that is, they give information about the smallest signal generator or the smallest rocket engine that will do the required job, as an aid in system design. If one already has a specific signal generator or a specified rocket engine at hand, the pertinent problem is then not the synthesis problem, "What is the minimum amplitude signal that will do the job?" but the analysis problem, "Will this device do the job, and if so, how well?", with performance being evaluated on some other appropriate basis, such as with a time or a fuel consumption criterion. Since numerous techniques for the solution of such analysis problems are available and widely known (Refs. 1 through 20), these problems will not be discussed in detail here except as they relate to the problem at hand.

1.2 Review of the Literature

1.2.1 Historical Foundations. The problem described above is known as an

---

[4] Examples of engines that are primarily thrust-limited, rather than, for example, energy-limited, are the engines which do not carry their own energy source, such as a "solar sail" or a light-powered ion engine.

optimal control problem.[5] Optimal control problems are the subject of an extensive and rapidly expanding body of literature. Interest of the engineering and scientific community in such problems began in earnest within the last fifteen years (Refs. 1,2,3,4). The mathematical foundation for much of the work being done today is found in the calculus of variations, developed long before this, beginning with the study of first-order necessary conditions for extremality (i. e., optimality) by Euler and Lagrange in the eighteenth century (Ref. 5, p. 54). Second-order conditions, which allow one to classify extremals as maxima, minima, or inflection points, (analogous to the classification of stationary points of ordinary functions by consideration of second derivatives) were then developed (Ref. 6, p. 23). With the publication of Bolza's work on variational problems with differential side conditions in 1913 (Ref. 7), all the needed elements were present for solution of optimal control problems which involve no restrictions on the state and control variables (except those implicit in the differential equations and boundary conditions). A discussion of these results and their relationship to optimal control problems (in their "modern" formulation) is given in Appendix A.

Mathematicians (Bliss, Refs. 6,8, and others) continued to refine and extend these results, and Valentine (Ref. 9, published in 1937) introduced a simple device which in effect allowed the n-component state vector $\underline{x}(t)$ and the r-component control vector $\underline{u}(t)$ to lie at each moment of time in closed sets in (n+r)-dimensional space, rather than <u>open</u> sets, as in previous works. This made possible a generalization of great potential application to engineering problems, because of the fact that in many practical problems $\underline{u}(t)$ or $\underline{x}(t)$ or both must be confined to closed regions in their corresponding Euclidean spaces. However, this method of Valentine received virtually no attention outside mathematics circles at the time.

Interest in these results became widespread among engineers only after current engineering problems, involving such things as minimum-response-time servo systems and thrust control of intercontinental or interplanetary rockets, stimulated the development of two new approaches to problems of this type: Pontryagin's maximum principle (Refs. 10 through 15) and Bellman's dynamic programming (Refs. 17,18).

---

[5] An optimal control problem, in the sense used here, is a problem in which a controlling variable must be chosen so as to produce some desired effect and so as to optimize (i. e., maximize or minimize) some performance or loss functional.

1.2.2 Pontryagin's Maximum Principle. After its original formulation, applicable to time-optimal problems (Refs. 10, 11), the maximum principle was extended to a formulation suitable for use with much more general "cost" functions (Refs. 12, 13), then to a version applicable to problems with bounded state variables (Refs. 13, 14), and, most recently, to an integral formulation (Ref. 15) that applies to problems in which the admissible controls are restricted to closed sets in some function space (rather than being restricted, moment by moment, to some set in Euclidean space).

Pontryagin and his associates have relied for the most part on variational methods to prove their results, using a particular kind of constrained or "one-sided" variation in place of the "two-way" variations of the classical calculus of variations.

Berkovitz and others have shown (Refs. 19, 20, 21) that many of these results can be obtained (under conditions on the system and on the class of admissible controls which are more restrictive than those assumed by Pontryagin and Gamkrelidze) from the classical calculus of variations using Valentine's method. (See Section A. 3. 2 of Appendix A). In fact, Berkovitz's results go beyond those of Gamkrelidze, in that they also apply to problems with joint constraints on the values of $\underline{x}(t)$ and $\underline{u}(t)$ rather than separate constraint sets for $\underline{x}(t)$ and $\underline{u}(t)$. Here again, however, Berkovitz's conditions are more restrictive than those of Gamkrelidze.

The major contributions of Pontryagin and his associates to the optimal control field fall into three categories:

(a) Mathematical results: The new theoretical contributions (i. e., those not contained or implied in previous works) are relatively few, the major contribution being not the maximum principle itself, but the proof of the maximum principle under less restrictive conditions on the system differential equations than those imposed by Bliss and other earlier workers (see Ref. 22, p. 4, and Section A. 3. 1 of Appendix A). Another contribution of considerable importance lies in the Pontryagin formulation of optimal control problems. By defining the control variables as separate entities, distinct from the state variables and subject to different types of constraints (a distinction that is not made in the usual calculus of variations formulation), Pontryagin has been able to simultaneously reduce the order of the

system of differential equations that must be considered, provide a concise notation for expressing the necessary conditions for optimality, and lay the foundation for analogies with Hamiltonian mechanics.

(b) Optimal control philosophy: Classical calculus of variations tends to center its attention on the cost functional--the quantity to be minimized--relegating everything else to the status of "side conditions." Pontryagin's approach, in common with other modern treatments, puts more emphasis on the physical system (represented by its differential equations). For many practical engineering problems, this latter orientation seems more appropriate, since the system is an objective thing (presumably known or knowable), while the "cost" is much more subjective and open to choice. This orientation leads plausibly to the problem of determining the optimal control for the same system with various cost functionals, to the so-called inverse problem of optimal control, and other practical considerations. (These points are discussed in more detail below.) Furthermore, the emphasis that the Pontryagin approach places on the control variables is appropriate to engineering problems, where the specification of these variables is often the primary goal.

(c) "Popularization": There can be no doubt about the fact that the work of the Russian school has caused much attention throughout the world to be given to the whole field of optimal control and the calculus of variations, and has stimulated considerable effort in these areas. This is in itself a contribution of some importance.

The maximum principle is not without drawbacks, of course. Some of these are discussed below.

1.2.3 Dynamic Programming. Dynamic programming, developed by Bellman (Refs. 17, 18) at about the same time that the Russian school was developing the maximum principle, serves to eliminate four disadvantages inherent in conventional variational methods: a) the two (or more) point boundary value problems that result from the application of such methods, b) the need for continuity (with respect to the state variables)

in the functions used to describe system behavior and in their partial derivatives, c) the indirectness of these methods (i. e. , the fact that they work with auxiliary variables which are even further removed from "physical reality" than are the equations which presumably describe system behavior), and d) the inapplicability of such methods to stochastic problems. These points are discussed in some detail in Appendix B, which also gives a brief exposition of the method of dynamic programming as a computational technique. Discussion here will be limited to a statement of the fundamental idea behind the method:

> For cost functionals in which the total cost (the cost over the whole interval) is the sum of the costs on a partitioning of the complete trajectory into subarcs, each of the subarcs is itself an optimal trajectory connecting the end points of the subarc. Therefore, if the total interval is from t_0 to t_1, and if t' is some intermediate time, it is a property of optimal trajectories that the control action from t' to t_1 must be optimal for that subarc, regardless of what happended before t'. This is a statement of the "principle of optimality. "

This principle leads to a computational technique which, in its discrete-time form, starts at the final time and computes the optimal trajectory in the vicinity of the final point and then works backward toward the initial time, step-by-step (i. e. , subarc by subarc), tabulating the optimal trajectories at each step. As pointed out in Appendix B, this process yields a family of optimal trajectories with various initial and/or final conditions, from which is selected the trajectory with the desired boundary conditions. While dynamic programming, as described above, is applied mainly in the numerical solution of problems, the same approach can be used to generate a partial differential equation, the solution of which gives the optimal control for continuous-time optimization problems. One would expect that this technique, when applied to a problem to which calculus of variations is applicable, would turn out to be equivalent to the calculus of variations. This is indeed the case for suitably restricted classes of problems, and many authors (Bellman, Ref. 28, Kopp, Ref. 23, Dreyfus, Ref. 24, Pontryagin, et al. , Ref. 13) have used this continuous-time version of dynamic programming to derive (with more or less rigor and with various restrictions on the class of problems considered) the Euler equations, the maximum principle, and/or other results of the calculus of variations.

This connection was further explored by Kalman (Refs. 25,26), who has shown (Ref. 27) that the dynamic programming approach is analogous to the Hamilton-Caratheodory approach to the calculus of variations.

As noted above, one of the major advantages of dynamic programming is that it can be used in problems to which variational methods are not applicable; for example, problems in which the derivatives required by the variational methods do not exist. However, as pointed out by Pontryagin, Boltyanskii, et al. [Ref. 13, pp. 69-73], dynamic programming does not have the rigorous mathematical foundation possessed by the variational methods, so that in some cases in which dynamic programming may be used (as a heuristic tool) there is no assurance that the results obtained are indeed optimal. These same authors go on to discuss a weakness in the continuous-time version of dynamic programming, which is that a certain function which appears in the derivation is required to be differentiable. They show examples of a very simple and basic sort (solvable by means of the maximum principle) for which this requirement is not met.

The discrete-time computational version also suffers from a weakness that is not apparent from the discussion above and in Appendix B, which can be summed up as a problem of system modeling: In this computational technique, the system is characterized by difference equations. If the original system characterization is in terms of differential equations, it can happen that important aspects of the behavior of the system are lost in the conversion to difference equations.[6] As a result of this, dynamic programming might in some cases produce a control which was far from optimal, without giving any indication that such an event had taken place.

---

[6] The following example, while admittedly somewhat strained in that it involves a differential equation which does not have a unique solution for certain initial conditions, nonetheless shows that drastic things can happen in the conversion from differential equations to difference equations: Note that the differential equation $\dot{x} = 2|x|^{\frac{1}{2}} + u(t)$, with initial conditions $x(0) = 0$ has two unforced solutions (i. e., solutions for which the forcing function $u(t)$ is equal to zero): $x(t) = 0$ and $x(t) = t^2$. In contrast, the difference equation

$$[x(t_i + \Delta t) - x(t_i)]/\Delta t = 2\left|x(t_i)\right|^{\frac{1}{2}} + u(t_i)] \quad \text{or}$$

$$x(t_i + \Delta t) = x(t_i) + \Delta t[2\left|x(t_i)\right|^{\frac{1}{2}} + u(t_i)]$$

has only the single unforced solution $x(t) = 0$ for the initial condition $x(0) = 0$. Suppose that we were given the problem of choosing $u(t)$ so as to force this system from $x(0) = 0$ to $x(1) = 1$ while minimizing the integral of $u^2(t)$. The differential equation formulation indicates that this can be accomplished with zero cost [i. e., $u(t) = 0$ throughout] but the difference equation formulation overlooks this possibility.

1.2.4 Direct Methods. Classical variational calculus and its modern extensions as well involve the solving of a set of auxiliary equations (which go under various names, depending on the formulation: Euler-Lagrange equations, Hamilton equations, adjoint equations, Lagrange-multiplier equations, etc.). There are also other optimization methods, lumped under the title of "direct methods," which do not involve the solution of such auxiliary systems, but instead determine (or approximate) the optimal control by working directly with the functional to be optimized.

One important class of direct methods, of which the Ritz method and the method of finite differences are representative, involves three basic steps (see Akheizer, Ref. 72, or Gelfand and Fomin, Ref. 73):

a) the construction of a minimizing sequence; that is, a sequence of controls $[\underline{u}_j(t)]$ having the following properties:
   i) each control $\underline{u}_j(t)$ is in some given metric space U,
   ii) each control $\underline{u}_j(t)$ causes the system to satisfy the desired boundary conditions,
   iii) the "cost" associated with the controls $\underline{u}_j(t)$ approaches as a limit the infimum of the "cost" over all controls satisfying i) and ii) as j approaches infinity;

b) a proof that the elements of the sequence converge (in the metric defined for the space U) to a function (called the limiting function) which satisfies i) and ii);

c) a proof that the limiting function has as its cost the infimum mentioned in step a).

Fortunately, theorems are available which eliminate the need for carrying out each of these steps in every case. For example, step c) can be eliminated if the cost functional can be shown to be lower semicontinuous (in the metric defined for the space U) at the limiting function (see Gelfand and Fomin, Ref. 73, pp. 8 and 194). And, as another example, under conditions which are from a practical standpoint neither extremely restrictive nor difficult to test for the satisfaction of, the Ritz method produces a sequence which is guaranteed to satisfy all the conditions of steps a), b), and c). The reader is referred to the stated references, and the references given therein, for detailed expositions of these and similar methods. We limit ourselves here to the following statements: Both the Ritz

method and the method of finite differences involve the construction of a minimizing sequence by the minimization of a succession of functions (not functionals) of m variables, for progressively larger values of m. From a computational standpoint, this is a simpler problem than the original problem, involving the minimization of a functional, and as a result these methods have been used successfully in a number of practical problems. In fact, depending on the degree of approximation desired, the process can often be terminated after only a few steps. On the other hand, convergence can in some cases by very slow, so that these methods by no means eliminate the need for the indirect variational methods discussed above.

The approach used by Faulkner (Ref. 40), Bryson (Refs. 41 and 42), and Kelley (Refs. 43 and 44) might also be considered a direct method [7] in that the adjoint equations are not explicitly solved in these methods. Instead, a successive-approximation approach is used to approximate a control which causes the adjoint equations (Lagrange multiplier equations) and the maximum principle to be satisfied. These methods are primarily computational implementations of classical or modern variational methods, and are therefore classified here with the computational advances discussed in the next section.

1.2.5 Current Work. Since the development of dynamic programming and the maximum principle, work in the field of optimal control systems has tended to fall into three categories:

a) Mathematical reformulations, extensions, and reinterpretations: the viewing of existing results in new ways; the proving of existing theorems by different or more elegant methods.

b) Computational advances: improvement of existing computational methods and the development of new methods.

c) Restatements and revision of basic goals: a recognition of the unrealistic nature of many optimal control problems, as posed; efforts to use optimal control theory in ways more meaningful to practical problems.

These three categories are discussed in more detail below.

---

[7] Faulkner refers to his approach as a direct method.

1.2.5.1 Mathematical Reformulations, Extensions, and Reinterpretations.

Considerable effort has been devoted to exploring the interrelationships between the calculus of variations, the maximum principle, and dynamic programming (as mentioned in preceding sections), resulting in a widespread appreciation of the common foundation underlying the whole of optimal control theory.

Geometrical interpretations of the optimal control problem (Roxin, Ref. 28, Flügge-Lotz and Halkin, Ref. 29) have provided a great deal of insight into the nature of optimal trajectories and into the techniques for solution of optimal control problems. The key idea here is the concept of the reachable set: the set of all points in state space that can be attained in a given time starting from a given initial point and using controls from the admissible set. With this interpretation, the Lagrange multiplier vector $\underline{\psi}(t)$ of the calculus of variations and the maximum principle becomes simply the outward normal to the reachable set, the principle of optimality is a consequence of the fact that an optimal trajectory must stay on the boundary of this set, and the maximum principle is simply a statement of the fact that the Lagrange multiplier vector must have a nonpositive inner product with the optimal state velocity vector, which is tangent to the surface of the reachable set.

Krassovski (Refs. 30,78), Kulikowski (Refs. 31-34), Kirillova (Ref. 35), Kranc and Sarachik (Refs. 36,37), Neustadt (Ref. 77), and others have used certain results from functional analysis (Hölder's inequality, principally) to solve linear time-optimal problems in which the admissible controls are restricted to closed sets in some function space, rather than having their values restricted, moment-by-moment, to some set (however complicated) in Euclidean space. (This approach, because it is closely related to the subject matter of the present work, is presented in some detail in Appendix C.) The contribution of these papers lies not so much in the problems that they are able to solve (which can be solved in other ways) or in increased convenience of obtaining numerical solutions (which is lacking), but in the simplicity and elegance of the mathematical proofs of optimality that they involve, and in their application of functional analysis concepts to a field which is after all the study of the extremization of functionals. Formulations involving more general function spaces than the Banach spaces used in these papers can be expected to follow.

In a treatment which involves elements of functional analysis in its formulation and proofs, and which utilizes geometric ideas throughout (including a generalized reachable set and many other ideas from the geometric interpretations mentioned above), Gamkrelidze (Ref. 15) has achieved a synthesis of many concepts and results which were previously only loosely related.

The question of existence of optimal controls is also receiving some study (Markus and Lee, Ref. 38), but much remains to be done in this area. It is obvious that an optimal control cannot exist unless there exists at least one control which causes the desired conditions to be satisfied. Therefore, the work on the controllability of systems (Refs. 25,60,69) is also pertinent to this problem. The theory for linear systems is virtually complete, but little has been done on the controllability of nonlinear systems.

Sufficient conditions for optimal control have also received some attention (Refs. 13,76). Here again, the results for linear systems are much more complete than those for nonlinear systems.

1.2.5.2 Computational Advances. The advent of large-scale digital computers has made it possible to solve many optimization problems that were once too long or complex for solution by existing methods. Even now, however, problems of rather low dimension--10 or so--may sometimes outstrip the capabilities of the largest and fastest computers available. Thus, there is frequently a need for computational techniques which reduce computation time or storage requirements, relative to straightforward implementations of the basic mathematical methods.

In the case of dynamic programming, which typically requires only a modest amount of arithmetic capacity but large amounts of storage, this has led to methods for reducing storage requirements at the expense of extended computing time. This and other computer programming techniques have made possible the solution of engineering problems of considerable size by means of dynamic programming.

The two-point boundary value problems which result from the application of variational methods must be solved iteratively, in general. To this end, interest in the method of gradients, or method of steepest descent, first applied to variational problems in 1908 by Hadamard (Ref. 39), has revived -recent work on the subject having been done

by Bryson, et al. (Refs. 41 and 42) and Kelley (Refs. 43, 44). Graham (Ref. 45) has generalized this technique to apply to "multiple-arc" problems--problems in which a finite number of discontinuities exists in the time derivatives of the state variables.

1.2.5.3 Restatements and Revision of Basic Goals. As can be seen from the previous discussion, much of the work that has been done in optimal control theory is highly theoretical and mathematical in nature. In applying these results to practical engineering problems, it is often necessary to make many simplifying assumptions (or just plain guesses). The question then arises, "What good are techniques which give very precise and often intricate solutions to mathematical formulations which are only rough approximations of the real problem/?"

One of the principal areas of difficulty involves the cost function itself. Corresponding to each cost function there is a specific optimal control (or set of optimal controls), but the choice of cost function is often a highly subjective matter, and it may be difficult or impossible to incorporate all the important factors into a single mathematical function in a meaningful way. (A similar statement could be made about the constraints on the state and control variables, also.)

One attempt to overcome this is the use of a vector-valued cost function (Zadeh, Ref. 46), which measures each control by a number of different criteria. Zadeh argues that, while such an approach may not produce a complete ordering of the controls (i.e., there may be no control which is as good as or better than every other control with respect to all of the criteria), it may eliminate a good many controls from consideration (as being not optimal on any count). It may then be possible to define a single cost functional which is meaningful over the class of controls which survive the first test, and in any case the system designer is better able to see how the various factors affect the total cost--a relationship that is often obscured when a single cost functional must be specified a priori. Another approach is the study of what is usually called the inverse problem of optimal control (Refs. 47, 48, for example): Determine the class of problems for which a given control or control law or type of control law (e.g., a linear control law) is optimal. If this class is large enough to encompass the uncertainties in the mathematical model and the

different cost functions that can meaningfully be applied, then one is justified in using that control or control law.

Many systems of engineering importance include significant stochastic variables, measurement errors, or unavoidable computation errors, which render the typical calculus of variations solution to the completely deterministic control problem inappropriate or impractical. Therefore, work is being done on optimization and prediction techniques applicable to stochastic problems (Refs. 49, 50, 51, to name a few). Also, the determination of control laws, which specify the control that should be used as a function of some or all of the state variables, is being emphasized (Refs. 52, 53).

In large or complex systems (complete chemical factories or public utilities systems, for example), another factor becomes important: While it may be possible to find an over-all optimum control policy for the whole system, such a policy might become so complicated and interrelated as to require excessive "hardware" or computer capacity for implementation; it may have an adverse effect on stability under parameter changes; etc. A more practical approach (and hopefully yielding results not too far from the over-all optimum) would be to optimize subsystems, which presumably require simpler control systems, and then have one or more levels of supervisory control systems to set goals for the subsystems. Such multilevel control is being investigated in many of its aspects (Refs. 54, 55, 56).

1.3 The Literature and the Minimum Peak Amplitude Problem

With the exception of the functional analysis approach, the optimization techniques discussed thus far involve minimization of some functional expressible as an integral (or a summation). In the problems solvable by these techniques, the total cost is, in general, affected by every segment of the trajectory, however small. In the minimum peak amplitude problem, on the other hand, the total cost is not expressible as an integral, and may be determined by the magnitudes of the control variables on a very small subinterval of the total interval, with the controls at all other times having no (direct) effect on the cost. Thus, of the available techniques, only the functional analysis approach is directly applicable to the minimum peak amplitude problem.

In the present work, functional analysis methods are used to derive the basic results. These results are then shown to be closely related to those of a certain time optimal problem studied by Krasovskii (Ref. 78). The significant differences between the two problems are discussed. Finally, a limiting process used by Kirillova (Ref. 35) is generalized to the problem at hand, which leads to a technique that is applicable to certain nonlinear systems.

CHAPTER II

THE LINEAR PROBLEM

2.1 Detailed Statement of the Problem

The problem to be stated in this section already includes some of the generalizations noted in Section 1.1. Other generalizations and variations will be introduced at appropriate points in the discussion. The reasons for and the implications of the various assumptions and restrictions made here are discussed in the section immediately following this statement of the problem:

Consider a time-varying deterministic linear dynamic system with input or control vector[8] $\underline{u} = \underline{u}(t) = [u_1(t), \ldots, u_r(t)]^T$, state vector $\underline{x} = \underline{x}(t) = [x_1(t), \ldots, x_n(t)]^T$, and output vector $\underline{y} = \underline{y}(t) = [y_1(t), \ldots, y_k(t)]^T$ related by

$$\dot{\underline{x}} = A(t)\,\underline{x} + B(t)\,\underline{u} \tag{2.1}$$

$$\underline{y}(t) = D\,\underline{x}(t) \tag{2.2}$$

where $\dot{\underline{x}}$ is the time derivative of the state $\underline{x}$; A(t) and B(t) are given n x n and n x r matrices, respectively; the elements of A(t) and B(t) are bounded measurable functions of time defined on the interval $T = [t_0, t_1]$; and D is a given constant k x n matrix ($1 \leq k \leq n$) of rank k. The matrices A(t) and B(t) are further assumed to be such that the output $\underline{y}$ is completely controllable on the interval T; that is, it is assumed that corresponding to every point $\underline{a}$ in n-dimensional Euclidean space E^n and every point $\underline{b}$ in k-dimensional Euclidean space E^k, there exists at least one bounded measurable control defined on T which causes the system to

---

[8] To conserve space in the body of the text, column vectors such as $\underline{u}$ and $\underline{x}$ will often be written out as transposed row vectors, the superscripted T indicating the transposition operation.

transfer from the given initial state $\underline{x}(t_0) = \underline{a}$ to any state for which the output y takes on the the value $\underline{b}$ at specified[9] time t_1. This complete controllability assumption, for which a necessary and sufficient condition is given in Appendix D, is different from the one used by Kalman and others (Refs. 25, 60, 69). This point is further discussed below.

In terms of these definitions and assumptions, the minimum peak amplitude problem to be considered initially can be stated as follows:

From the set $U(\underline{a}, \underline{b}, T)$ of bounded measurable controls $\underline{u}(t)$ which cause the system characterized by Eq. 2.1 to transfer from the given initial state $\underline{x}(t_0) = \underline{a}$ at given initial time t_0 to any state for which $\underline{y}(t_1) = \underline{b}$ at given final time t_1, choose as an optimal control $\bar{\underline{u}}$ any one for which the cost functional

$$C(\bar{\underline{u}}) = \sup_{t \in T} \left| G(t)\, \bar{\underline{u}}(t) \right| \tag{2.3}$$

is a minimum with respect to all $\underline{u} \in U(\underline{a}, \underline{b}, T)$. Here G(t) is an r x r matrix of bounded measurable functions defined on T, the inverse of which exists for all t in T and also consists of bounded measurable functions defined on T. As in Section 1.1, the symbols $|\ |$ denote the Euclidean norm of the vector enclosed.

2.2 Discussion of Assumptions and Restrictions

The assumptions and restrictions made in the above problem statement are discussed and explained below:

The restriction that A(t) and B(t) consist of bounded measurable functions is sufficiently weak to allow a large number of physical systems to be modeled by Eq. 2.1, and yet strong enough to guarantee that the solutions of this equation can be written in a particularly simple form[10]--a form which plays an important role in the results obtained here. This restriction could be relaxed in minor ways (e.g., by requiring only essential boundedness instead of boundedness) if desired, but such relaxed conditions are of little interest in most practical problems.

---

[9] The original problem was stated in terms of achieving a certain goal at a specified final time, and therefore the complete controllability assumption has been formulated in the same terms. The implications of this assumption are discussed in the next section.

[10] This is discussed in detail in Section 3.1.

The matrix G(t) allows different weighting factors to be applied to the various terms which appear in the cost functional, and allows these factors to be varied with time. In the rocket problem mentioned in Section 1.1, the flexibility provided by the inclusion of this matrix was not needed, since the thrust was simply the Euclidean length (norm) of the vector $\underline{u}$. In such a problem, G(t) is set equal to the identity matrix. By including G(t) in the initial formulation, we can apply our results both to this problem and to other problems with more general cost functionals--for example, to certain electrical networks in which it is desired to minimize the maximum instantaneous input power and in which the components of $\underline{u}$ represent the current and/or voltage inputs to the circuit. The components of G(t) are then related to the input resistances and/or conductances of the circuit, in such a way that the expression $\underline{u}^T(t)\, G^T(t)\, G(t)\, \underline{u}(t)$ (the square root of which is the quantity involved in the cost functional) becomes the total input power at time t.

The matrix G(t) is required to be invertible in order to avoid the possibility of degenerate problems.[11] Such problems are excluded from consideration here because of the mathematical difficulties that they present. This is not to say that degenerate problems are not valid problems--they may arise naturally when one or more of the control variables "costs nothing," i.e., is available in such quantities that it may be applied in unlimited amounts without incurring any cost. Examples might be the use of atmospheric air and sunlight as control variables in a chemical process: compared to that of the other input variables, the cost of these might be so small that it could be ignored.

The cost functional has been defined in terms of the supremum, but for most practical purposes we could use the essential supremum just as well. For reasons of convenience, the results derived here are obtained using a cost functional defined

---

[11] A degenerate problem is one in which the boundary conditions can be satisfied by a control which is different from zero on a set of measure greater than zero and which nonetheless results in zero cost. If G(t) has a nontrivial null space, then it could happen that the set $U(\underline{a}, \underline{b}, T)$ contained a nonzero control $\underline{u}(t)$ which lay entirely in this null space at each moment of time, resulting in $G(t)\, \underline{u}(t)$ being identically zero. This would imply in turn that the cost $C(\underline{u})$ was zero.

in terms of the essential supremum. The optimal control for such a functional turns out to be identical to that for the functional originally specified, except for possible differences on sets of measure zero. We shall consider all controls which differ from each other only on a set of measure zero as being equivalent,[12] and therefore in this problem there is no significant difference between these two cost functionals.

The specification of the final conditions on the system in terms of a final condition $\underline{b}$ on the output vector $\underline{y}$ is actually a statement to the effect that the final value of the state $\underline{x}(t_1)$ must lie in some (n-k)-dimensional linear manifold M in n-dimensional state space E^n, where M is defined as the set of all points $\underline{x}(t_1) \in E^n$ for which $D\,\underline{x}(t_1) = \underline{b}$. The formulation in terms of the output $\underline{y}$ lends itself to application in a number of practical problems. For example, if the system in question is a linear bandpass filter, some of the state variables (those representing internal voltages or currents, for example) may be of no direct concern in a given application. The matrix D is then chosen so that the "output" consists of only those state variables which are of interest (or some desired linear combination of them). If the problem is one of steering a projectile to intercept a moving target in three-dimensional space, only the position of the projectile at the moment of interception (and not the velocity, angular velocity, etc.) might be important--a fact which could be taken into account by an appropriate choice of D. In such a case, the final condition $\underline{b}$ would represent the position that the projectile must occupy if the interception is to take place at the desired final time t_1.

With this idea in mind, we can see that the requirement that the output be completely controllable on the interval T is less restrictive than the complete controllability condition often imposed (Refs. 77,78), in that those state variables which do not enter into the output vector $\underline{y}$ (through the matrix D) need not be controllable at all. We require controllability of only those parts of the system which concern us. However, in another sense this type of complete controllability is more restrictive than that defined by Kalman (Ref. 26)

---

[12] As we shall see in the next chapter, the control functions enter into the expression for the output only in the integrand of a Lebesgue integral. Thus, control functions which differ only on a set of measure zero all yield the same output. See Natanson, Ref. 79, p. 137.

in that it requires that every possible output be attainable at precisely the specified final time t_1, (rather than merely at some time greater than t_0). Kalman's definition of controllability is appropriate to minimum-time problems and other problems in which the final time is unspecified, but is of little use in a problem with fixed final time. To illustrate this point further, suppose we had a system which was not completely controllable on the interval T, as defined above, but was completely controllable in Kalman's sense; i. e., for any given initial state and final state, there exists some time t_1' at which the final state can be achieved. If the smallest possible value of t_1' is greater than the specified t_1, then obviously the requirements of the original problem can not be met, since we can not achieve the desired goal at the specified time t_1. On the other hand, if it is known that a $t_1' < t_1$ exists, it does not necessarily follow that the goal can be achieved at t_1 (and achieving the goal at t_1 may be the essence of the problem, as in the swinging of a baseball bat across the plate at the precise moment when the baseball thrown by the pitcher is crossing the plate and in the proper position to be hit past the second baseman on the hit-and-run play). To see that complete controllability at t_1' does not necessarily imply complete controllability at $t_1 > t_1'$, consider the following (admittedly rather artificial) example: Let the system in question have two parts, a completely controllable part characterized by the state variables $x_1(t), \ldots, x_m(t)$ and an uncontrollable[13] part characterized by the state variables $x_{m+1}(t), \ldots, x_n(t)$. Let the output matrix D be a continuous matrix function of time (as is done in Section 3.8 below) such that on the interval $[t_0, t_1']$ the output $\underline{y}(t)$ involves only the controllable state variables $x_1(t), \ldots, x_m(t)$ and such that on the interval $[t_1'', \infty)$ the output $\underline{y}(t)$ involves only the uncontrollable state variables $x_{m+1}(t), \ldots, x_n(t)$. (Here t_1'' is larger than t_1' to allow for a continuous transition from one form of D to the other.) This system is clearly not controllable for any $t_1 \geq t_1''$, but is completely controllable for all t_1 in $(t_0, t_1']$.

The complete controllability condition stated in Section 2.1 is essential to all the results obtained here for the linear problem in that it plays a key role in the existence theorem of Section 3.2, which is in turn needed in all other results. Problems which do not

---

[13] See Kalman, Ref. 26.

satisfy this condition may well be valid and important problems, but they generally cannot be solved by the methods presented here.

Two possible variations of the problem, as posed, will now be discussed: A more general system characterization than that represented by Eqs. 2.1 and 2.2 is one in which the control variables influence the output not only through their influence on the state variables, but also directly, i.e., Eq. 2.2 is replaced by

$$\underline{y}(t) = D\,\underline{x}(t) + E(t)\,\underline{u}(t) \tag{2.4}$$

where E(t) is a k x r matrix function of time defined on the appropriate interval which is bounded in norm[14] at each time t in this interval. Such a situation can arise even in very simple cases. An example is the idealized system[15] shown in Fig. 2.1. For this system, the equations corresponding to Eqs. 2.1 and 2.2 are

$$\dot{x} = -\frac{1}{RC}\,x + \frac{1}{RC}\,u$$

$$y = -x + u$$

The equation for the output y is obviously of the form of Eq. 2.4 and not that of Eq. 2.2

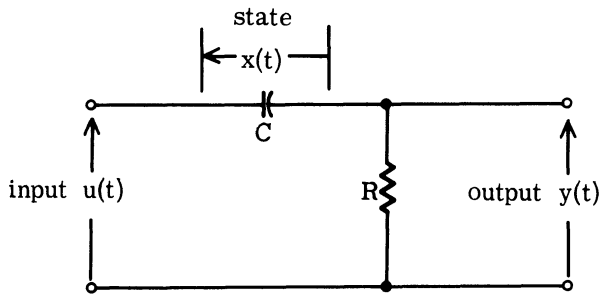

Fig. 2.1. An example in which the input directly influences the output.

---

[14] The norm $\|E(t)\|$ of the matrix E(t) is here defined at each moment of time as $\sup_{|\underline{u}|=1} |E(t)\,\underline{u}|$ where as always the symbols $|\ |$ denote the Euclidean norm of the enclosed vector.

[15] This system is idealized in the sense that lead inductances have been neglected. Inclusion of a lead inductance in series with the input capacitor would eliminate the effect described here. Nonetheless, this situation can be closely approximated in practice.

Another variation that has some practical applicability involves problems in which both the initial state and the final state are specified as lying in linear manifolds in state space.

Both of these variations include interesting and useful problems, but they will not be treated in detail here because of the increased complexity that they introduce into the derivations and proofs. However, at appropriate points in the discussion, comments will be made to show how the techniques used here can be extended to apply to these variations.

CHAPTER III

SOLUTION OF THE LINEAR PROBLEM

3.1 Preliminary Definitions and Results

The basic results of this chapter are presented as a series of lemmas and theorems. Before proceeding with these, we make certain definitions and note certain results from the theory of linear differential equations which will be needed in the remainder of this work:

For A(t), B(t), and $\underline{u}(t)$ consisting of bounded measurable functions defined for $t \in T$, the solution $\underline{x}(t)$ of the vector-matrix differential equation $\dot{\underline{x}} = A(t)\,\underline{x} + B(t)\,\underline{u}$ with boundary condition $\underline{x}(t_0) = \underline{a}$ can be written as

$$\underline{x}(t) = X(t,t_0)\underline{a} + \int_{t_0}^{t} \mathbf{X}(t,s)\, B(s)\, \underline{u}(s)\, ds \qquad (3.1)$$

where $\mathbf{X}(t,s)$ is an absolutely continuous invertible n x n matrix, the solution of the matrix differential equation

$$\frac{d}{dt} X(t,s) = A(t)\, X(t,s); \qquad t,s \in T$$

with boundary condition $X(t,t) = I$ (the identity matrix) for all $t \in T$, (see, for example, Coddington and Levinson, Ref. 57, Chapters I and II). Since the emphasis of this work is on optimization problems, rather than on the solution of systems of linear differential equations, $\mathbf{X}(t,s)$ will be assumed to be known in most of the following.[16]

---

[16] If the matrix A is constant, independent of t, this assumption presents no great difficulty, since straightforward methods for obtaining $X(t,s)$ are available in such cases (Refs. 57, 75). We note, however, that for arbitrary A(t), the determination of $\mathbf{X}(t,s)$ in analytical form may be difficult or impossible, so that for such problems the assumption that $X(t,s)$ is known is not to be taken lightly. In such cases, numerical solution is often the only practical approach, and even this can be very complex.

Applying the above expression for $\underline{x}(t)$ to the problem at hand (see Section 2.1), and noting that $\underline{y}(t_1) = D\underline{x}(t_1)$, we see that the desired final condition $\underline{y}(t_1) = \underline{b}$ will be satisfied if and only if $\underline{u}(t)$ is such that the equation

$$\underline{y}(t_1) = \underline{b} = D\left[X(t_1, t_0)\underline{a} + \int_T X(t_1, s)\, B(s)\, \underline{u}(s)\, ds\right] \tag{3.2}$$

is satisfied. (Here and in the sequel the integral sign with subscripted T implies integration over the interval $T = [t_0, t_1]$).

This problem will now be restated in a form more convenient for analysis: Upon defining[17]

$$\underline{g} = \underline{b} - DX(t_1, t_0)\underline{a} \tag{3.3}$$

$$V(t_1, s) = DX(t_1, s)\, B(s)\, G^{-1}(s) \tag{3.4}$$

$$\underline{z}(t) = G(t)\, \underline{u}(t) \tag{3.5}$$

$$\|\underline{z}\| = \underset{t \in T}{\text{ess. sup.}} \left|\underline{z}(t)\right| \tag{3.6}$$

we can rewrite Eq. 3.2 as

$$\underline{g} = \int_T V(t_1, s)\, \underline{z}(s)\, ds \tag{3.7}$$

By assumption, G(t) is a matrix of bounded measurable functions, the inverse of which is also a matrix of bounded measurable functions. Therefore, every bounded measurable $\underline{u}(t)$ corresponds to a unique bounded measurable $\underline{z}(t)$, given by Eq. 3.5, and every bounded measurable $\underline{z}(t)$ corresponds to a unique bounded measurable $\underline{u}(t)$, given by $\underline{u}(t) = G^{-1}(t)\, \underline{z}(t)$. There is, therefore, a one-to-one correspondence between functions $\underline{u}(t)$ satisfying Eq. 3.2

---

[17] Here we use the essential supremum rather than the supremum (as originally specified) in the definition of the norm, because the essential supremum is here more convenient from a mathematical standpoint. As noted in Section 2.2, these norms are considered to be equivalent for purposes of this work, in that any control which is optimal with respect to the essential supremum norm can be converted into a control which is optimal with respect to the supremum norm simply by reducing the amplitude of the control at any moment at which it exceeds its essential supremum to a value not exceeding this essential supremum.

and functions $\underline{z}(t)$ satisfying Eq. 3.7, which establishes the complete equivalence of the original problem formulation (see Section 2.1) and the following formulation:

From the set $Q(\underline{g})$ of bounded measurable vector functions $\underline{z}(t)$, $t \in T$, which satisfy Eq. 3.7 for a particular choice of $\underline{g}$ (determined from $\underline{a}$ and $\underline{b}$ be means of Eq. 3.3), choose as an optimal control[18] $\bar{\underline{z}}(t)$ any one for which

$$\|\bar{\underline{z}}\| \leq \|\underline{z}\| \text{ for all } \underline{z} \in Q(\underline{g}) \tag{3.8}$$

The complete controllability condition on the original system (see Section 2.1) likewise has an equivalent formulation in terms of these definitions, as follows:

> Lemma: The system defined in Section 2.1 has an output which is completely controllable on the interval T a) if and b) only if for every $\underline{g} \in E^k$, there exists at least one bounded measurable vector function $\underline{z}(t)$ satisfying Eq. 3.7.

Proof: The proof of part a) follows simply by taking $u(t) = G^{-1}(t)\,\underline{z}(t)$. Part b) is proven by assuming that for some $\underline{g}$ there exists a bounded measurable $\underline{u}'(t)$ satisfying Eq. 3.2, but that there is no bounded measurable $\underline{z}(t)$ satisfying Eq. 3.7. But $\underline{z}'(t) = G(t)\,\underline{u}'(t)$ is a bounded measurable function satisfying Eq. 3.7, which is a contradiction. Q.E.D.

We note for future reference that any bounded measurable control differing from the $\underline{z}(t)$ of this lemma only on a set of measure zero will also satisfy Eq. 3.7 so that this lemma actually guarantees the existence of at least one class of controls - a class of controls which differ only on sets of measure zero.

3.2 Existence of an Optimal Control

Whether or not an optimal control exists for a given optimal control problem is a question not only of theoretical interest but of considerable practical importance as well, since from an engineering standpoint it is important to know whether or not we are seeking to do the impossible when we try to find an optimal control. In clarification of this

---

[18] For convenience, $\underline{z}(t)$ and $\underline{u}(t)$ will both be referred to as "controls" in this work. The distinction in any particular case will be clear from the context.

point, let $U(\underline{a}, \underline{b}, T)$ be the set of admissible[19] controls which cause the system in question to satisfy the boundary conditions $\underline{x}(t_0) = \underline{a}$, $\underline{y}(t_1) = \underline{b}$, and consider the two possible situations in which an optimal control may fail to exist: a) the set $U(\underline{a}, \underline{b}, T)$ may be empty, a possibility which is ruled out for problems considered in this work because of the complete controllability assumption; and b) $U(\underline{a}, \underline{b}, T)$ may contain an infinite number of controls but the cost functional $C(\underline{u})$ may not take on a minimum over this set, i.e., there may be no control $\underline{\bar{u}}(t)$ such that

$$C(\underline{\bar{u}}) = \underset{\underline{u} \in U(\underline{a}, \underline{b}, T)}{\text{infimum}} C(\underline{u})$$

[If $U(\underline{a}, \underline{b}, T)$ contains only a finite number of controls, then clearly the infimum is taken on for some $\underline{u} \in U(\underline{a}, \underline{b}, T)$ and hence an optimal control exists.]

In the second situation, no matter what control we choose, there exist other controls in $U(\underline{a}, \underline{b}, T)$ with lower cost. The engineering solution to such a problem is to use a control with cost sufficiently close[20] to the infimum noted above, rather than seek a true optimal control.

In this section a theorem is presented which guarantees the existence of a minimum peak amplitude control for the specified-final-time problem presented in Chapter II. This theorem is important not only for the reasons stated above, but because it underlies all succeeding results in this chapter as to the form and properties of minimum peak amplitude controls. We alert the reader to the fact that statements below such as "Let $\underline{\bar{z}}(t)$ denote any one of the minimum peak amplitude controls..." are meaningful only because of the existence theorem given here. Such statements occur again and again in the proofs and examples to follows.

---

19 What constitutes an admissible control of course varies from problem to problem. A typical definition might be the set of vector functions of time defined on the interval T which are piecewise continuous and have values lying in some prescribed set in the appropriate Euclidean space at each moment of time $t \in T$. Another is the one given in Chapter II for the class of problems considered in this work.

20 How close this must be will of course depend on the problem at hand. Typically, other factors, such as ease of generation of the proposed control, are brought into play to help determine the final choice of control.

Theorem 3.1: Every linear system of the type described in Section 2.1 has a minimum peak amplitude control for every set of finite boundary conditions

$$\underline{x}(t_0) = \underline{a},\ \underline{y}(t_1) = \underline{b};\ \underline{a} \in E^n,\ \underline{b} \in E^k .$$

Discussion: Every pair of boundary conditions $\underline{x}(t_0) = \underline{a}$, $\underline{y}(t_1) = \underline{b}$ determines a specific $\underline{g}$ by means of Eq. 3.3. Furthermore, there is at least one pair $\underline{a}$, $\underline{b}$ (namely $\underline{a} = \underline{0}$, $\underline{b} = \underline{g}$) which yields any given value of $\underline{g}$. In light of this and the remarks of Section 3.1, it is clear that Theorem 3.1 will be true if and only if, for each system of the stated form, and for every $\underline{g} \in E^k$, there exists a control $\bar{\underline{z}}(t)$ satisfying Eq. 3.8.

The existence of such a $\bar{\underline{z}}(t)$ will be shown in a series of steps as follows:[21]

a) Controls $\underline{z}(t)$ will be considered as points in a Banach space Z with norm defined by Eq. 3.6, and Eq. 3.7 will be shown to be a linear continuous open mapping $L(\underline{z})$ of such points into points in E^k.

b) It will be shown that the closed unit hypersphere R_1 (often called the closed unit ball) in Z maps into a closed convex absorbing set S_1 in E^k, the boundary points of which are the images of points on the boundary of R_1.

c) The hypersphere in Z will then be expanded (or contracted) until the goal $\underline{g}$ lies precisely on the boundary of the corresponding image set in E^k. This point is then the image of a point on the surface of the hypersphere in Z, and the existence of the desired $\bar{\underline{z}}$ is thereby established.

---

[21] Since the minimum peak amplitude problem is a special case of the problem treated by Neustadt in his Theorem 1, Ref. 77, (although different from any of the special cases he considered), this result follows immediately from said theorem. Certain of the conclusions of the theorem in the next section likewise are immediate consequences of the same theorem. The proofs presented here use many of Neustadt's arguments, but in a different order and with a somewhat different emphasis. The existence proof below is based on a suggestion by Prof. W. A. Porter of The University of Michigan, but follows Neustadt on the key point of the compactness of the image set in E^k.

Proof of Theorem 3.1:

a) The properties of the mapping defined by Eq. 3.7:

Consider all essentially-bounded measurable r-component vector functions of time defined on T. Group these vector functions into equivalence classes according to the equivalence relation that functions which differ from each other only on a subset of T of measure zero are in the same class. Let $\underline{z}(t)$ denote a representative member of the generic equivalence class, with norm defined by Eq. 3.6. This space Z is a Banach space[22] -- an r-dimensional version of the familiar L_∞ space.

For $V(t_1,t)$ a k x r matrix of bounded measurable functions of t, the operation

$$\int_T V(t_1,t)\,\underline{z}(t)\,dt$$

defines a linear mapping $L(\underline{z})$ of points $\underline{z}$ in Z into points in E^k. It is, in fact, a mapping of the space Z <u>onto</u> the space E^k, since, by the complete controllability assumption (as restated in the Lemma of Section 3.1), there exists at least one $\underline{z}(t) \in Z$ which maps into any given $\underline{g} \in E^k$.

The norm of this mapping is defined as

$$\|L\| = \sup_{\|\underline{z}\| = 1} L(\underline{z}) = \sup_{\|\underline{z}\| = 1} \left| \int_T V(t_1,t)\,\underline{z}(t)\,dt \right|.$$

Using the definition of the Euclidean norm of a vector (see Section 1.1), the inequality $\underline{v}\,\underline{z} \leq |\underline{v}|\,|\underline{z}|$ (which is valid for all r-component row vectors $\underline{v}$ and r-component column vectors $\underline{z}$), and the fact that, if $\|\underline{z}\| = 1$, $|\underline{z}(t)| \leq 1$ for almost all $t \in T$, it is easily shown that

$$\|L\| \leq (t_1 - t_0)\,kr\,|v_{ij}|_{sup}$$

where $|v_{ij}|_{sup}$ is the supremum over all $t \in T$ and over all i,j (i = 1,...,k; j = 1,...,r) of the absolute values of the elements $v_{ij}(t_1,t)$ of the matrix $V(t_1,t)$. Since every element of

---

[22] See Appendix E for a further discussion of this space.

$V(t_1, t)$ is bounded for all $t \in T$, because of the assumption of Chapter II, $|v_{ij}|_{sup} < \infty$ and therefore L is bounded. This implies that the set S_1 is bounded also.

From the fact that S_1 is bounded, (and from Thm. 47.A, Ref. 80, which states that if the image under a mapping from one normed linear space to another of the closed unit sphere is bounded, then the mapping is continuous), it follows that L is a continuous mapping, i.e., if $\underline{z} \in Z$ and $\underline{g} \in E^k$ are such that $L(\underline{z}) = \underline{g}$, then for every neighborhood N_g of $\underline{g}$ in E^k there exists a neighborhood N_z of $\underline{z}$ of Z such that N_g is contained[23] in $L(N_z)$. But since L is a continuous linear mapping of Z onto E^k, it is also an open mapping (see Thm. 50.A, p. 236, Simmons, Ref. 80), i.e., every open set in Z maps into an open set in E^k.

b) The image in E^k of the unit hypersphere in Z:

Now consider the closed unit hypersphere R_1 centered on the origin of Z, and let S_1 be the image in E^k of R_1 under the mapping L. That the set S_1 is convex can be shown as follows: Let $\underline{s}_1$ and $\underline{s}_2$ be any two points in S_1, and let $\underline{z}_1$ and $\underline{z}_2$ be any two points in R_1 such that $L(\underline{z}_1) = \underline{s}_1$ and $L(\underline{z}_2) = \underline{s}_2$. Because $\underline{z}_1$ and $\underline{z}_2$ are in R_1, $\|\underline{z}_1\| \leq 1$ and $\|\underline{z}_2\| \leq 1$. Convexity of S_1 will have been shown if we can show that $\theta \underline{s}_1 + (1-\theta)\underline{s}_2 \in S_1$ for all real numbers θ, $0 \leq \theta \leq 1$. Denote $\theta \underline{s}_1 + (1-\theta)\underline{s}_2$ by $\underline{s}$. Since $\underline{s} = \theta L(\underline{z}_1) + (1-\theta) L(\underline{z}_2) = L(\theta \underline{z}_1) + L[(1-\theta)\underline{z}_2] = L[\theta \underline{z}_1 + (1-\theta)\underline{z}_2]$, because of the linearity of L, and since

$$\|\theta \underline{z}_1 + (1-\theta)\underline{z}_2\| \leq \|\theta \underline{z}_1\| + \|(1-\theta)\underline{z}_2\| = \theta \|\underline{z}_1\| + (1-\theta)\|\underline{z}_2\| \leq \theta + (1-\theta) = 1$$

by the triangle inequality and the bounds on $\|\underline{z}_1\|$ and $\|\underline{z}_2\|$ noted above, it follows that $\underline{s}$ is indeed the image of a point in R_1, and hence that $\underline{s} \in S_1$.

The set S_1 is also symmetrical about the origin in E^k (which implies that if $\underline{s} \in S_1$, then $-\underline{s} \in S_1$), since $L(-\underline{z}) = -L(\underline{z})$, because of the linearity of L, and $-\underline{z} \in R_1$ if $\underline{z} \in R_1$.

The set S_1 is an absorbing set (i.e., for every nonzero vector $\underline{g} \in E^k$, S_1 contains at least one vector $\underline{s}_g$ with Euclidean norm bounded away from zero which is colinear

---

[23] The symbol $L(N_z)$ represents the image of the set N_z under the mapping L; that is, $L(N_z)$ is the set of all points in E^k which are images under L of points in N_z. This notation for the image of a set is used throughout this work.

with $\underline{g}$) because of the complete controllability assumption. To show this, we exhibit one such vector $\underline{s}_g = L(\underline{z}_g)$, which we obtain by normalizing (to unit norm in Z-space) the control used in Appendix D to prove the sufficiency of the condition for complete controllability on the interval T. This procedure yields

$$\underline{z}_g(t) = \frac{1}{K} V^T(t_1,t)\, W^{-1}\, \underline{g}$$

$$\underline{s}_g = L(\underline{z}_g) = \frac{1}{K} \int_T V(t_1,t)\, V^T(t_1,t)\, W^{-1}\, \underline{g}\, dt = \frac{1}{K} WW^{-1}\, \underline{g} = \frac{1}{K}\, \underline{g}$$

and

$$\left|\underline{s}_g\right| = \frac{1}{K} \left|\underline{g}\right|,$$

where W is a constant k x k matrix given by

$$W = \int_T V(t_1,t)\, V^T(t_1,t)\, dt$$

and K, the constant which adjusts $\underline{z}_g$ to unit norm in Z-space, is defined as

$$K = \underset{t \in T}{\text{ess. sup.}} \left| V^T(t_1,t)\, W^{-1}\, \underline{g} \right|.$$

Because of the complete controllability assumption (see Appendix D), W is positive definite, and therefore, its inverse exists and has bounded eigenvalues. Since every element of $V^T(t_1,t)$ is uniformly bounded in absolute value on T by the finite number $|v_{ij}|_{sup}$ defined above, the matrix $V^T(t_1,t)\, W^{-1}$ is uniformly bounded in norm[24] on T. Denote any such bound by N. Then $\left|V^T(t_1,t)W^{-1}\underline{g}\right| \leq N\left|\underline{g}\right|$ on T, which yields $K \leq N\left|\underline{g}\right|$, which in turn yields $\left|\underline{s}_g\right| \geq \frac{1}{N}$. Since N is less than infinity, $\underline{s}_g$ is a vector in the same direction as $\underline{g}$ and with Euclidean norm bounded away from zero, as claimed.

Every point of S_1 which is on the boundary of S_1 is the image of some point on the boundary of R_1. This can be shown as follows: Assume that there exists some

---

[24]The norm $\|A\|$ of any matrix A, operating on vectors $\underline{x}$, is here defined as

$$\|A\| = \sup_{|\underline{x}| = 1} |A\, \underline{x}|.$$

boundary point $\underline{s}'$ of S_1 which is in S_1 and which is the image of an interior point[25] $\underline{z}'$ of R_1. Because L is an open mapping, every neighborhood of $\underline{z}'$ maps into an open set containing $\underline{s}'$. Since $\underline{z}'$ is an interior point of R_1, there is a neighborhood of $\underline{z}'$ which is entirely contained within R_1 which maps into an open set containing $\underline{s}'$. But every open set containing $\underline{s}'$ contains points which are outside of S_1. This is a contradiction, since every point of R_1 maps into some point of S_1, and, therefore, $\underline{s}'$ cannot be the image of an interior point of R_1. Since $\underline{s}'$ is the image of some point in R_1, by definition, it then follows that it is the image of a point on the boundary of R_1, as claimed.

Finally, the set S_1 is closed[26] (in terms of the usual Euclidean metric on E^k). To show this, we adopt for the moment a different point of view from that taken above: Consider the set of r-component measurable row vector functions of time $\underline{v}(t)$ defined on the interval T, the Euclidean norms of which are integrable on T. Define the norm of $\underline{v}(t)$ by

$$\|\underline{v}\|_1 = \int_T |\underline{v}(t)|\, dt$$

and denote by B the vector space consisting of these functions with this norm. This is a Banach space[27]--an "r-dimensional" version of the familiar L_1 space. The rows of the matrix $V(t_1,t)$, considered as functions of t, are elements of this space.

The conjugate space B* of this space is the space of bounded linear functionals $z(\underline{v})$ defined over B, with norm given by

$$\|z\| = \sup_{\|\underline{v}\|_1 \leq 1} |z(\underline{v})|$$

It is easily shown that all functionals of the form

---

[25]The set R_1 has interior points, of course. For example, the open sphere of unit radius centered on the origin consists wholly of interior points of R_1.

[26]General references for this part of the proof: Neustadt, Ref. 77; Hille and Phillips, Ref. 84, pp. 26-39; Simmons, Ref. 80, pp. 211-242; Taylor, Ref. 81, pp. 160-252.

[27]This space and the representation theorem for functionals on this space are discussed in Appendix E.

$$z(\underline{v}) = \int_T \underline{v}(t)\,\underline{z}(t)\,dt$$

are in the space B*, where $\underline{z}(t)$ is an essentially bounded r-component column vector function of time defined on T. Not so obvious, but nonetheless true,[28] is the fact that every functional in B* is representable in this form.

Furthermore, the above definition of the norm on B* takes on a particularly convenient form in terms of this representation, namely,[29]

$$\|z\| = \underset{t \in T}{\text{ess. sup.}} \left|\underline{z}(t)\right| = \|\underline{z}\|$$

where $\|z\|$ is defined in Eq. 3.6.

Thus, every functional z in B* corresponds to an equivalence class[30] in Z, and every equivalence class corresponds to a functional in B*. Furthermore, the norms of the corresponding members of the two spaces are the same. The two spaces are therefore isometrically equivalent to each other (Dunford and Schwartz, Ref. 85, p. 65), which allows certain conclusions which are drawn below about B* to be applied to Z also.

Let B** denote the second conjugate space of B; i. e., the space of all bounded linear functional v**(z) defined on the elements z of B* (i. e., on the elements $\underline{z}(t)$ of Z), and then consider the natural embedding of B in B**, in which a correspondence is established between each element of B and an element of a subset of B** by means of the relation

$$v^{**}(\underline{z}) = z(\underline{v})$$

---

[28] See Appendix E.

[29] See Appendix E.

[30] The representation for a given functional is not unique, since any vector $\underline{z}'(t)$ which differs from $\underline{z}(t)$ only on a set of measure zero could be used in place of $\underline{z}(t)$. Thus, the situation is analogous to that encountered in the definition of the space Z above, and the kernel functions $\underline{z}(t)$ can be grouped into equivalence classes in exactly the same way as were the controls occurring in the definition of Z.

Now consider the weak * topology on B*; that is, the weakest topology on B* for which all the functionals in B** which correspond to elements of B under the natural embedding remain continuous. We now complete the proof that S_1 is closed in three steps:

1) The closed unit sphere R_1 of Z is a compact Hausdorff space in the weak* topology. This follows from Theorem 49.A, Simmons, Ref. 80, p. 233, (which states that the closed unit sphere of the conjugate space of a normed linear space is a compact Hausdorff space in the weak * topology), and from the fact that Z is isometrically equivalent to B*.

2) The continuous linear transformation $L(\underline{z})$ defined above, which maps elements of Z (i.e., B*) into E^k, is continuous in the weak* topology on Z and the topology generated by the Euclidean norm on E^k. To show this, we note that $L(\underline{z})$ actually consists of k functionals from B**, each of which maps B* into E^1. By the definition of the weak* topology, each of these k functionals is continuous in the weak* topology. This implies (by the definition of a continuous mapping, Simmons, Ref. 80, p. 93) that for each of these functionals the inverse image of each open interval in E^1 is an open set in Z in the weak* topology. Consider any one of these functionals $v_i^{**}(\underline{z})$, and any open interval M_i on the ith coordinate axis of E^k, and let Q_i be the open set in Z which is the inverse image of M_i under the mapping v_i^{**}. Then Q_i is also the inverse image under the mapping $L(\underline{z})$ of the "open strip" in E^k defined by $[\underline{g}:\underline{g} \in E^k,\ g_i\ G\ M_i]$, since $v_i^{**}(\underline{z})$ is just the ith component of $L(\underline{z})$. The set of all such open strips for all $i = 1, \ldots, k$ forms an open subbase for E^k(Simmons,

Ref. 80, pp. 99-104). We have thus shown that the inverse image of each subbasic open set of E^k with respect to the mapping $L(\underline{z})$ is an open set. This implies that $L(\underline{z})$ is continuous in the stated topologies, as claimed (Simmons, Ref. 80, Theorem 18.E, p. 103).

3) The set S_1 is closed (in terms of the usual Euclidean metric on E^k). To show this, we first apply Theorem 21.B, Simmons, Ref. 80, p. 111, which states that if L is a continuous mapping from a compact topological space into another topological space, then the image of the first space under this mapping is a compact subset of the second. Since, from part 1) above, R_1 is a compact Hausdorff space in the weak* topology, and since from part 2) above, L is a continuous mapping of R_1 into E^k, (in the topologies already mentioned for these spaces), it follows that S_1 is a compact subset of E^k. But from Lemma 7, c), Section I. 5. 6, Dunford and Schwartz, (Ref. 85), and from the fact that the metric space E^k is also a Hausdorff space (Simmons, Ref. 80, p. 134), it then follows that S_1 is closed (in the topology generated by the usual Euclidean norm on E^k).

c) Determination of the minimum cost:

Now consider any given goal point $\underline{g} \in E^k$. This point is either inside, on the boundary, or outside of S_1. If it is on the boundary of S_1, it is in S_1, since S_1 is closed. Because every point of S_1 which is on the boundary of S_1 is the image of a point on the boundary of R_1, and because there is no interior point of R_1 which maps into a boundary point of S_1, the existence of an optimal control $\bar{\underline{z}}(t)$ is proven in this case, and $\|\bar{\underline{z}}\| = 1$. If $\underline{g}$ is outside (inside) of S_1, we choose a larger (smaller) closed hypersphere R in Z, centered on

the origin, such that its image S in E^k contains $\underline{g}$ as a boundary point. In precise mathematical terms, this process involves the evaluation of the Minkowski functional[31] of the set S_1 at the point $\underline{g}$ (denote its value at $\underline{g}$ by $\overline{C}$.) and then the choosing of R as the closed hypersphere in Z centered on the origin and with radius equal to $\overline{C}$. Then S = L(R) contains $\underline{g}$ as a boundary point, S has the same properties of compactness, convexity, etc., as S_1, and the reasoning used above for the case in which $\underline{g}$ was on the boundary of S_1 again applies. Since S_1 contains vectors in every possible direction which are bounded away from zero in norm, $\overline{C}$ will always be finite if $|\underline{g}|$ is finite. The existence of an optimal control $\underline{\overline{z}}(t)$ with norm $\overline{C}$ is thus established. This completes the proof of Theorem 3.1.

3.3 Some Generalizations

At the end of Section 2.2, two generalizations of the minimum peak amplitude problem, as originally stated, were discussed. In the first of these, a more general system description, in which the control entered directly into the output, was involved. Note from Eq. 3.1 (which applies to these problems also) that the control can be perturbed arbitrarily at a single moment in time (which is of course a set of measure zero) without affecting the state $\underline{x}(t)$, but that for this class of problems such a perturbation <u>does</u> affect the output $\underline{y}(t)$ at that moment (see Eq. 2.4). Note also that in the problem considered here the value of $\underline{y}(t)$ is important <u>only</u> at the final time t_1, no requirements having been placed on it at any other time. This implies the following: Suppose we consider some control $\underline{u}(t)$ for which $\underline{y}(t_1)$ is <u>not</u> equal to the desired goal $\underline{g}$. If no component of the vector $\underline{g} - \underline{y}(t_1)$ lies in the null space of the matrix E(t) at time t_1, then we can choose another control $\underline{u}'(t)$ which differs from $\underline{u}(t)$ only at the time t_1 and which causes the corresponding $\underline{y}'(t_1)$ to equal $\underline{g}$; namely

$$\underline{u}'(t) = \underline{u}(t) \quad t \in [t_0, t_1)$$

$$\underline{u}'(t_1) = \underline{u}(t_1) + \Delta\underline{u}$$

where $\Delta\underline{u}$ is such that $E(t_1)\,\Delta\underline{u} = \underline{g} - \underline{y}(t_1)$. In terms of the essential supremum cost functional, $\underline{u}(t)$ and $\underline{u}'(t)$ have the same cost. Thus, if this is the cost functional which is meaningful in a given problem, it is clear that the optimal control problem for systems of the type

---

[31] See pp. 134-136, Taylor, Ref. 81.

shown in Eq. 2.4 can be solved in the same way as the problem as originally posed (Section 2.1), if we first subtract from the goal $\underline{g}$ any component of $\underline{g}$ which lies in the space spanned by the columns of E(t) at time t_1. We leave this component of $\underline{g}$ to be achieved by an appropriate choice of $\Delta\underline{u}$ at $t = t_1$. If $E(t_1)$ happens to be of rank k at $t = t_1$, then the optimal control consists only of a $\Delta\underline{u}$ of the appropriate magnitude and "direction" at $t = t_1$; with the control being zero for all other $t \in T$. The essential supremum of the Euclidean norm of such a control function is of course zero.

If, on the other hand, we wish to use the supremum cost functional, the situation is different, since the required value of $|\underline{u}'(t_1)|$ may be greater than the supremum of $|\underline{u}'(t)|$ over the rest of the interval. We must therefore carefully choose $\underline{u}(t)$ on $[t_0, t_1)$ so as to leave unaccomplished only that much of the over-all goal $\underline{g}$ as can be attained using a $\underline{u}'(t_1)$ for which $|\underline{u}'(t_1)|$ does not exceed $\sup_{t \in [t_0, t_1)} |\underline{u}(t)|$. Further reasoning along this line then leads to the following conclusions:

The existence proof of the previous section can be modified to apply to this problem by replacing the image S_1 of the closed unit hypersphere R of Z with a set obtained by constructing at each point of the boundary of S_1 a hyperellipsoid corresponding to all the points that can be reached from that point as a result of the direct contribution of the term $E(t)\,\underline{u}(t)$ in Eq. 2.4 using controls with norm not exceeding unity at this last instant. The union of S_1 and all these hyperellipsoids is then the reachable set for the generalized problem, and all the rest of the steps in the existence proof go through with little change. To compute numerical values of the optimal controls for such problems, we can follow the procedure just outlined for the existence proof, or else use the following iterative technique: A guess C is made as to the value of the minimum peak amplitude, and the goal $\underline{g}$ is replaced by the set of all points in E^k from which $\underline{g}$ can be reached using a control $\underline{u}'(t_1)$ for which $|\underline{u}'(t_1)| \leq C$. The problem of choosing the optimal control on the interval $[t_0, t_1)$ to force the system into this set is then solved, and the corresponding peak amplitude C' is determined. If C' is less than the original guess C, the original guess is reduced and the procedure is repeated. If it is greater, the original guess is increased, and so on, until the two values are equal.

The second generalization mentioned at the end of Section 2.2 involved initial states as well as final states in some linear manifold. This likewise presents no problems with respect to the existence of optimal controls, for the following reasons: If the initial state lies in some linear manifold M (rather than at a single given point), then the goal point $\underline{g}$ must likewise be replaced by a linear manifold, defined as all points in E^k of the form

$$\underline{b} - X(t_1, t_0)\, \underline{x}(t_0);\ \underline{x}(t_0) \in M$$

Such a manifold is a closed set. The step in the existence proof in which the reachable set S_1 is expanded (or contracted) until it just touches $\underline{g}$ thus has an analog here, and the existence proof once again goes through with little change. Actual computation of optimal solutions is considerably more difficult, however, and is not investigated here.

If the initial time t_0 is not specified in advance, but must be chosen as part of the problem solution, it is no longer necessarily true that an optimal solution exists. The difficulties here are the same as those which occur when the final time is not specified in advance. These difficulties are discussed in Section 3.8.

3.4 The Form of the Optimal Solution

The following lemma, which may be regarded as a vector version of Hölder's inequality with $p = \infty$, will be useful in the proof of Theorem 3.2 below.

Lemma: Let $\underline{v}(t)$ and $\underline{z}(t)$ be r-component vector functions[32] of time defined on the interval T and belonging to the spaces B and Z (defined in the proof of Theorem 3.1), respectively. Then

$$\int_T \underline{v}(t)\,\underline{z}(t)dt \leq C \int_T |\underline{v}(t)|\,dt \tag{3.10}$$

where $C = \|\underline{z}\| = \underset{t \in T_1}{\text{ess. sup.}} |\underline{z}(t)|$, and where T_1 is the subset of T on which $|\underline{v}(t)| \neq 0$. Furthermore, the equality is taken on in this inequality if and only if $\underline{z}(t)$ satisfies

[32] Note that $\underline{v}$ is a row vector and $\underline{z}$ is a column vector.

$$\underline{z}(t) = C \frac{\underline{v}^T(t)}{\left|\underline{v}^T(t)\right|} \quad \text{for almost all } t \in T_1 \tag{3.11}$$

This lemma places no restrictions on $\underline{z}(t)$ on the set $T - T_1$.

Proof:

$$\int_T \underline{v}(t)\,\underline{z}(t)dt \leq \left| \int_T \underline{v}(t)\,\underline{z}(t)dt \right| \leq \int_T \left|\underline{v}(t)\,\underline{z}(t)\right| dt$$

$$\leq \int_T \left|\underline{v}(t)\right| \left|\underline{z}(t)\right| dt \leq C \int_T \left|\underline{v}(t)\right| dt$$

The first inequality becomes an equality if and only if $\int_T \underline{v}(t)\,\underline{z}(t)dt$ is nonnegative. The second becomes an equality if and only if the scalar quantity $\underline{v}(t)\,\underline{z}(t)$ is of the same sign (excluding points at which it is zero). The third becomes an equality if and only if $\underline{v}^T(t)$ and $\underline{z}(t)$ are colinear almost everywhere on T. The last inequality, which is just the usual scalar form of Hölder's inequality, becomes an equality if and only if $\underline{z}(t) = \underset{t \in T_1}{\text{ess. sup.}} \left|\underline{z}(t)\right| = C$ almost everywhere on T_1. These four conditions combine to show that the equality can be taken on in inequality 3.10 if and only if Eq. 3.11 is satisfied. We note in passing that the condition of equality says nothing about $\underline{z}(t)$ on the set $T - T_1$; i.e., the set on which $\left|\underline{v}(t)\right| = 0$. The implications of this fact for the problem at hand will be discussed later.

Theorem 3.2: Every minimum peak amplitude control $\overline{\underline{z}}(t)$, with norm $\|\overline{\underline{z}}\| = \overline{C}$, the existence of at least one of which is guaranteed by Theorem 3.1, is of the form

$$\left.\begin{aligned} &\overline{\underline{z}}(t) = \overline{C} \frac{V^T(t_1,t)\overline{\underline{c}}}{\left|V^T(t_1,t)\overline{\underline{c}}\right|} \quad \text{for almost all } t \in T_1(\overline{\underline{c}}) \\ &\text{On the set } T - T_1(\overline{\underline{c}}),\ \overline{\underline{z}}(t) \text{ is restricted only by the condition} \\ &\left|\overline{\underline{z}}(t)\right| \leq \overline{C} \quad \text{for almost all } t \in T - T_1(\overline{\underline{c}}) \end{aligned}\right\} \tag{3.12}$$

Here $\overline{\underline{c}}$ is a nonzero vector in E^k and $T_1(\overline{\underline{c}})$ is the subset of T on which $\left|V^T(t_1,t)\overline{\underline{c}}\right| \neq 0$. Moreover, $\overline{\underline{c}}$ must be an outward normal to the set S at the point $\underline{g}$, where S is the image under the mapping L of the set $R = [\underline{z} : \underline{z} \in Z,\ \|\underline{z}\| \leq \overline{C}]$. Also, if the outward normal to

S at $\underline{g}$ is not uniquely defined,[33] then any one of the outward normals may be used as $\underline{\bar{c}}$. Finally,

$$\bar{C} = \sup_{\underline{c} \neq 0} \frac{(\underline{c}, \underline{g})}{\int_T \left| V^T(t_1, t)\underline{c} \right| dt} \tag{3.13}$$

and this supremum is taken on only for $\underline{c}$'s which are outward normals to S at $\underline{g}$.

Proof: From Theorem 3.1, we know that an optimal control exists for every goal $\underline{g} \in E^k$. Let the minimum peak amplitude corresponding to the given $\underline{g}$ be denoted by $\bar{C}$, and let $\underline{\bar{z}}(t)$ denote any one of the minimum peak amplitude controls yielding this $\underline{g}$; i.e., $L(\underline{\bar{z}}) = \underline{g}$. Let $\underline{\bar{c}}$ be any outward normal to the set S at the point $\underline{g}$. (Because of the way S is constructed, $\underline{g}$ is on the boundary of S.) Finally, let $\underline{z}(t)$ be any control in the set R, and let $\underline{s} = L(\underline{z})$.

Clearly, $\underline{s}$ is in S.

Now form the inner product[34] $(\underline{\bar{c}}, \underline{g})$ of $\underline{\bar{c}}$ and $\underline{g}$. Then

$$(\underline{\bar{c}}, \underline{g}) = \sup_{\underline{s} \in S} (\underline{\bar{c}}, \underline{s}) \tag{3.14}$$

as can be seen by considering the hyperplane of support of S at $\underline{g}$ to which $\underline{\bar{c}}$ is the normal, and noting that because of the convexity of S (established in the proof of Theorem 3.1) no vector in S has a component normal to this hyperplane which is greater than that of $\underline{g}$. Now, since every point in R maps into S, taking the supremum over all $\underline{s} \in S$ is equivalent to taking the supremum over all $\underline{z} \in R$. Equation 3.14 can thus be rewritten

$$(\underline{\bar{c}}, L(\underline{\bar{z}})) = \sup_{\underline{z} \in R} (\underline{\bar{c}}, L(\underline{z}))$$

or

$$\int_T \underline{\bar{c}}^T V(t_1, t)\, \underline{\bar{z}}(t) dt = \sup_{\underline{z} \in R} \int_T \underline{\bar{c}}^T V(t_1, t)\, \underline{z}(t) dt \tag{3.15}$$

---

[33] Where S does not have a unique outward normal, such as at a corner, we say that any vector which is an outward normal to a hyperplane of support of S at that point is an outward normal to S at that point.

[34] Here and in the sequel, the inner product of two vectors $\underline{\theta}$ and $\underline{\eta}$ in m-dimensional Euclidean space is denoted by $(\underline{\theta}, \underline{\eta})$ and is given by

$$(\underline{\theta}, \underline{\eta}) = \sum_{i=1}^{m} \theta_i \eta_i$$

Using $V^T(t_1,t)\bar{c}$ as $v(t)$ in the above lemma, we have that

$$\int_T \bar{c}^T V(t_1,t)\, z(t)dt \leq \operatorname*{ess.sup.}_{t \in T_1} \left|z(t)\right| \int_T \left|V^T(t_1,t)\bar{c}\right|dt$$

where T_1 is the subset of T on which $\left|V^T(t_1,t)\bar{c}\right| \neq 0$. Furthermore, the equality is taken on only if $z(t)$ is of the form $C\, V^T(t_1,t)\bar{c} \Big/ \left|V^T(t_1,t)\bar{c}\right|$ a. e. on T_1. Since we must find the supremum of this integral over all $z(t)$ in R, it is thus clear that we must choose $z(t)$ of the form $\bar{C}\, V^T(t_1,t)\bar{c} \Big/ \left|V^T(t_1,t)\bar{c}\right|$ a. e. on T_1. Also, since $z(t)$ must be in R, it follows that $\left|z(t)\right| \leq \bar{C}$ a. e. on $T - T_1$. Substituting this result into Eq. 3. 15 then gives

$$\int_T \bar{c}^T V(t_1,t)\, \bar{z}(t)dt = \bar{C} \int_T \bar{c}^T V(t_1,t) \frac{V^T(t_1,t)\bar{c}}{\left|V^T(t_1,t)\bar{c}\right|} dt = \bar{C} \int_T \left|V^T(t_1,t)\bar{c}\right|dt$$

But this is precisely the situation covered by the above lemma, with $\bar{c}^T V(t_1,t)$ taken as the row vector $v(t)$. Therefore, from this lemma, it follows that $\bar{z}(t)$ must be of the form

$$\bar{z}(t) = \bar{C} \frac{V^T(t_1,t)\bar{c}}{\left|V^T(t_1,t)\bar{c}\right|} \quad \text{for almost all } t \in T_1(\bar{c})$$

Since $\bar{z}(t)$ was defined as <u>any</u> one of the optimal controls yielding g, the first part of Eq. 3. 12 is established. The second part of Eq. 3. 12 is obviously true, since the essential supremum of $\left|z(t)\right|$ is known to be $\bar{C}$, and a function can exceed its essential supremum at most on a set of measure zero.

We now show that no vector which is not an outward normal to S at g can be used in place of $\bar{c}$ in Eq. 3. 12: Let c be any vector in E^k which is <u>not</u> an outward normal to S at g, and note that

$$(c,g) < \max_{s \in S} (c,s)$$

(Like Eq. 3. 14, this follows from the convexity of S: If c is not normal to S at g, then it is normal to S at some other point, call it h. Then (c,g) will be less than (c,h), by the same argument used to establish Eq. 3. 14.) Repeating the steps used in deriving Eq. 3. 15 from Eq. 3. 14 then gives

$$(c,g) = \int_T c^T V(t_1,t)\, \bar{z}(t)dt < \bar{C} \int_T \left|V^T(t_1,t)c\right|dt$$

We now use the above lemma to conclude that $\bar{z}(t)$ can not be of the form $\bar{C}\,\dfrac{V^T(t_1,t)\underline{c}}{|V^T(t_1,t)\underline{c}|}$, since if it were of this form the equality would be taken on in the lemma, which would contradict the above inequality.

The final statement of the theorem is now easily proven: Since from the above inequality

$$\bar{C} > \frac{(\underline{c},\underline{g})}{\int_T |V^T(t_1,t)\underline{c}|\,dt}$$

for all $\underline{c}$ which are not outward normals to S at $\underline{g}$, and since, from Eq. 3.15,

$$\bar{C} = \frac{(\bar{\underline{c}},\underline{g})}{\int_T |V^T(t_1,t)\bar{\underline{c}}|\,dt}$$

for all $\bar{\underline{c}}$ which are outward normals to S at $\underline{g}$, the final assertion of the theorem follows at once. Q.E.D.

Discussion: A comparison of the foregoing theorems and proofs with the results for other problems and approaches reveals a close similarity in the underlying concepts. For instance, the set S is clearly the reachable set at the time t_1 for the given system, starting from the origin at time t_0, and for $\underline{z}(t)$ restricted by $|\underline{z}(t)| < \bar{C}$. The fact that the optimal control involves $\bar{\underline{c}}$, a normal to this reachable set, is equivalent to the fact that the Lagrange multiplier vector $\underline{\psi}(t)$ arising in the calculus of variations solution of many linear systems problems is normal to the final manifold at the final time. Finally, the fact that $|\bar{\underline{z}}(t)| = \bar{C}$ on $T_1(\bar{\underline{c}})$ reflects the fact that the trajectory lies on the boundary of the reachable set at such moments.

The question naturally arises as to the nature of the optimal control on the set $T - T_1(\bar{\underline{c}})$. In fact, since $\bar{\underline{c}}$ itself may not be uniquely determined (i.e., the reachable set may have a "corner" or "edge" at the point in question), a second question concerns the different optimal controls which result from different choices of $\bar{\underline{c}}$. These two questions will be investigated in turn. The examples in the next section show that if the set $T - T_1(\bar{\underline{c}})$ has measure greater than zero, the optimal control is not uniquely determined, in general.

3.5 Two Examples Involving Nonunique Optimal Control

Example 1:

The system to be considered here is characterized by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_{11}(t) & b_{12}(t) \\ 0 & b_{22}(t) \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

where

$$b_{11}(t) = \begin{cases} 1-t & 0 \leq t \leq 1 \\ 0 & 1 \leq t \leq 2 \end{cases}$$

$$b_{12}(t) = \begin{cases} 0 & 0 \leq t \leq 1 \\ -(t-1)(2-t) & 1 \leq t \leq 2 \end{cases}$$

$$b_{22}(t) = \begin{cases} 0 & 0 \leq t \leq 1 \\ t-1 & 1 \leq t \leq 2 \end{cases}$$

The state $\underline{x}(t)$ is taken as the output for this example, so that D = I. We note in passing that the matrix B(t) is a continuous function of time on the given interval.

From the set $U(\underline{0}, \underline{b}, T)$ of bounded measurable controls $\underline{u}(t)$ defined on the interval T = [0, 2] which cause this system to transfer from the initial state $\underline{x}(0) = \underline{0}$ to the desired final state $\underline{x}(t_1) = \underline{b}$ at specified final time $t_1 = 2$, choose as an optimal control any control $\bar{\underline{u}}(t)$ for which $C(\bar{\underline{u}}) = \underset{t \in T}{\text{ess. sup.}} \left|\bar{\underline{u}}(t)\right|$ takes on its minimum value with respect to all $\underline{u}(t) \in U(\underline{0}, \underline{b}, T)$.

For this system

$$X(t_1, t) = \begin{bmatrix} 1 & t_1 - t \\ 0 & 1 \end{bmatrix}$$

Hence, for $t_1 = 2$,

$$V(t_1,t) = V(2,t) = X(2,t)\,B(t) = \begin{cases} \begin{bmatrix} 1-t & 0 \\ 0 & 0 \end{bmatrix} & 0 \le t \le 1 \\ \begin{bmatrix} 0 & 0 \\ 0 & t-1 \end{bmatrix} & 1 \le t \le 2 \end{cases}$$

Using Eq. 3.1,

$$\begin{bmatrix} x_1(2) \\ x_2(2) \end{bmatrix} = \int_0^1 \begin{bmatrix} 1-t & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} dt + \int_1^2 \begin{bmatrix} 0 & 0 \\ 0 & t-1 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} dt$$

$$= \int_0^1 \begin{bmatrix} (1-t)\,u_1(t) \\ 0 \end{bmatrix} dt + \int_1^2 \begin{bmatrix} 0 \\ (t-1)\,u_2(t) \end{bmatrix} dt \qquad (3.16)$$

Therefore,

$$x_1(2) = \int_0^1 (1-t)\,u_1(t)dt$$

and

$$x_2(2) = \int_1^2 (t-1)\,u_2(t)dt \qquad (3.17)$$

From these expressions it follows at once[35] that the set S_1 defined in the proof of Theorem 3.1 (in this case, the set of all points $\underline{x}(t)$ in E^k which can be obtained from Eq. 3.16 using $\underline{u}(t)$ for which $|\underline{u}(t)| \le 1$, $t \in T$) is the square S_1 shown in Fig. 3.1. Suppose that the desired final condition on $\underline{x}(2)$ is $\underline{x}(2) = \underline{b} = \underline{g} = [0.8, 0.6]^T$. Following the procedure of part c) of the proof of Theorem 3.1, or alternatively, using Eq. 3.13, we obtain that $\overline{C} = 1.6$. An outward normal to the set S at $\underline{g}$ is $\underline{\overline{c}} = [1,0]^T$ and its direction is uniquely defined as can be seen

---

[35] Any point (g_1, g_2) inside or on the boundary of this square can clearly be obtained by choosing $\underline{u}(t) = [2g_1, 0]^T$, $0 \le t \le 1$ and $\underline{u}(t) = [0, 2g_2]^T$, $1 < t \le 2$, for example. Such controls satisfy the condition $|\underline{u}(t)| \le 1$, $t \in T$. By applying the lemma of Section 3.4 to Eqs. 3.17, we can also show that neither $|x_1(2)|$ nor $|x_2(2)|$ can exceed $\frac{1}{2}$, that the value $x_1(2) = \frac{1}{2}$ can be taken on only if $\underline{u}(t) = [1,0]^T$ almost everywhere on $[0,1]$, and that similarly the value $x_2(2) = \frac{1}{2}$ can be taken on only if $\underline{u}(t) = [0,1]^T$ almost everywhere on $[1,2]$.

from inspection of Fig. 3.1. Using these values of $\bar{C}$ and $\bar{c}$ in Theorem 3.2, we have that

$$\bar{u}(t) = \bar{z}(t) = (1.6)\begin{bmatrix}1\\0\end{bmatrix}, \quad 0 \leq t \leq 1$$

with $\bar{u}(t)$ being thus-far undefined on the interval $1 < t \leq 2$ except for the requirement that $|\bar{u}(t)| \leq 1.6$ almost everywhere on this interval. The optimum control is not uniquely determined on this interval, and in fact there are an infinite number of optimal controls for thie particular choice of final condition. One family of optimal controls is given by

$$\bar{u}(t) = \begin{cases} (1.6)\begin{bmatrix}1\\0\end{bmatrix} & 0 \leq t \leq 1 \\ (1.6)\begin{bmatrix}0\\K\end{bmatrix} & 1 < t \leq 1 + \sqrt{\frac{3}{4K}} \\ \begin{bmatrix}0\\0\end{bmatrix} & 1 + \sqrt{\frac{3}{4K}} < t \leq 2 \end{cases}$$

where $\frac{3}{4} \leq K \leq 1$, as can be verified by direct substitution in Eq. 3.16. Any number of other optimal controls could be devised.

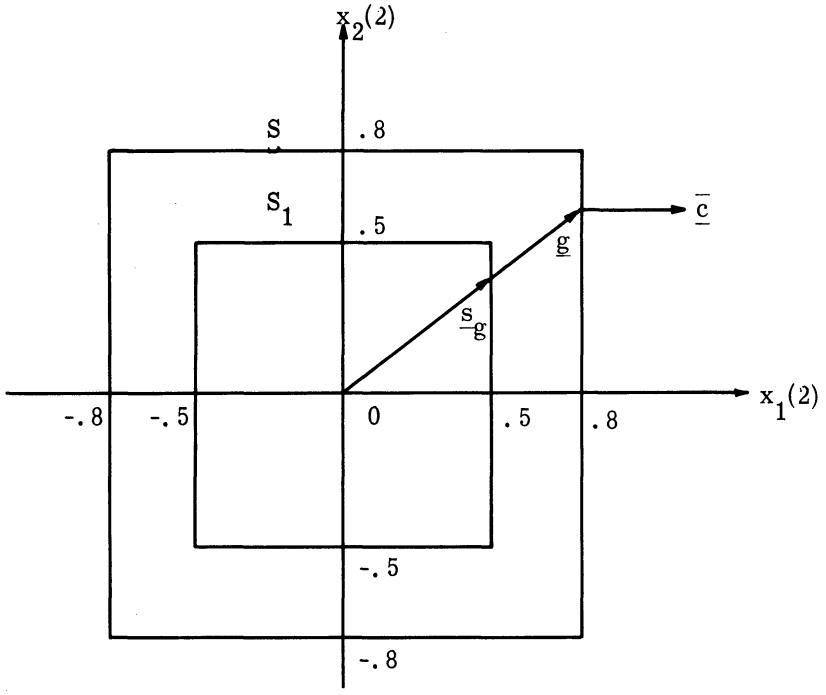

Fig. 3.1 The sets S_1 and S for Example 1.

Let us now consider the situation at one of the corners of S_1. To be specific, choose the corner at $(\frac{1}{2}, \frac{1}{2})$. Here the direction of the normal to S_1 is not uniquely defined, so that all vectors of the form $\underline{c} = [\cos\theta, \sin\theta]^T$, $0 \leq \theta \leq \frac{\pi}{2}$, are outward normals here.

It is interesting to note that, except for the two $\underline{c}$ corresponding to $\theta = 0$ and $\theta = \frac{\pi}{2}$, every one of these normal vectors yields the same control $\underline{u}(t) = [1, 0]^T$, $0 \leq t < 1$; $\underline{u}(t) = [0, 1]^T$, $1 < t \leq 2$, defined uniquely except at the point $t = 1$. (This can be verified by direct substitution in Eq. 3.12 of Theorem 3.2.) Furthermore, this control is included in the set of incompletely specified controls determined by the substitution of the two limiting normals in Eq. 3.12 of Theorem 3.2, so that the theorem is not contradicted. Finally, any control which differs from this control on a set of measure greater than zero can not yield $\underline{x}(2) = [\frac{1}{2}, \frac{1}{2}]^T$, as indicated in Footnote 35 following Eq. 3.17.

In summary, this example has shown both a situation in which the direction of the normal to S (or S_1) is uniquely determined but the optimal control is not, and a situation in which the direction of the normal is not uniquely defined but the control is (to within a set of measure zero).

One must not conclude from this that the optimal control is always unique when the direction of the normal is not uniquely defined, as can be shown by the following example, which is similar to this one, but in three-dimensional space:

Example 2:

From the set $U(\underline{0}, \underline{b}, T)$ of bounded measurable controls defined on the interval $T = [0, 3]$ which cause the system characterized by

$$\dot{\underline{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \underline{x} + \begin{bmatrix} b_{11}(t) & b_{12}(t) & b_{13}(t) \\ 0 & b_{22}(t) & b_{23}(t) \\ 0 & 0 & b_{33}(t) \end{bmatrix} \underline{u}$$

to transfer from the given initial state $\underline{x}(0) = \underline{0}$ to the final state $\underline{x}(t) = \underline{b}$ at specified final time $t_1 = 3$, choose as an optimal control any one for which $C(\underline{u}) = \underset{t \in T}{\text{ess. sup.}} \left|\underline{u}(t)\right|$ is minimum with respect to all $\underline{u} \in U(\underline{0}, \underline{b}, T)$. Here, the coefficients $b_{ij}(t)$ are all zero except as follows: $b_{11}(t) = (1-t)$ on the interval $[0, 1]$; $b_{12}(t) = -3(t-1)(2-t)(3-t)$ and $b_{22}(t) = 3(t-1)(2-t)$ on the

interval $[1, 2]$; $b_{12}(t) = \frac{1}{2}(t-2)(3-t)^2$, $b_{23}(t) = -(t-2)(3-t)$, and $b_{33}(t) = t-2$ on the interval $[2, 3]$. Since for this problem

$$X(t_1,t) = \begin{bmatrix} 1 & t_1-t & \frac{1}{2}(t_1-t)^2 \\ 0 & 0 & t_1-t \\ 0 & 0 & 1 \end{bmatrix},$$

it is easily shown that

$$V(2,t) = X(2,t)\,B(t) = \begin{cases} \begin{bmatrix} 1-t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & 0 \le t \le 1 \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3(t-1)(2-t) & 0 \\ 0 & 0 & 0 \end{bmatrix} & 1 \le t \le 2 \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & t-2 \end{bmatrix} & 2 \le t \le 3 \end{cases}$$

Repeating the procedure of the first example, we obtain as the set S_1 a cube centered on the origin of E^3 and with vertices at $(\pm\frac{1}{2}, \pm\frac{1}{2}, \pm\frac{1}{2})$ in E^3. Equations analogous to Eqs. 3.17 can then be analyzed to show that

a) on any face of the cube, the situation is analogous to the first part of the first example--the normal direction is uniquely defined but it does not uniquely specify a control when substituted into Eq. 3.12 of Theorem 3.2, and the optimal control is, in fact, not unique.

b) on any corner of the cube the situation is analogous to the second part of the first example--the normal direction is not uniquely defined, but all normals (except the limiting ones--those which are normal to one of the edges or faces) uniquely specify (to within a set of measure zero) a unique optimal control.

c) on any edge of the cube a new situation arises--the normal direction is not uniquely defined, and the optimal control is not unique. However, even here all the different normals (except the limiting normals - those also normal to one of the faces) are equivalent, in that, upon substitution into Eq. 3.12, they all determine the optimal control to the same degree--i.e., at moments in time at which they determine a specific control, all normals determine the same control, and all leave the control undefined at the same moments.

These examples illustrate that the optimal control is not necessarily unique, that the existence of a unique normal to the set S_1 (or S) is not a sufficient condition for the existence of a unique optimal control, and that the existence of a unique optimal control does not imply that S_1 has a unique normal direction at the given point. However, as we shall see later, all three of these conclusions become invalid for suitably restricted classes of systems.

These examples also pose an interesting possibility, which we note as an aside at this point: We can regard the cost functional $C(\underline{u})$ as imposing an ordering on the controls in the set $U(\underline{a}, \underline{b}, T)$. In these examples the given cost function does not necessarily impose a complete ordering on these controls. We are then free to impose a secondary cost function on the set of controls which are optimal with respect to the original cost function. (In fact, if we are to end up with a specific optimal control, we must apply some rule for deciding which of the optimal controls to use.) For example, we might use a minimum-energy criterion to further order the optimal controls. Alternatively, however, we could use the freedom available to us to satisfy some other objective, such as the avoidance (if possible) of some specified region in state space or the minimization of the excursion into such a region.

Furthermore, it is possible that even this secondary criterion might not provide a complete ordering, so that additional criteria could be applied. This problem (which we designate as a "hierarchical" optimal control problem) can, therefore, be viewed as a sequence of optimization problems each with a more restricted set of admissible control functions, the new admissible set being the set of control functions which are optimal with respect to the previously applied criterion. Whenever this admissible set has only one member, the problem terminates. If, at a particular stage, the admissible set is empty,

an inappropriate or overly restrictive criterion must have been applied previously. This process has certain similarities with the vector-valued cost function approach of Zadeh (Ref. 46), but is not identical in that the various criteria are applied sequentially here, rather than simultaneously.

The fact is, however, that in a significant fraction of the optimal control problems that have been studied, the original cost functional is such that a single optimal control is determined, thus leaving no opportunity for the application of a hierarchical approach. The modification of this approach outlined below might still be useful in engineering problems, nonetheless: In Section 1.2.5.3 we mentioned the difficulties involved in the choosing of a single cost functional which accurately reflects, in the proper proportion, all of the factors that are important in a given problem. The functional actually chosen is often a compromise of some sort. From an engineering standpoint, it thus seems unreasonable that the control which emerges as having the minimum "cost" (in this compromise sense) be regarded as preferable to every other control, including those with "costs" almost as small as the optimal cost. That is, since the cost functional is a guess or a compromise anyway, we are putting too much faith in our mathematics if we insist on using only the control which is optimal this sense. A more reasonable approach would be to look at all the controls which have costs close to the optimal cost (within some reasonable engineering tolerance, say 10%, or whatever figure is appropriate to the problem at hand), and choose among these on the basis of some other appropriate criterion, such as minimum energy, ease of generation by existing devices, minimum frequency bandwidth, etc.

This idea forms the basis for a revised hierarchical optimization technique:

a) Find a control which is optimal with respect to the primary cost functional, (chosen, as best one can, to embody the most important elements of the cost) and determine the appropriate cost.

b) Take as the admissible set of controls all those with costs within some tolerance of the minimum cost.

c) Apply the secondary cost functional to determine an optimal control from this set, and the appropriate cost.

d) Take as the admissible set of controls all those that "survived" step b) and are within some tolerance [not necessarily the same as in part b)] of the minimum cost determined in part c).

e) Apply a third cost functional, and proceed as in the previous steps until the process terminates in the selection of a single control.

The resulting control will be within the chosen tolerence of the minimum cost (with respect to the original cost functional) and yet might be considerably better (in the secondary, tertiary, etc., senses) than the control which is truly optimal in the first sense.

3.6 Proper Systems

The conditions imposed thus far on the types of systems considered have been sufficient to guarantee the existence of an optimal control, and have allowed the establishment of Theorem 3.2, which gives information about the form of the optimal control. But, as we have seen from the examples of the last section, they are not sufficient to guarantee that the optimal control be unique. While it is usually not of prime engineering importance that optimal controls be unique, it is sometimes worthwhile to know whether or not there is a unique optimal control for a given problem. For example, if a certain control $\underline{\bar{u}}(t)$ is known to be optimal, but this control is undesirable for some reason (e.g., difficult to generate, requires wideband equipment, etc.), then we might wish to seek a "better" optimal control if we knew that $\underline{\bar{u}}(t)$ was not the only optimal control. If, on the other hand, $\underline{\bar{u}}(t)$ were known to be the unique optimal control, such efforts would be futile.

Another difficulty with the class of problems considered thus far which is of greater practical importance than the lack of uniqueness of the optimal controls is the fact that when the optimal control is not uniquely defined by Theorem 3.2, we have no simple and convenient method of determining what numerical values of the control to specify. In contrast, when Theorem 3.2 completely specifies the optimal control almost everywhere on T, the problem of finding an optimal control reduces to the relatively straightforward one of finding some vector $\underline{\bar{c}}$ in E^k such that the control determined by Eq. 3.12 satisfies Eq. 3.7.

In order to insure that Theorem 3.2 completely specifies an optimal control, we must make additional restrictions on the class of systems considered. (Such restrictions will also guarantee that the optimal control is unique, a result which is incidental to our purposes here.)

If we restrict attention to systems which satisfy all the conditions of Section 2.1 plus the additional condition that, for all $\underline{\bar{c}} \in E^k$, $\underline{\bar{c}} \neq \underline{0}$, the quantity $\left|V^T(t_1,t)\underline{\bar{c}}\right|$ is different from zero almost everywhere on T, then the difference between the set $T_1(\underline{\bar{c}})$ defined in Theorem 3.2 and the set T will be a set of measure zero. Theorem 3.2 can then be applied to determine an optimal control defined almost everywhere on T for every choice of $\underline{\bar{c}}$. Furthermore, this will be a useful result, since many systems of engineering interest meet the above stated requirement.

This line of reasoning motivates the following definition and theorem:

DEFINITION: A system of the type described in Section 2.1 is proper on the interval T if, for every vector $\underline{\bar{c}} \in E^k$, $\underline{\bar{c}} \neq \underline{0}$, the quantity $\left|V^T(t_1,t)\underline{\bar{c}}\right|$ is different from zero almost everywhere on the interval T.

We note in passing that this definition is analogous to LaSalle's definition (Ref. 59) of a proper system, which can be stated as follows:

A system characterized by an equation of the form of Eq. 2.1, and with A(t) and B(t) continuous matrix functions of time defined for all t, is proper if, for every $\underline{\bar{c}} \in E^n$, $\underline{\bar{c}} \neq \underline{0}$, the quantity $\left|B^T(t)\, X^T(t_1,t)\underline{\bar{c}}\right|$ is not identically zero on any time interval of length greater than zero.

These two definitions are not completely comparable, since the second definition applies to a more restricted class of systems than the first, but for problems to which they both apply, the second is more restrictive than the first in two respects:

a) The second definition involves every possible time interval on the whole time axis, while the first makes restrictions on the system only on the interval T.

b) For $k < n$, the matrix D, which enters into the definition of $V(t_1,t)$, might be such that a given system would not satisfy the second definition, but would satisfy the first. To show this, we have only to write out the two quantities $|B^T(t)\, X^T(t_1,t)\underline{\bar{c}}_n|$ and $|G^{-T}(t)\, B^T(t)\, X^T(t_1,t)\, D^T\underline{\bar{c}}_k|$ that appear in the two definitions. The n-vectors $D^T\underline{\bar{c}}_k$ lie in a k-dimensional subspace of E^n, because D is of rank k, by assumption. If all the vectors $\underline{\bar{c}}_n$ in E^n which caused

$|B^T(t)\, X^T(t_1, t)\, \bar{c}_n|$ to be identically zero on some subinterval of T happened to be orthogonal to this manifold, then the stated result would take place. On the other hand, the second definition cannot be satisfied if the first is not, since every $\bar{c}_k$ in the null space of $G^{-T}(t)\, B^T(t)\, X^T(t_1, t)\, D^T$ corresponds to some $\bar{c}_n$ in the null space of $B^T(t)\, X^T(t_1, t)$, namely $\bar{c}_n = D^T \bar{c}_k$.

Theorem 3.3: Every system of the type described in Section 2.1 which is proper on the interval T has an optimal control which is defined almost everywhere on T by Eq. 3.12 of Theorem 3.2. Furthermore, this control is unique to within a set of measure zero.

Proof: We first dispose of the cases in which $\bar{C} = 0$. In such cases, Eq. 3.13 shows that the optimal control is zero almost everywhere on T, which obviously satisfies this theorem. Therefore, we need to consider only cases for which $\bar{C} > 0$.

For such cases, we first note that the first assertion of the theorem is obviously true, since the condition that the system be proper on the interval T is precisely the condition that Eq. 3.12 determine $\bar{z}(t)$ almost everywhere on T. To prove the uniqueness of this control, we consider any two outward normals $\underline{c}_1$ and $\underline{c}_2$ to the set S at the point $\underline{g}$ and the corresponding optimal controls $\underline{z}_1(t)$ and $\underline{z}_2(t)$. (By Theorem 3.2, every optimal control has some such normal associated with it so that by considering all such normals we are indeed considering all the optimal controls.)

Both $\underline{z}_1(t)$ and $\underline{z}_2(t)$ must satisfy Eq. 3.7, and therefore,

$$\underline{g} = \int_T V(t_1, t)\, \underline{z}_1(t)\, dt$$

$$\underline{g} = \int_T V(t_1, t)\, \underline{z}_2(t)\, dt$$

Subtracting these gives

$$\underline{0} = \int_T V(t_1, t)\, [\underline{z}_1(t) - \underline{z}_2(t)]\, dt$$

Forming the inner product of each side of this equation with $\underline{c}_1$ then yields

$$0 = \int_T \underline{c}_1^T V(t_1,t)\,[\underline{z}_1(t) - \underline{z}_2(t)]\,dt\,.$$

But since

$$\underline{z}_1(t) = \bar{C}\,\frac{V^T(t_1,t)\underline{c}_1}{\left|V^T(t_1,t)\underline{c}_1\right|}$$

almost everywhere on T, from Eq. 3.12, the quantity $\underline{c}_1^T V(t_1,t)$ can be written as

$$\frac{\left|V^T(t_1,t)\underline{c}_1\right|}{\bar{C}}\,\underline{z}_1^T(t)$$

almost everywhere on T. Thus,

$$0 = \int_T \frac{\left|V^T(t_1,t)\underline{c}_1\right|}{\bar{C}}\left[\underline{z}_1^T(t)\,\underline{z}_1(t) - \underline{z}_1^T(t)\,\underline{z}_2(t)\right]dt\,.$$

We note from Eq. 3.12 that $|\underline{z}_1(t)| = \bar{C}$ and $|\underline{z}_2(t)| = \bar{C}$ almost everywhere on T. Using this result and the fact that the inner product $\underline{z}_1^T\underline{z}_2$ of any two vectors $\underline{z}_1$ and $\underline{z}_2$ is no greater than the product of the Euclidean norms of the two vectors, we have that

$$\underline{z}_1^T(t)\,\underline{z}_1(t) - \underline{z}_1^T(t)\,\underline{z}_2(t) \geq \bar{C}^2 - \bar{C}^2 = 0$$

almost everywhere on T. Thus, the integrand of the above integral is never negative (except perhaps on a set of measure zero), and yet the integral is equal to zero. Therefore, the integrand must be zero almost everywhere. Since $\frac{1}{\bar{C}}\left|V^T(t_1,t)\underline{c}_1\right|$ is different from zero almost everywhere on T, it must follow that $\underline{z}_1^T(t)\,\underline{z}_1(t) - \underline{z}_1^T(t)\,\underline{z}_2(t) = 0$ a.e. on T. Exactly similar reasoning with the roles of $\underline{z}_1(t)$ and $\underline{z}_2(t)$ interchanged shows that $\underline{z}_2^T(t)\,\underline{z}_2(t) - \underline{z}_2^T(t)\,\underline{z}_1(t) = 0$ a.e. on T. Adding these two expressions then gives

$$\underline{z}_1^T(t)\,\underline{z}_1(t) - \underline{z}_1^T(t)\,\underline{z}_2(t) - \underline{z}_2^T(t)\,\underline{z}_1(t) + \underline{z}_2^T(t)\,\underline{z}_2(t) = 0 \text{ a.e. on T.}$$

But this is just the square of the Euclidean norm of the vector $\underline{z}_1(t) - \underline{z}_2(t)$. Therefore,

$$|\underline{z}_1(t) - \underline{z}_2(t)| = 0 \text{ a.e. on T,}$$

which implies that $\underline{z}_1(t) = \underline{z}_2(t)$ a. e. on T. Since $\underline{z}_1(t)$ and $\underline{z}_2(t)$ were any two optimal controls, the theorem is proven.

> Lemma: If a given system of the type described in Section 2. 1 is proper on the interval T, then the requirement of Section 2. 1 that the output be completely controllable on the interval T is redundant.

Proof: Consider a system which meets the description of Section 2. 1 with the complete controllability requirement omitted, and suppose that $\left|V^T(t_1,t)\underline{c}\right| \neq 0$ a. e. on T for all $\underline{c} \in E^k$, $\underline{c} \neq \underline{0}$. Then

$$\int_T \underline{c}^T V(t_1,t)\, V^T(t_1,t)\underline{c}\, dt > 0 \quad \text{for all } \underline{c} \in E^k, \ \underline{c} \neq \underline{0}.$$

This is a statement of the fact that the matrix

$$\int_T V(t_1,t)\, V^T(t_1,t)dt$$

is positive definite, which is shown in Appendix D to be a necessary and sufficient condition that the given system output be completely controllable on the interval T. Q. E. D.

In testing a given system to see if it meets the conditions of Theorems 3. 1 and 3. 2, we may thus omit the test for complete controllability on the interval T if the system satisfies the condition used to define systems which are proper on the interval T.

Before proceeding with some examples of proper systems, we note certain implications involved in this approach: Had the original Banach space Z been reflexive and rotund, the existence and uniqueness of the optimal control would have followed at once (see Porter and Williams, Ref. 83, p. 42). In this reference it is shown that it is the rotundity of the space that results in a unique optimal control.

In the problem considered here, the control space is not rotund, and the optimal control is not unique in every case. By restricting attention to proper systems, we obtain the same result as if the space had been rotund, namely, a unique optimal control. This is more than just a chance similarity, as we see from the following considerations: In a rotund space, the unit hypersphere has no "flats," or, in other words, is strictly convex. The image of such a hypersphere under a continuous linear mapping of the type studied

in Section 3.2 is also strictly convex, as is easily shown using the linearity of the mapping and the fact that points on the boundary of the image set can be the images only of points on the boundary of the original hypersphere (as proved in Section 3.2). It is the strict convexity of the image set (the reachable set) which results in a unique optimal control. In the problem considered here, the unit hypersphere in Z has "flats." (As a simple example of this, consider the two controls $z_1(t) = 1$, $0 \leq t \leq 2$ and $z_2(t) = 1$, $0 \leq t \leq 1$; $z_2(t) = 0$, $1 < t \leq 2$. Both have peak amplitude equal to 1, and hence,are on the surface of the unit hypersphere in Z. Any linear combination of these of the form $\theta z_1(t) + (1-\theta) z_2(t)$, $0 \leq \theta \leq 1$, also has peak amplitude equal to 1. Hence,Z is not strictly convex. Examples of this type for vector-valued controls $\underline{z}(t)$ are easily concocted along the same lines.) Examples 1 and 2, Section 3.5, show that without the restriction that the system be proper on the interval T, the set S_1 may also have "flats." With this restriction, however, S_1 is always strictly convex. This can be shown as follows: Theorem 3.2 states that every optimal control has associated with it at least one normal vector. Theorem 3.3 states that for proper systems each such normal uniquely and completely determines an optimal control. Since each input to a linear system of the type considered in this work gives rise to a unique output (Ref. 57), this means that each normal vector can correspond to one and only one point on the surface of S_1. If no two points on the surface of the convex set S_1 in E^k have the same normal, then S_1 must be strictly convex. Thus, be restricting attention to proper systems, we guarantee that the image S_1 of the unit hypersphere in Z is strictly convex.

A further restriction could be made also, to obtain properties analogous to those of spaces having "smooth" unit hyperspheres; that is, hyperspheres with no "corners" (see Porter and Williams, Ref. 83), which guarantees that the reachable set has a unique normal at every point of its boundary. It can be shown that the unit hypersphere for the control space Z used here also has "corners." But a suitably restricted class of systems, which we shall refer to as "smooth" systems, are such that the reachable set has no corners. Figure 3.2 shows examples of the appearance of the reachable set for systems which are proper and/or smooth, or neither. These points are also discussed by Kriendler (Ref. 61). Smooth systems are not investigated in detail here, however, because the distinction between smooth and nonsmooth systems is not of great engineering significance.

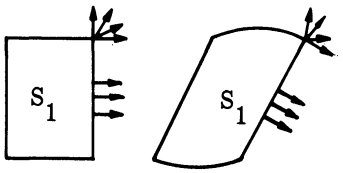

Fig. 3.2(a) Typical sets S_1 for systems which are neither proper (rotund) nor smooth.

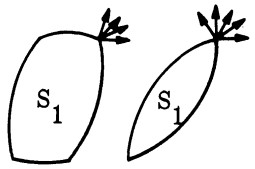

Fig. 3.2(b) Typical sets S_1 for systems which are proper (rotund) but not smooth.

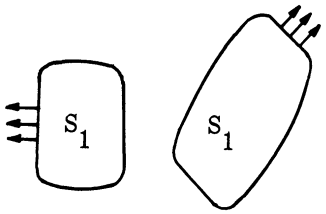

Fig. 3.2(c) Typical sets S_1 for systems which are smooth but not proper.

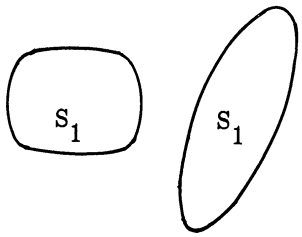

Fig. 3.2(d) Typical sets S_1 for systems which are both smooth and proper (rotund).

3.7 Two Examples Involving Proper Systems

Example 3:

From the set $U(\underline{0},\underline{b},T)$ of bounded measurable scalar functions u(t) defined on the interval $T = [0,1]$ which cause the system (a double integrator) characterized by

$$\dot{\underline{x}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \underline{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

to transfer from the initial state $\underline{x}(0) = \underline{0}$ to the desired final state $\underline{x}(1) = \underline{b}$, choose as an optimal control any one for which

$$C(u) = \operatorname*{ess.\,sup.}_{t \in T} \left| u(t) \right|$$

is minimum with respect to all $u(t) \in U(\underline{0},\underline{b},T)$.

For this system the matrix $X(t_1,t)$ is

$$X(t_1,t) = \begin{bmatrix} 1 & t_1-t \\ 0 & 1 \end{bmatrix}$$

so that $V(t_1,t) = V(1,t) = X(1,t)\,B(t) = \begin{bmatrix} 1-t \\ 1 \end{bmatrix}$. This system is proper on the interval T, since $\left| V^T(t_1,t)\underline{c} \right| = \left| (1-t)c_1 + c_2 \right|$, which, for nonzero $\underline{c}$, can be zero on the interval T only at the isolated point

$$t = 1 + \frac{c_2}{c_1}, \qquad c_1 \neq 0, \qquad -1 \leq \frac{c_2}{c_1} \leq 0 .$$

For this problem the set S_1 defined in the proof of Theorem 3.1 is as shown in Fig. 3.3, from which it can be seen that in this case S_1 is rotund but not smooth. That S_1 is as shown can be verified by substituting vectors $\bar{\underline{c}}$ corresponding to all possible normal directions in Eq. 3.12, using $\bar{C} = 1$ (this gives all points on the boundary of S_1), and making use of the fact that all interior points of S_1 can be obtained for some appropriate choice of $\bar{C} < 1$. The optimal control for a particular goal $\underline{b}$ can be obtained from this plot of the set S_1 as follows: Draw a line from the point $\underline{b}$ to the origin, extending it outward if necessary until it intersects the boundary of S_1. Determine the direction of the normal to S_1 at the point of intersection, and use this result in Eq. 3.12 to determine the optimal control. For more accuracy, an iterative computational algorithm of the type discussed in Chapter VII could be used.

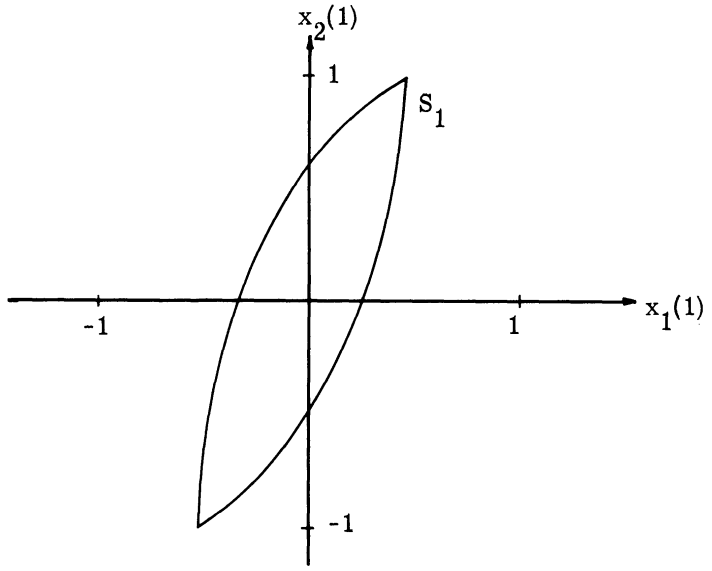

Fig. 3.3 The set S_1 for Example 3.

We note at this point that Example 3, in addition to being a minimum peak amplitude problem in the sense used in this work, is also a member of the class of problems considered by Neustadt (Ref. 77) and others, for which the cost functional was defined as,

$$\max_{i=1,\ldots,r} \sup_{t \in T} \left| u_i(t) \right|$$

where $u_i(t)$ is the ith component of the vector $\underline{u}(t)$. (For scalar $\underline{u}(t)$, this cost functional is identical to the one used in this work.) The next example, on the other hand, is one in which these two cost functionals are not the same, since it involves a vector-valued control.

Example 4:

From the set $U(\underline{0}, \underline{b}, T)$ of bounded measurable controls $\underline{u}(t)$ defined on the interval $T = [0,1]$ which cause the system characterized by

$$\dot{\underline{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \underline{x} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \underline{u}$$

$$\underline{y}(t) = \begin{bmatrix} 1 & 0 & 0 \\ & & \\ 0 & 0 & 1 \end{bmatrix} \underline{x}(t)$$

to transfer from the state $\underline{x}(0) = \underline{0}$ to the desired final condition $\underline{y}(1) = \underline{b}$, choose as an optimal control any one for which $C(\underline{u}) = \operatorname{ess.sup}_{t \in T} \left| \underline{u}(t) \right|$ is a minimum with respect to all $\underline{u}(t) \in U(\underline{0}, \underline{b}, T)$.

For this problem

$$X(t_1, t) = \begin{bmatrix} 1 & t_1 - t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

so that

$$V(t_1, t) = V(1, t) = DX(1, t)\, B(t) = \begin{bmatrix} 1-t & 0 \\ & \\ 0 & 1 \end{bmatrix}.$$

This system is easily shown to be proper[36] on the interval T, so that Theorems 3.1, 3.2, and 3.3 apply. The resulting set S_1, which is both rotund and smooth, is as shown in Fig. 3.4. The points on the boundary of this set are obtained by evaluating Eq. 3.12 for vectors

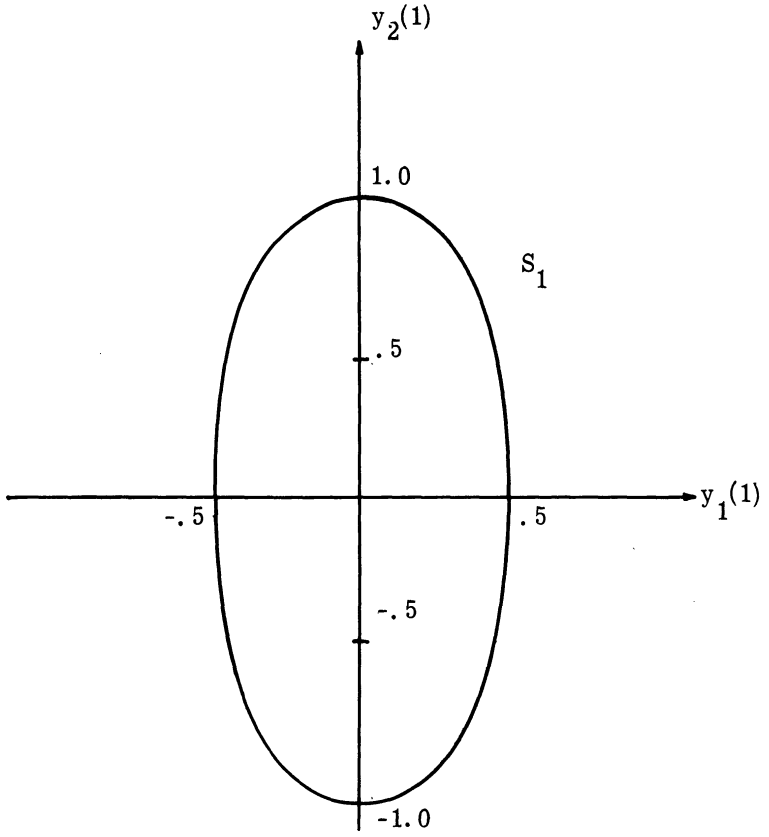

Fig. 3.4 The set S_1 for Example 4.

$\underline{c} \in E^2$ in every possible direction. If we denote by θ the angle that the vector $\underline{c}$ makes with the x_1 axis, then the boundary of the set S_1 is given by

[36] It is, in fact, proper in LaSalle's sense also (Ref. 59).

$$\underline{y}(1) = \begin{cases} \begin{bmatrix} \frac{1}{2}\sec\theta - \frac{1}{2}\,\mathrm{sgn}\,[\cos\theta]\,\tan^2\theta\,\log\dfrac{1+\sec\theta}{\tan\theta} \\ \mathrm{sgn}\,[\sin\theta]\,\tan\theta\,\log\dfrac{1+\sec\theta}{\tan\theta} \end{bmatrix} & -\pi < \theta < \pi \\ \quad \theta \neq 0, \pm\frac{\pi}{2} \\ \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix}, \quad \theta = 0; \quad \begin{bmatrix} 0 \\ \pm 1 \end{bmatrix}, \quad \theta = \pm\frac{\pi}{2}; \quad \begin{bmatrix} -\frac{1}{2} \\ 0 \end{bmatrix}, & \theta = \pi \end{cases}$$

as can be verified by substituting the $\underline{z}(t)$ obtained from Eq. 3.12 into Eq. 3.7 and carrying out the integration. The optimal control for any given goal $\underline{b}$ can be obtained as described in Example 3.

Comparison of Figs. 3.3 and 3.4 shows that, while it may happen that a system which is proper on the interval T is "smooth," i.e., has a set S_1 with uniquely defined normal direction at each point of its boundary (see Fig. 3.4), this is not necessarily true of such systems (see Fig. 3.3).

3.8 Problems With Unspecified Final Time

In this section we consider minimum peak amplitude problems in which the final time t_1 is not specified in advance. That is, we are to choose not only the control $\underline{u}(t)$ but the time interval over which it acts, so as to minimize the peak amplitude of the control. Since the control can be determined from previous results once the optimal value[37] of t_1 is known, the only new aspect of this problem is that finding the optimal t_1.

Two versions of this problem are considered here. In the first version, no restrictions are placed on t_1 except that it be greater than the initial time t_0. In the second version, t_1 is restricted to lie in a closed and bounded[38] interval, i.e., $t_0 < t' \leq t_1 \leq t'' \leq \infty$,

---

[37] The optimal value of t_1 is the one which allows the peak amplitude $\overline{C}$ to take on the lowest possible value, compared to the values of $\overline{C}$ corresponding to any other choice of t_1.

[38] The interval is required to be closed and bounded as one condition of an existence proof, as well as to give a reasonable model of those engineering problems in which unlimited time intervals are not allowed. The existence of solutions is discussed in detail below.

where t' and t'' are given. In both versions, problems involving "moving targets" are included by allowing the desired final condition vector $\underline{b}$ to be a function of t_1. For further generality, the k x n matrix D appearing in the definition of the output vector $\underline{y}$ (see Eq. 2.2) is also allowed to be a function of t_1. So that the results of the previous sections can be applied to this problem, we require, in analogy to Section 2.1, that for each fixed value of t_1 in the allowed range, the k x n matrix D be of rank k and the output $\underline{y}$ of the system be completely controllable on the interval $[t_0, t_1]$.

Under these restrictions, it follows from Theorem 3.1 that there exists at least one minimum peak amplitude control (with unique minimum peak amplitude $\bar{C}$) for each allowed choice of final time t_1. Thus, the minimum peak amplitude $\bar{C}$ can be regarded as a function of t_1. Upon rewriting Eq. 3.13 to show the t_1-dependence explicity, and making use of the fact that the magnitude of $\underline{c}$ does not affect $\bar{C}(t_1)$, so long as $\underline{c} \neq \underline{0}$, we have that

$$\bar{C}(t_1) = \max_{\underline{c} = 1} \frac{\left(\underline{c}, \underline{g}(t_1)\right)}{\int_{t_0}^{t_1} \left| V^T(t_1, t)\underline{c} \right| dt} \tag{3.18}$$

or, from Eq. 3.15,

$$\bar{C}(t_1) = \frac{\left(\bar{\underline{c}}(t_1), \underline{g}(t_1)\right)}{\int_{t_0}^{t_1} \left| V^T(t_1, t)\bar{\underline{c}}(t_1) \right| dt} \tag{3.19}$$

Here $\bar{\underline{c}}$ is an outward normal to the boundary of the reachable set at time t_1, as defined in Theorem 3.2.

Since $\bar{C}(t_1)$ is a <u>function</u> of t_1 (not a functional), the problem of determining the minimum value of $\bar{C}(t_1)$ over all allowed values of t_1 is a problem in the ordinary <u>calculus</u> (not the calculus of variations). As in all such problems, it may happen that $\bar{C}(t_1)$ does not take on a minimum value on the open interval (t_0, ∞). The following simple examples illustrate this point:

Example 5:

Consider the problem of determining the minimum peak amplitude control u(t) which forces the system characterized by

$$\dot{x} = -x + u$$

from the initial state $x(0) = a = 0$ to the final state $x(t_1) = b(t_1) = 1$, where $0 < t_1 < \infty$. One can easily show (Ref. Section 3.1) that $\mathbf{X}(t,s) = e^{-t+s}$, and hence that u(t) must satisfy $g(t_1) = b(t_1) - e^{-t_1} a = \int_0^{t_1} e^{-t_1+s} u(s)\,ds$. From Theorem 3.2

$$u(t) = \bar{C}(t_1) \frac{e^{-t_1+t} \underline{\bar{c}}}{\left| e^{-t_1+t} \underline{\bar{c}} \right|} = \pm \bar{C}(t_1),$$

with the plus sign being used if $g(t_1)$ is positive and the minus sign if $g(t_1)$ is negative. Thus

$$g(t_1) = \pm \bar{C}(t_1) \int_0^{t_1} e^{-t_1+s}\, ds = \pm \bar{C}(t_1)\left(1-e^{-t_1}\right)$$

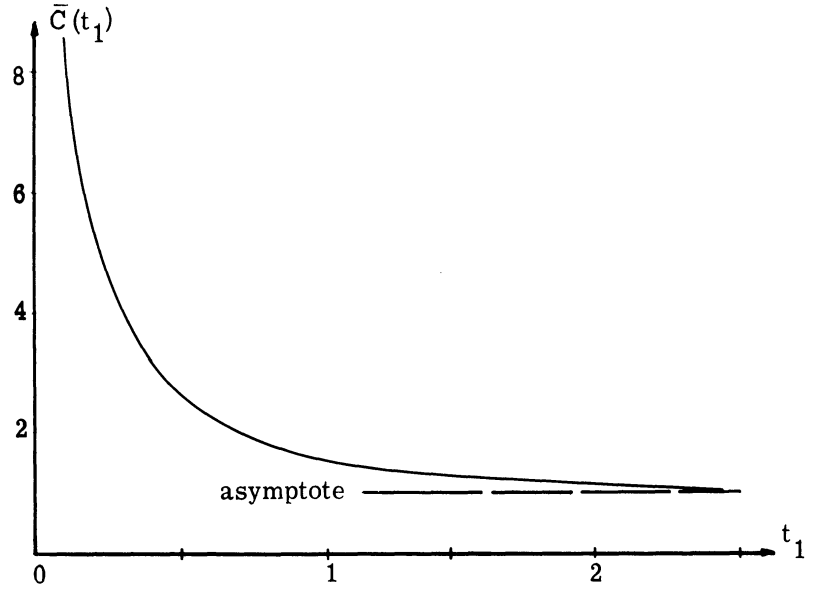

Fig. 3.5 $\bar{C}(t_1)$ vs t_1 for Example 5.

which yields

$$\bar{C}(t_1) = \frac{|g(t_1)|}{1-e^{-t_1}} = \frac{|b(t_1) - e^{-t_1} a|}{1-e^{-t_1}} = \frac{1}{1-e^{-t_1}}$$

This function is plotted in Fig. 3.5, from which we see that there is no value of $t_1 \in (0,\infty)$ which causes $\bar{C}(t_1)$ to be minimum. We note, however, that if t_1 is restricted to lie in some closed and bounded interval $[t',t'']$, then the problem does have a solution, namely $t_1 = t''$, $u(t) = \bar{C}(t'')$, etc.

Example 6:

In this example, we keep the same simple system as in Example 5 but change the boundary condition as follows:

$$x(0) = a = 1 \qquad x(t_1) = b(t_1) = e^{t_1}$$

(Since the final condition varies with t_1, this is a "moving target" problem.) Using the results of the previous example,

$$\bar{C}(t_1) = \frac{|g(t_1)|}{1-e^{-t_1}} = \frac{|b(t_1) - e^{-t_1} a|}{1-e^{-t_1}} = \frac{e^{t_1} - e^{-t_1}}{1-e^{-t_1}} = 1 + e^{t_1}$$

This function is plotted in Fig. 3.6, from which we see that there is no value of $t_1 \in (0,\infty)$ which causes $\bar{C}(t_1)$ to be minimum. As before, however, if t_1 is restricted to lie in some closed interval $[t',t'']$, where $t' > 0$, then the problem has a solution, namely $t_1 = t'$, $u(t) = \bar{C}(t')$, etc. If the interval to which t_1 is restricted is $[0,t'']$, then another interesting result is observed: The minimum cost is zero, and the corresponding final time is $t_1 = 0$, since no control effort is required to force the system from $x(0) = 1$ to $x(t_1) = x(0) = e^0 = 1$. Thus, the curve of $\bar{C}$ versus t_1 is discontinuous at $t_1 = 0$ in this case.

These two problems involve functions $\bar{C}(t_1)$ which are not only continuous in t_1 except at $t_1 = 0$, but are also continuously differentiable except at $t_1 = 0$, and yet they have no solution in the case where t_1 is allowed to lie anywhere in $(0,\infty)$. On the other hand,

it can happen that problems having a very "badly-behaved" $\bar{C}(t_1)$ function may nonetheless possess a solution, even on the open interval $(0,\infty)$. The following example is case in point:

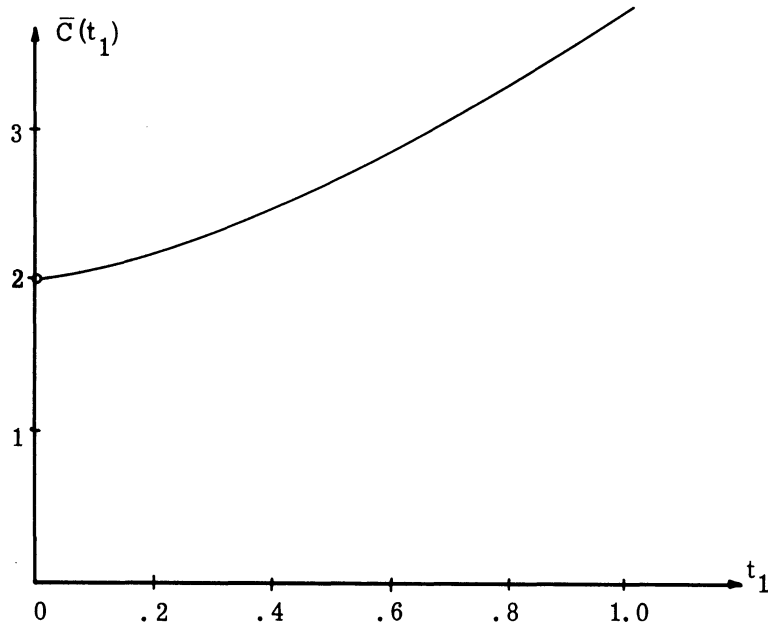

Fig. 3.6 $\bar{C}(t_1)$ vs t_1 for Example 6.

Example 7:

Let the system be the same as in Examples 5 and 6, and let the boundary conditions be

$$x(0) = a = 0 \qquad x(t_1) = b(t_1) = e^{t_1}\left[\operatorname{sgn}\left(\cos\frac{0.125\,\pi}{t_1-0.5}\right) + 2\right]$$

where the signum function sgn (y) is here defined as follows:

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

Proceeding as in the previous examples we get that

$$\bar{C}(t_1) = e^{t_1}\left[\operatorname{sgn}\left(\cos\frac{0.125\,\pi}{t_1-0.5}\right) + 2\right]$$

Investigation of this function for all $t_1 > 0$ shows that it takes on its absolute minimum value $\bar{C} = 4$ at $t_1 = \log_e 2$. Figure 3.7 shows $\bar{C}(t_1)$ plotted as a function of t_1. A slight modification of this problem, namely, one in which the final condition is

$$x(t_1) = b(t_1) = e^{t_1}\left[\operatorname{sgn}\left(\cos\frac{0.125\,\pi}{t_1 - \log_e 2}\right) + 2\right],$$

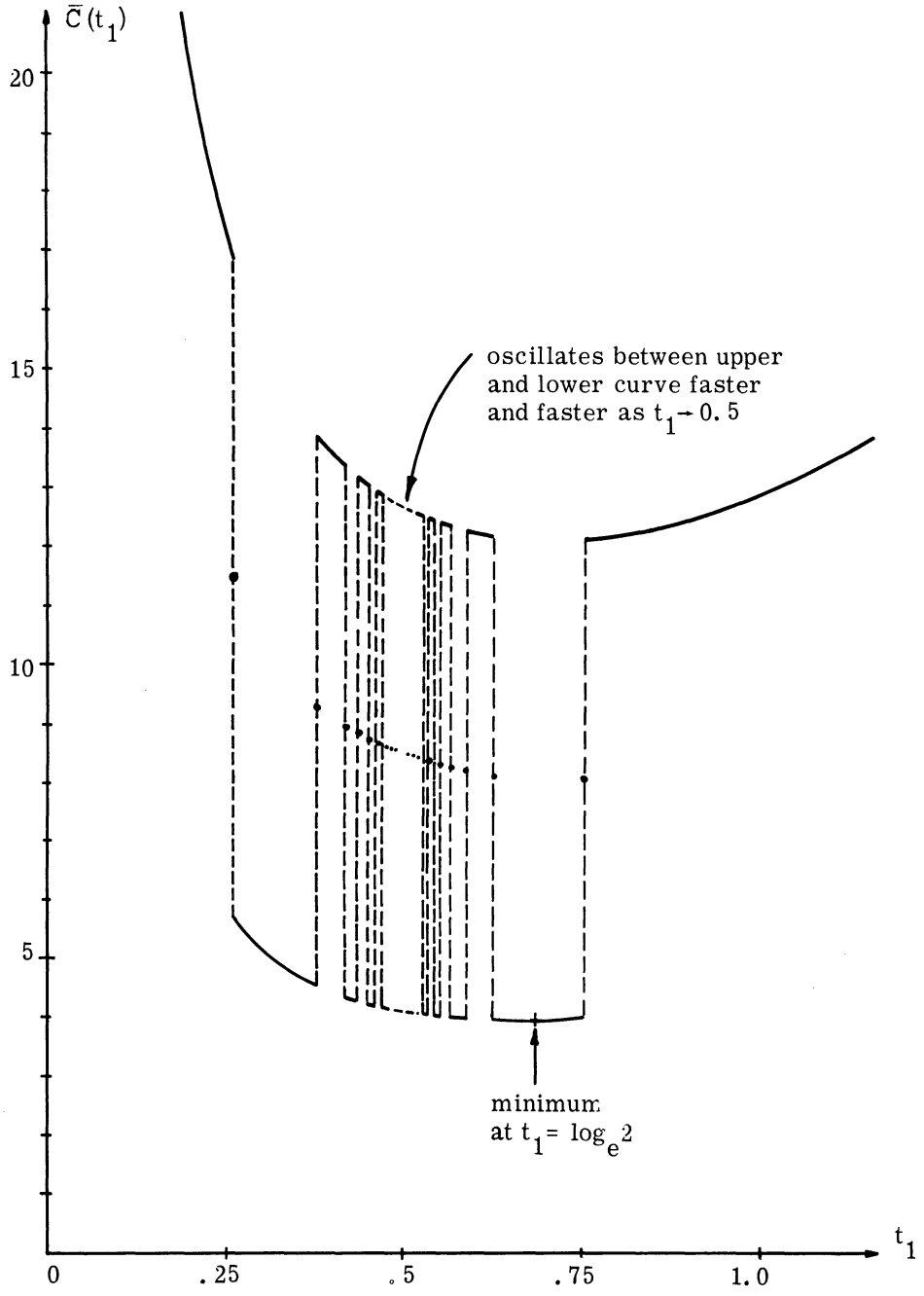

Fig. 3.7 $\bar{C}(t_1)$ vs t_1 for Example 7.

would result in a problem which had no solution on any interval, open or closed, which included $t_1 = \log_e 2$ as one of its points.

These three examples, taken together, show that the question of the existence of nonexistence of solutions to unspecified final time minimum peak amplitude problems is not a simple one, and that the rules obtained for fixed final time problems do no apply. However, we can state the following lemma:

> Lemma: For problems satisfying the conditions of Section 2.1 at each t_1 in the allowed interval $[t',t'']$, $t_0 < t' < t'' < \infty$ and in which $\bar{C}(t_1)$ is continuous in t_1 on this interval, the unspecified final time problem has at least one solution.

Proof: This lemma is merely a statement of the fact that a continuous real function defined on a closed interval takes on its infimum. One way to prove this is to note that a closed bounded interval on the real line is a compact metric space (with the distance between two points t_a and t_b being taken as $|t_a - t_b|$). Then, from Theorems 4.15 and 4.16 of Rudin (Ref. 82, p. 77), the image of this space under the continuous mapping assumed to exist between t_1 and $\bar{C}(t_1)$ is also a compact metric space, and therefore $\bar{C}(t_1)$ takes on its infimum. Q.E.D.

The following lemma gives a sufficient condition that $C(t_1)$ be continuous in t_1:

> Lemma: $\bar{C}(t_1)$ is continuous in t_1 on the interval $[t',t'']$ if $\underline{b}(t_1)$ and $D(t_1)$, the vector and matrix which define the final manifold, are continuous in t_1 on the same interval.

Proof: We note that if $\underline{b}(t_1)$ and $D(t_1)$ are continuous in t_1, then $\underline{g}(t_1) = \underline{b}(t_1) - D(t_1)\, X(t_1,t_0)\, \underline{a}$ and $V(t_1,t) = D(t_1)\, X(t_1,t)\, B(t)\, G^{-1}(t)$ are also, since $X(t_1,t)$ is absolutely continuous (and hence, continuous) in t_1 (see Coddington and Levinson, Ref. 57, pp. 42-43). The quantity

$$\frac{(\underline{c},\, \underline{g}(t_1))}{\int_{t_0}^{t_1} \left| V^T(t_1,t)\, \underline{c} \right| dt}$$

appearing in Eq. 3.18 [which we denote by $F(\underline{c},t_1)$] is continuous in t_1 on the interval in question, since it is the ratio of continuous functions and the denominator is bounded away from zero (uniformly with respect to vectors $\underline{c}$ satisfying the contraint $|\underline{c}| = 1$) for $t_1 \geq t' > t_0$ due to the complete controllability assumption of Section 2.1. Note also that $F(\underline{c},t_1)$ is independent of the magnitude of $\underline{c}$, for all $\underline{c} \neq \underline{0}$. Thus, for every t_1 in $[t',t'']$, every $\underline{c} \neq \underline{0}$, and every $\epsilon > 0$ there exists a $\delta(t_1) > 0$ independent of $\underline{c}$ such that $\left|F(\underline{c},t_1+\Delta t_1) - F(\underline{c},t)\right| < \epsilon$ for $\Delta t_1 < \delta(t_1)$.

From this it follows that for $|\Delta t_1| < \delta(t_1)$,

$$\left| \max_{|\underline{c}_1| = 1} F(\underline{c}_1,t_1+\Delta t_1) - \max_{|\underline{c}_2| = 1} F(\underline{c}_2,t_1) \right| < \epsilon .$$

But since

$$\left|\bar{C}(t_1+\Delta t_1) - \bar{C}(t_1)\right| = \left| \max_{|\underline{c}_1| = 1} F(\underline{c}_1,t_1+\Delta t_1) - \max_{|\underline{c}_2| = 1} F(\underline{c}_2,t_1) \right|$$

from Eq. 3.18 and from the definition of $F(\underline{c},t_1)$, it is clear that $\bar{C}(t_1)$ is continuous in t_1.

Q.E.D.

As a final point in this section, we investigate the possibility of determining the minimum of $\bar{C}(t_1)$ by differentiating $\bar{C}(t_1)$ with respect to t_1 and setting this derivative equal to zero. We assume that $\bar{C}(t_1)$ is piecewise continuous and has a piecewise-continuous derivative with respect to t_1. (In order to be able to carry out the required operations, we must also assume that $\underline{\bar{c}}(t_1)$, the normal to the reachable set which appears in Eq. 3.19, satisfies the same continuity and differentiability conditions.) In order to locate the minimum of $\bar{C}(t_1)$, if it exists, we must then investigate the stationary points and points of discontinuity of $\frac{d\bar{C}(t_1)}{dt}$, and, in problems involving t_1 restricted to a closed interval, the end points of this interval.

The derivative of $\bar{C}(t_1)$ with respect to t_1 can be computed from Eq. 3.19. In this derivation we make use of the facts that $V(t_1,t) = D(t_1)\, X(t_1,t)\, B(t)\, G^{-1}(t)$, $\underline{g}(t_1) = \underline{b}(t_1) - D(t_1)\, X(t_1,t_0)\underline{a}$, and $\frac{d}{dt_1} X(t_1,t) = A(t_1)\, X(t_1,t)$. To simplify the notation, explicit mention of the functional dependencies of $\bar{C}(t_1)$, $\underline{\bar{c}}(t_1)$, $\underline{g}(t_1)$, $V(t_1,t)$, $D(t_1)$, $X(t_1,t)$, $A(t_1)$, and $B(t)$ will be omitted in this derivation unless needed to avoid confusion:

$$\frac{d\bar{C}}{dt_1} = \frac{\left(\underline{c}, \frac{d\underline{g}}{dt_1}\right) + \left(\frac{d\underline{\bar{c}}}{dt_1}, \underline{g}\right)}{\int_{t_0}^{t_1} \left| V^T \underline{\bar{c}} \right| dt}$$

$$- \frac{(\underline{\bar{c}}, \underline{g})}{\left[\int_{t_0}^{t_1} \left| V^T \underline{\bar{c}} \right| dt\right]^2} \left\{ \left| V^T(t_1, t_1)\underline{\bar{c}} \right| + \int_{t_0}^{t_1} \frac{1}{2} \frac{2 \frac{d\underline{\bar{c}}^T}{dt_1} VV^T \underline{\bar{c}} + 2\underline{\bar{c}}^T \frac{dV}{dt_1} V^T \underline{\bar{c}}}{\left| V^T \underline{\bar{c}} \right|} dt \right\}$$

$$\frac{d\bar{C}}{dt_1} = \frac{1}{\int_{t_0}^{t_1} \left| V^T \underline{\bar{c}} \right| dt} \left\{ \begin{array}{l} \frac{d\underline{\bar{c}}^T}{dt_1} \left[\underline{g} - \bar{C} \int_{t_0}^{t_1} V \frac{V^T \underline{\bar{c}}}{\left| V^T \underline{\bar{c}} \right|} dt \right] - \bar{C} \left| V^T(t_1, t_1)\underline{\bar{c}} \right| \\ + \underline{\bar{c}}^T \frac{d\underline{b}}{dt_1} - \underline{\bar{c}}^T \frac{dD}{dt_1} X(t_1, t_0)\underline{a} - \underline{\bar{c}}^T DAX(t_1, t_0)\underline{a} \\ - \bar{C}\, \underline{\bar{c}}^T \frac{dD}{dt_1} \int_{t_0}^{t_1} XBG^{-1} \frac{V^T \underline{\bar{c}}}{\left| V^T \underline{\bar{c}} \right|} dt - \bar{C}\, \underline{\bar{c}}^T DA \int_{t_0}^{t_1} XBG^{-1} \frac{V^T \underline{\bar{c}}}{\left| V^T \underline{\bar{c}} \right|} dt \end{array} \right.$$

We note that the term multiplying $\frac{d\underline{\bar{c}}^T}{dt_1}$ is identically zero, due to the satisfaction of the final condition. Further collecting of terms, plus substitution of Eq. 3.1 (evaluated at $t = t_1$) and use of the $\underline{u}(t)$ given by Theorem 3.2 in two places, gives

$$\frac{d\bar{C}}{dt_1} = \frac{1}{\int_{t_0}^{t_1} V^T \underline{\bar{c}}\, dt} \left\{ \begin{array}{l} - \bar{C} \left| V^T(t_1, t_1)\underline{\bar{c}} \right| - \left(\underline{\bar{c}}, \frac{dD}{dt_1} \underline{x}(t_1) - \frac{d\underline{b}}{dt_1} \right) \\ - \underline{\bar{c}}^T DA\underline{x}(t_1) \end{array} \right\} \quad (3.20)$$

The stationary points of $\bar{C}(t_1)$ occur when this quantity is equal to zero. Thus, $\bar{C}(t_1)$ will have a stationary point at a particular value of t_1 if the following condition is satisfied:

$$\underline{\bar{c}}^T D(t_1) A(t_1) \underline{x}(t_1) + \bar{C}(t_1) \left| V^T(t_1, t_1)\underline{\bar{c}} \right| + \left(\underline{\bar{c}}, \frac{dD(t_1)}{dt_1} \underline{x}(t_1) - \frac{d\underline{b}(t_1)}{dt_1} \right) = 0 \quad (3.21)$$

Unfortunately, this condition is of limited usefulness, since there is no way to tell in advance whether or where $\bar{C}(t_1)$ will have a discontinuous derivative. As noted above, the minimum value of $\bar{C}(t_1)$ may occur at just such a point. The following example, involving a very simple system description and constant boundary conditions, nonetheless gives rise to a $\bar{C}(t_1)$ with a discontinuous derivative. The same thing can happen in higher-order systems also.

Example 8:

Let the system be the same as in Examples 5, 6, and 7, and let the boundary conditions be $\underline{x}(0) = a = 1$, $\underline{x}(t_1) = b(t_1) = \frac{1}{2}$. Then, following the same procedure as before, we have that $g(t_1) = \frac{1}{2} - e^{-t_1}$ and hence that

$$\bar{C}(t_1) = \frac{\left| \frac{1}{2} - e^{-t_1} \right|}{1 - e^{-t_1}}$$

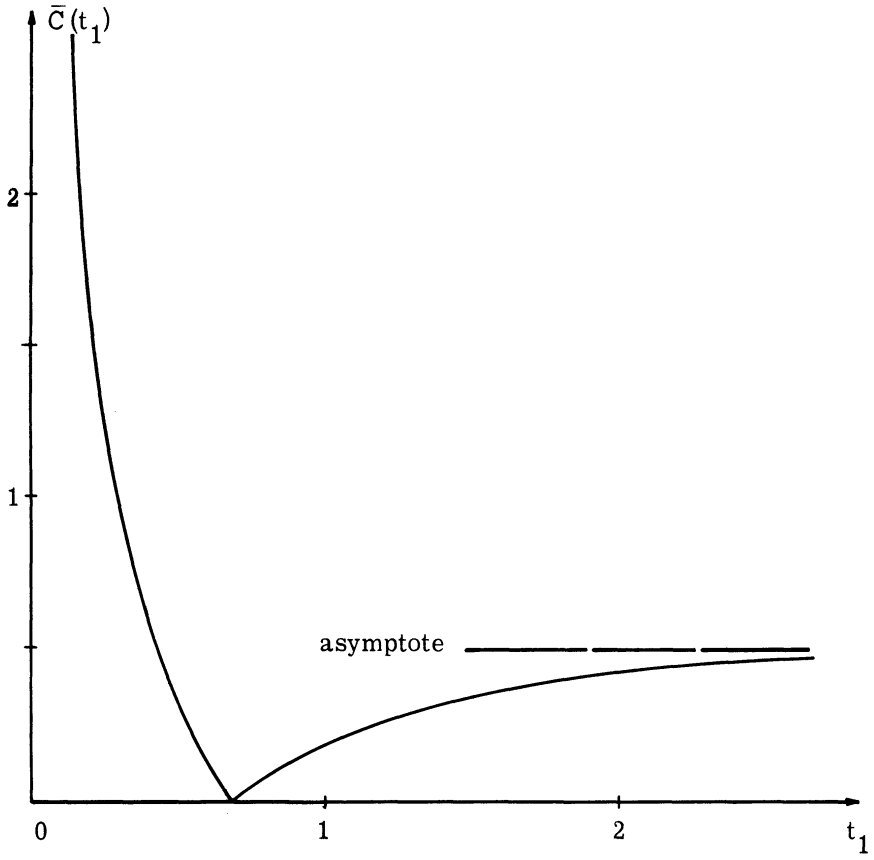

Fig. 3.8 $\bar{C}(t_1)$ vs t_1 for Example 8, showing discontinuous derivative at $t_1 = \log_e 2$.

This function is plotted in Fig. 3.8, from which we see that the optimal solution is $t_1 = \log_e 2$, $\bar{C}(t_1) = 0$, $u(t) = 0$, $x(t) = e^{-t}$. However, it is clear that $\bar{C}(t_1)$ does not possess a derivative at $t_1 = \log_e 2$, and that this is thus obviously not a stationary point of $\bar{C}(t_1)$.

One conclusion that can be drawn from the discussion and examples of this section is as follows:

Unspecified final time minimum peak amplitude problems may exhibit behavior which is difficult to analyze or predict, even in the simplest cases. Therefore, a good engineering approach to such problems uses a combination of graphical and analytic techniques--graphical techniques to determine the range or ranges of t_1 in which minima should be sought, followed by analytical techniques (where applicable) to locate the minima exactly. This is in fact the procedure that was used in Example 7 to find the minimum at $t_1 = \log_e 2$.

3.9 Optimal Control Laws

As defined earlier, an optimal control law is a rule which states what control function should be applied to the controlled system as a function of the state of the system, so that the resulting trajectory will be optimal. This law sometimes takes a very simple form, as in the by-now classical "bang-bang" second-order servo problem (Ref. 13), in which the phase plane is divided into two regions, one in which the optimal control always takes on the maximum allowed value, and the other in which it takes on the minimum allowed value. More complicated examples have also been worked out, involving more than two regions, but still preserving the simple rule that every state in that region is associated with the same value of the control (see Ref. 13, pp. 23-45).

In the problems considered here, the control law generally takes a more complicated form than this. Theorem 3.2 states that at those moments at which a minimum peak amplitude control is uniquely defined, it is of the form

$$\underline{\bar{z}}(t) = G(t)\,\underline{\bar{u}}(t) = \bar{C}\,\frac{V^T(t_1,t)\underline{\bar{c}}}{\left|V^T(t_1,t)\underline{\bar{c}}\right|} \tag{3.22}$$

where $\bar{C}$ and $\underline{\bar{c}}$ are determined from the boundary condition equation

$$\underline{b} - DX(t_1,t_0)\underline{a} = \overline{C} \int_{t_0}^{t_1} V(t_1,t) \frac{V^T(t_1,t)\underline{\overline{c}}}{\left|V^T(t_1,t)\underline{\overline{c}}\right|} \, dt$$

We can also write this in terms of the state $\underline{x}(t)$ at time t, thus:

$$\underline{b} - DX(t_1,t)\,\underline{x}(t) = \overline{C} \int_{t}^{t_1} V(t_1,s) \frac{V^T(t_1,s)\underline{\overline{c}}}{\left|V^T(t_1,s)\underline{\overline{c}}\right|} \, ds \qquad (3.23)$$

Since this expression does not involve the initial state $\underline{a}$ or initial time t_0, we can in principle construct a device which determines the optimal control at each moment as a function of the state $\underline{x}(t)$ at that moment, plus of course the desired goal state $\underline{b}$ and final time t_1; such a device must determine the state $\underline{x}(t)$ at time t, then solve Eq. 3.23 to determine $\overline{C}$ and $\underline{\overline{c}}$, and then use these values in Eq. 3.22 to determine $\underline{\overline{u}}(t)$. This is not as formidable a problem as it may appear to be at first, since if the given equations were a perfect model of the physical system and if the initial computation were errorless, $\overline{C}$ and $\underline{\overline{c}}$ would remain the same throughout the whole trajectory. The effect of imperfect modeling of the system, unforseeable external disturbances, and computation errors is to require, in practice, that $\overline{C}$ and $\underline{\overline{c}}$ be corrected during the time interval of the problem, but if these errors and disturbances are typically small, infrequent correction would normally suffice, so that the computation of $\overline{C}$ and $\underline{\overline{c}}$ need not necessarily be carried out in real time.

If Eq. 3.22 does not uniquely specify the optimal control at every moment of time, [i.e., if $V^T(t_1,t)\underline{\overline{c}}$ is zero on some nontrivial subset of the interval], complications are introduced, in that some other means must be provided for computing the control at such times, but, in principle, the above discussion still applies.

Another way of specifying the optimal control law is possible, and is especially convenient if the system is proper, which implies that the quantities $\overline{C}$ and $\underline{\overline{c}}$ completely determine the optimal control. This other way amounts to a tabulation of the optimal control for each point in state space; that is, we consider the (k+1)-dimensional Euclidean space

having k of its coordinate axes associated with the k components of the goal state[39] $\underline{g}(t_1,t) = \underline{b} - DX(t_1,t)\,\underline{x}(t)$ and the (k+1)st axis associated with the time remaining, $t_1 - t$, and tabulate, for each point in the half space for which $t_1 - t$ is positive, the numbers $\bar{C}$, $\underline{\bar{c}}$ appearing in the expression for the optimal control (Eq. 3.22) corresponding to that goal and amount of time remaining. Such a tabulation is impossible (since an infinite number of points are involved), but two simplifications allow the same end to be accomplished more reasonably:

a) Only the values of $\bar{C}$ and $\underline{\bar{c}}$ for unit vectors $\underline{g}(t_1,t)/|\underline{g}(t_1,t)|$ need be tabulated, since the control for other values of $\underline{g}(t_1,t)$ can be obtained simply by multiplying by $|\underline{g}(t_1,t)|$ the value of $\bar{C}$ corresponding to a unit vector in the desired direction.

b) In the usual engineering problem, some tolerance is allowed, both with respect to the satisfaction of the desired final condition and with respect to the achieving of minimum cost. Thus, adjacent points in the above-mentioned (k+1)-dimensional space could be grouped together in regions and treated as equivalent, thus reducing to a finite number the number of values of $\bar{C}$ and $\underline{\bar{c}}$ that must be tabulated. Approaches other than straight tabulation are possible, also, such as a tabulation of level curves or (in two or three dimensional problems), the storing of the information as X-Y plots or 3-D cams.

In any case, the problem of tabulating a vector function of time for each $\underline{g}(t_1,t)$ has been reduced to the storing of k + 1 numbers $\bar{C}$, $\bar{c}_1,\ldots,\bar{c}_k$ (actually k numbers, since $\underline{\bar{c}}$ is a unit vector) for each $\underline{g}(t_1,t)$.

---

[39] Really, the amount of the goal that remains unaccomplished.

CHAPTER IV

THE INVERSE PROBLEM

4.1 A Time-Optimal Problem

Associated with each minimum peak amplitude problem there is a certain minimum time problem.[40] The following discussion of this minimum time problem offers additional insight into the original problem, and provides an alternate approach to the solution of that problem. It will be shown that under certain conditions the two problems are, in a sense, inverse to each other.

We begin by posing the time-optimal problem:

In the class of bounded measurable controls $\underline{u}(t)$ which cause the system

$$\dot{\underline{x}} = A(t)\,\underline{x}(t) + B(t)\,\underline{u}(t)$$

$$\underline{y}(t) = D(t)\,\underline{x}(t)$$

to satisfy the boundary conditions

$$\underline{x}(t_0) = \underline{a}\,;\ \underline{y}(t_1) = \underline{b}(t_1)$$

at some unspecified time t_1 and for which $|G(t)\,\underline{u}(t)| \leq C$ for all $t \in T$, find one for which t_1 is a minimum. A, B, D, G, $\underline{x}$, $\underline{y}$, and $\underline{u}$ are as specified in Sec. 2.1. C is now a given constant.

4.2 The Solution to the Time-Optimal Problem

We solve this problem with the aid of Pontryagin's maximum principle, which states that

---

[40] Krasovskii (Ref. 78) has considered a special case of this problem, approaching it from the functional analysis viewpoint.

a necessary condition that the control $\underline{u}(t)$ and the corresponding $\underline{x}(t)$ and $\underline{y}(t)$ be optimal is that there exist a constant vector $\underline{c}$, a non-positive constant ψ_o, and an absolutely continuous nonzero vector $\underline{\psi}(t)$ satisfying

$$\left.\begin{aligned} \dot{\psi}_i &= -\frac{\partial H(t)}{\partial x_i} \qquad t \in T \\ \psi_i(t_1) &= \left.\frac{\partial Q}{\partial x_i(t_1)}\right|_{t_1 = \text{const.}} \end{aligned}\right\} \; i = 1, \ldots, n \tag{4.1}$$

$$H(t_1) + \left.\frac{\partial Q}{\partial t_1}\right|_{\underline{x}(t_1) = \text{const.}} = 0 \tag{4.2}$$

such that at every moment of time $t \in T$, $\underline{u}(t)$ maximizes $H(t)$, (for fixed $\underline{x}$ and $\underline{\psi}$) over all controls for which $|G(t)\,\underline{u}(t)| \leq C$, where

$$H(t) = \psi_o + (\underline{\psi}(t),\, A(t)\underline{x}(t) + B(t)\,\underline{u}(t)) \tag{4.3}$$

$$Q = (\underline{c},\, \underline{y}(t_1) - \underline{b}(t_1)) = (\underline{c},\, D(t_1)\underline{x}(t_1) - \underline{b}(t_1)) \tag{4.4}$$

Since only the last term of Eq. 4.3 depends on $\underline{u}(t)$, the required maximization of $H(t)$ is carried out by maximizing the quantity $\underline{\psi}^T(t)\, B(t)\, \underline{u}(t)$ subject to the constraint $|G(t)\,\underline{u}(t)| \leq C$. Rewriting this quantity as $\underline{\psi}^T(t)\, B(t)\, G^{-1}(t)\, G(t)\, \underline{u}(t)$ reveals that whenever $G^{-T}(t)\, B^T(t)\, \underline{\psi}(t) \neq \underline{0}$, $\underline{u}(t)$ must be chosen so that $G(t)\,\underline{u}(t)$ is colinear with $G^{-T}(t)\, B^T(t)\, \underline{\psi}(t)$ and of the maximum allowed length. Thus

$$G(t)\,\underline{u}(t) = C\,\frac{G^{-T}(t)\, B^T(t)\, \underline{\psi}(t)}{|G^{-T}(t)\, B^T(t)\, \underline{\psi}(t)|} \tag{4.5}$$

whenever $G^{-T}(t)\, B^T(t)\, \underline{\psi}(t)$ is not equal to zero. At those times at which this quantity is equal to zero, the maximum principle cannot be used to determine $\underline{u}(t)$.

Carrying out the indicated operations in Eq. 4.1 gives $\dot{\underline{\psi}} = -A^T(t)\underline{\psi}$, $\underline{\psi}(t_1) = D^T(t_1)\,\underline{c}$, which has as its solution

$$\underline{\psi}(t) = X^T(t_1, t)\, D^T(t_1)\,\underline{c} \tag{4.6}$$

as can be verified by direct substitution and the use of the identity $\mathbf{X}(t_1, t)\,\mathbf{X}(t, t_1) = I$. Here $\mathbf{X}(t_1, t)$ is the state transition matrix defined in Section 3.1 in connection with the solution of the original problem. Substitution of this result into Eq. 4.5 gives

$$G(t)\,\underline{u}(t) = C\,\frac{V^T(t_1, t)\,\underline{c}}{|V^T(t_1, t)\,\underline{c}|} \qquad \text{whenever} \qquad |V^T(t_1, t)\,\underline{c}| \neq 0$$

$$|G(t)\,\underline{u}(t)| \leq C \quad \text{but otherwise undetermined whenever } |V^T(t_1, t)\,\underline{c}| = 0 \tag{4.7}$$

where $V(t_1, t)$ is as defined by Eq. 3.4. We note that this result is identical in form to the optimal control for the original problem, as given by Theorem 3.4, except that the unknown $\bar{C}$ has been replaced by the given value C, and t_1 is now considered as a variable, rather than a given constant. Since the boundary conditions for the two problems are also the same for each t_1, we might expect the two problems to be closely related. These considerations motivate the following definition:

4.3 The Inverse Relationship

Definition: A minimum peak amplitude problem and a minimum time problem are said to be in inverse relationship to each other (or, alternatively, to be inverses of each other) if they involve the same system and boundary conditions and if the control which is optimal for the minimum peak amplitude problem (for a certain t_1 and resulting peak amplitude $\bar{C}$) is also optimal (with minimum time t_1) for the minimum time problem in which $\bar{C}$ is chosen as the bound on the amplitude of $G(t)\,\underline{u}(t)$. That is, in the two problems the roles of cost functional and constraint on $\underline{u}$ are interchanged, the cost functional of one being the constraint on the other, and vice versa. We ignore controls which differ from the optimal control only on a set of measure zero.

Three questions now arise:

a) Are there any pairs of problems which are inverse to each other? The answer is yes. Every time-optimal problem involving a proper system and a

magnitude constraint on the Euclidean norm of G(t) $\underline{u}(t)$ which has a solution[41] is inverse to a minimum peak amplitude problem. This follows[42] from the identity of form of the solutions to the two kinds of problems, and from the uniqueness of the solution to the fixed final time minimum peak amplitude problem for systems which are proper on the interval T.

b) Are all minimum peak amplitude problems inverse to some minimum time problem; i. e., are the two classes of problems completely equivalent? The answer here is no. To illustrate this, suppose that we have solved a minimum time problem and obtained three distinct solutions (with corresponding final times t_{1a}, t_{1b}, and t_{1c}, $t_{1a} < t_{1b} < t_{1c}$) all of which satisfy all of the given necessary conditions for time optimality. Obviously, only t_{1a} is truly a time optimal solution, but, by reasoning identical to that above, all three correspond to minimum peak amplitude problems. Thus, the minimum peak amplitude problems corresponding to t_{1b} and t_{1c} do not possess minimum time inverses.

c) Since all minimum peak amplitude problems do not have minimum time inverses, can we find conditions which tell us whether or not a given minimum peak amplitude problem (for a specific choice of t_1) is in inverse relationship to some minimum time problem? In answer to this, we shall find a general <u>necessary</u> condition for the existence of the inverse relationship.

However, a general set of necessary and sufficient conditions which apply for all boundary conditions is not likely to be obtained.

---

41 Many such problems have no solution, since C may be too small. For instance, Example 5 of Sec. 3.8 shows a system for which there are no minimum-time solutions if C is chosen less than unity. See Fig. 3.4.

42 To illustrate this reasoning in more detail, assume that we have solved the minimum time problem with specified $\overline{C}$, and obtained a minimum time t_1. If we now consider a minimum peak amplitude problem with this same t_1 chosen as the final time, we know that there is a unique optimal solution. Since the control for the minimum time problem satisfies all the conditions of this problem, it must be the unique solution to the minimum peak amplitude problem. Thus, every minimum time solution is also the optimal solution to a minimum peak amplitude problem, as claimed.

To facilitate discussion of these points, we introduce for each minimum peak amplitude problem to be considered a cost-vs-final time graph of the type shown in Fig. 4. 1a). We imagine that we have solved the minimum peak amplitude problem for the given system and boundary conditions and for a series of choices of final time t_1. Along the abscissa of the graph we plot the final time t_1, and along the ordinate the corresponding values of cost $\overline{C}$.

Such a graph is used in minimum peak amplitude problems as follows: For any given t_1 we can read off at once the corresponding optimal cost $\overline{C}$. If t_1 is unspecified, we can select from the graph the point (or points) for which $\overline{C}$ takes on its absolute minimum value, or, alternatively, we can see that $\overline{C}$ has no absolute minimum value for finite t_1. Referring to Fig. 4. 1a), we note in passing that for this example the times t_{1c} and t_{1e} correspond to local minima and maxima of $\overline{C}$, respectively, and hence that both satisfy the zero-slope condition used for the determination of t_1 (see, Eq. 3. 21), even though neither corresponds to an absolute minimum of $\overline{C}$.

This graph also gives information about minimum time problems, for which it is interpreted as follows:

For any given bound $\overline{C}$ on $|G(t)\,\underline{u}(t)|$, the corresponding minimum time (if one exists) is determined by entering the graph at that point $\overline{C}$ on the ordinate and determining the minimum abscissa which gives that value of $\overline{C}$. For example, referring to Fig. 4. 1a), a problem for which $|G(t)\,\underline{u}(t)|$ must be less than or equal to the specified value $\overline{C}_b$, has a minimum time solution with final time t_{1b}. Solutions which require final times t_{1d} or t_{1f} are <u>not</u> time-optimal solutions.

This graph also clearly illustrates the above statement that all minimum peak amplitude problems do not have minimum time inverses. Minimum peak amplitude problems with times between t_{1c} and t_{1g}, in this example, do not possess minimum time inverses.

Fig. 4. 1b) provides a further clarification of these points. Every point on this curve corresponds to a solution of the inverse (time-optimal) problem satisfying all the conditions of Pontryagin's maximum principle. The gap in the curve between

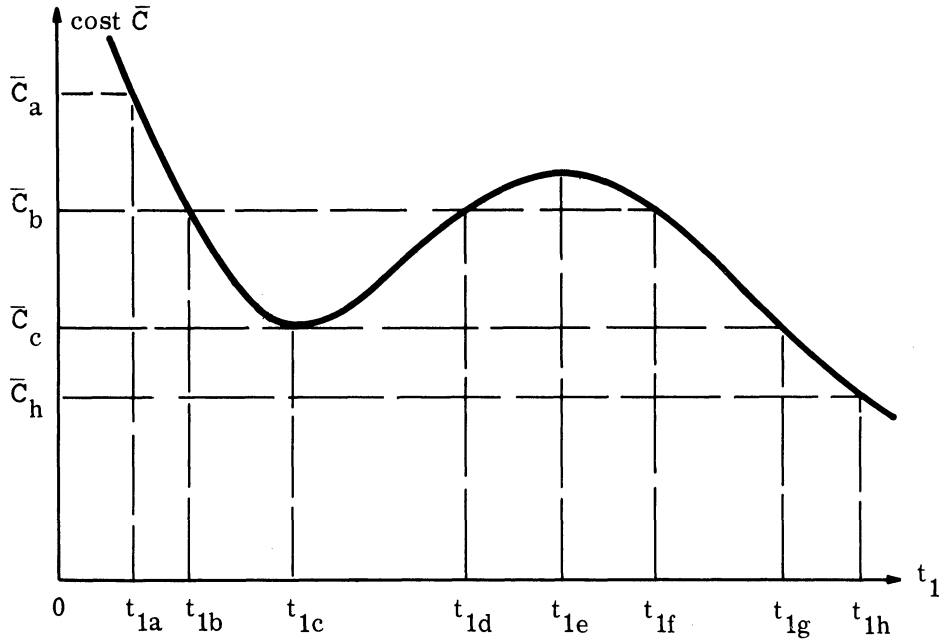

Fig. 4. 1(a). A typical curve of cost vs. final time for a minimum peak amplitude problem.

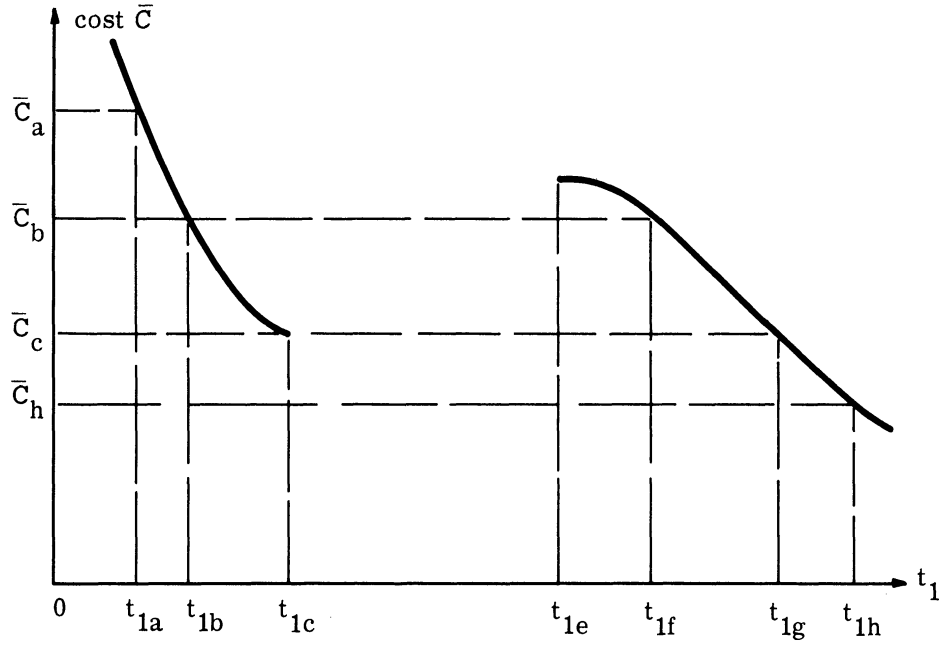

Fig. 4. 1(b). $\bar{C}$ vs. t_1 for the inverse problem, showing only those points which satisfy all the conditions of the maximum principle.

t_{1c} and t_{1e} implies that there are no trajectories involving these final times which satisfy the conditions of the maximum principle. (It turns out that the trajectories in Fig. 4.1a) corresponding to these times satisfy all the conditions of the maximum principle except the one that ψ_o be non-positive.) It is obvious that the trajectories corresponding to final times between t_{1e} and t_{1g} are not time optimal, but this is not a contradiction of the maximum principle, since it gives only necessary conditions.

4.4 A Necessary Condition for the Existence of the Inverse Relationship

From the above standpoint, one immediately sees that a necessary condition that a given minimum peak amplitude problem with specified t_1 have a minimum-time inverse is that the slope of the curve of $\bar{C}$ - vs - t_1 be non-positive at that value of t_1, assuming that the slope is defined there. That this is not sufficient is attested to by the problems with $t_{1e} < t_1 \leq t_{1g}$ in Fig. 4.1a). In Section 3.8 the slope of $\bar{C}$ was evaluated as

$$\frac{d\bar{C}}{dt_1} = \left[\int_{t_1}^{t_1} |V^T(t_1, s)\underline{\bar{c}}| \, ds \right]^{-1} \left[-\bar{C}|V^T(t_1, t_1)\underline{\bar{c}}| - \underline{\bar{c}}^T D(t_1)A(t_1)\underline{x}(t_1) - \underline{\bar{c}}^T \frac{dD(t_1)}{dt_1}\underline{x}(t_1) + \underline{\bar{c}}^T \frac{d\underline{b}(t_1)}{dt_1} \right] \tag{4.8}$$

The relationship of this slope to the necessary condition for time optimality given by Eq. 4.2 will now be derived: Upon substituting Eqs. 4.3, 4.4, 4.6, and 4.7 into Eq. 4.2 and making use of the fact that $X(t_1, t_1) = I$, we obtain

$$\psi_o + \underline{c}^T D(t_1)A(t_1)\underline{x}(t_1) + C\underline{c}^T D(t_1)B(t_1)G^{-1}(t_1)\frac{V^T(t_1, t_1)\underline{c}}{|V^T(t_1, t_1)\underline{c}|} + \underline{c}^T \frac{dD(t_1)}{dt_1}\underline{x}(t_1) - \underline{c}^T \frac{d\underline{b}(t_1)}{dt_1} = 0$$

which can be rearranged and simplified to give

$$\psi_o = -C|V^T(t_1, t_1)\underline{c}| - \underline{c}^T D(t_1)A(t_1)\underline{x}(t_1) - \underline{c}^T \frac{dD(t_1)}{dt_1}\underline{x}(t_1) + \underline{c}^T \frac{d\underline{b}(t_1)}{dt_1} \tag{4.9}$$

Thus, ψ_o is directly proportional to $\frac{d\bar{C}}{dt_1}$, with positive proportionality constant. This fact leads to the conclusion that a minimum peak amplitude problem for which $\frac{d\bar{C}}{dt_1}$ is positive corresponds to a trajectory which satisfies all the time-optimal necessary

conditions given in Section 4. 2 for the inverse problem, except the condition that ψ_o be non-positive. The failure of this condition implies that the solution is not time-optimal, and hence that the inverse relationship does not exist in this case.

When $\frac{dC}{dt_1}$ is negative or zero, the ψ_o for the corresponding time optimal problem is also negative or zero, but the inverse relationship may or may not exist. (We refer again to problems with final times t_{1b} and t_{1f} in Fig. 4. la) to illustrate this point.)

4. 5 Summary and Conclusions

It has been shown in this chapter that an intimate relationship (called here the "inverse relationship") exists between the original minimum peak amplitude problem and a certain class of time optimal problems, and that in many cases the answer to one problem can be obtained by solving the other, and vice versa. This would seem to indicate that we have obtained two distinct methods of solving the original linear minimum peak amplitude problem. The methods are actually the same, however, since both involve the solution of essentially the same set of equations; that is, one must find a control of the form given by Thm. 3. 2 (or Eq. 4. 7) which causes the boundary conditions to be satisfied. For nonlinear systems, on the other hand, there is no well-developed theory for the solution of minimum peak amplitude problems to correspond to that developed in Chapter III for linear systems. In such nonlinear problems, there is therefore some hope that the inverse problem could be useful. This is touched upon in Chapter VI.

In the next chapter (Chapter V) a method of solution of the linear problem is discussed which is in fact distinct from any of the methods discussed thus far, and which may have computational advantages in some cases. As will be seen, this method has certain things in common with the direct methods discussed in Section 1. 24.

CHAPTER V

THE RELATED LINEAR PROBLEM

5.1 Motivation

It was noted above that the usual variational techniques do not apply to minimum peak amplitude problems because the cost functional

$$C = \underset{t \in T}{\text{ess. sup.}} \; |G(t)\underline{u}(t)| \tag{5.1}$$

is not expressible as an integral. However, for G(t) and $\underline{u}(t)$ as described[43] in Sec. 2.1, it is true that functional[44]

$$\|G\underline{u}\|_p = \left[\frac{1}{T} \int_T |G(t)\underline{u}(t)|^p \, dt\right]^{\frac{1}{p}} \tag{5.2}$$

can be made to approximate the essential supremum of $|G(t)\underline{u}(t)|$ to any arbitrary degree of accuracy by choosing the constant p sufficiently large [Taylor, Ref. 81, p. 91]. This fact suggests a possible alternative approach to the minimum peak amplitude problem:

Suppose we set up a problem (to be referred to as the related problem) with the same system characterization (Eqs. 2.1 and 2.2) and the same boundary conditions as in the original problem, but with $\|G\underline{u}\|_p$ as the quantity to be minimized, instead of the quantity given by Eq. 5.1. (This is a problem that can be treated by

---

[43] Since $\underline{u}(t)$ and G(t) consist of bounded measurable functions (by assumption) the quantities $|G(t)\underline{u}(t)|$ and $|G(t)\underline{u}(t)|^p$ are also bounded measurably functions, and the integral shown here exists for all $p \geq 1$. Thus $\|G\underline{u}\|_p$ also exists for all $p \geq 1$.

[44] The symbol T used as an algebraic quantity denotes the length of the interval T, i.e., $t_1 - t_0$. The norm of a function is defined in the way shown here, rather than without the factor 1/T, because this definition allows considerable simplification in the presentation of certain later results.

either functional analysis or classical variational techniques.) Assume that this related problem could be solved, and that expressions could be obtained for $\underline{x}(t)$, $\underline{u}(t)$, etc., having p as a parameter. Since the cost functional for the related problem approaches that of the original problem as p approaches infinity, it is not unreasonable to hope that the solution(s) to the related problem will likewise approach (in some meaningful and useful sense) the solution(s) to the original problem.[45]

Should this indeed prove to be the case, it would open up two new possibilities:

a) It can happen in the minimum peak amplitude problem that there is no unique optimal control. (See, for instance, Example 1 of Chapter III.) This can give rise to computational difficulties, since Thm. 3.2 does not completely specify the optimal control in such cases. If it could be shown that the related linear problem always had a solution which could be obtained from some simple expression (such as the first line of Eq. 3.12), then there might be computational advantages inherent in a procedure which approximated the desired solution by a solution of the related problem for large p. This would be especially true if means were available for estimating how far from optimal the result was.

b) The functional analysis techniques used in Chapter III to solve the original linear problem do not apply directly to nonlinear problems. However, the related problem can be solved by classical variational techniques, which apply equally well to linear and nonlinear problems. Therefore, we might hope to be able to solve <u>nonlinear</u> minimum peak amplitude problems by setting up a related <u>nonlinear</u> problem, solving it by classical techniques, and then passing to the limit. This possibility is investigated in Chapter VI.

---

[45] This procedure is analogous to that used by Kirillova (Ref. 35) for proper systems with a single control variable.

The proposed approach is analogous to the interchanging of an integration and a limiting process in the calculus: Here we wish to interchange an optimization process and a limiting process; that is, instead of solving an optimization problem involving a limiting form of $\|G\underline{u}\|_p$, we wish to solve an optimization problem involving $\|G\underline{u}\|_p$ and then pass to the limit. And, as in the analogous interchanging of integration and limit, it is by no means certain that the procedure is valid. Its validity will be established for the linear problem. For the nonlinear problems considered in the next chapter, however, corresponding results are not available, so that for such problems this approach must be considered as a heuristic or practical tool, rather than a rigorous procedure for determining optimal controls.

5.2 Statement and Solution of the Related Linear Problem

In the original problem, as formulated in Chapter II, the optimal control was to be sought among the set of essentially bounded measurable controls which caused the system to satisfy the desired boundary conditions. For the related linear problem, it will be convenient[46] to choose as the class of admissible controls not the class of essentially bounded controls but the class of controls for which the integral

$$\int_T |G(t)\underline{u}(t)|^p dt$$

exists. (This is, of course, a different class for each different choice of p, but in every case it includes the class of essentially bounded measurable controls as a subclass.)

This broadening of the class of admissible controls would seem to open up the possibility that the optimal control for the related linear problem could turn out to be not essentially bounded, a result which would offer computational difficulties if nothing else. However, we shall show that for the kinds of systems considered here (see Sec. 2.1), the optimal controls for the related linear problem are essentially bounded, so that this difficulty does not occur.

---

[46] The results obtained here rest on the fact that the space of admissible controls is a Banach space. This choice of the class of admissible controls simplifies the proof that the control space is a Banach space. See Appendix E.

We now proceed to a statement of the related linear problem:

Let L_p^r be the space of r-component column vector functions of time $\underline{z}(t)$ defined on the interval T which are integrable in the above sense, and with norm defined by

$$\|\underline{z}\|_p = \left[\frac{1}{T}\int_T |\underline{z}(t)|^p dt\right]^{\frac{1}{p}} \tag{5.3}$$

Then L_p^r is a Banach space.[47]

Let $U_p(\underline{a}, \underline{b}, T)$ be the set of all controls $\underline{u}(t)$ for which $G(t)\underline{u}(t)$ is in L_p^r and which cause the system defined in Sec. 2.1 to satisfy the boundary conditions $\underline{x}(t_o) = \underline{a}$, $\underline{y}(t_1) = \underline{b}$. Choose as an optimal control any element of $U_p(\underline{a}, \underline{b}, T)$ for which $\|G\underline{u}\|_p$ takes on its minimum value over all $\underline{u}(t)$ in $U_p(\underline{a}, \underline{b}, T)$.

In order to simplify the notation, we proceed as in Chapter III and define $\underline{z}(t) = G(t)\underline{u}(t)$. The related linear problem is then equivalent to the problem of choosing a $\underline{z}(t)$ in L_p^r satisfying Eq. 3.7 for which $\|\underline{z}\|_p$ is a minimum.

The existence of an optimal control for this problem follows at once from Neustadt's Theorem 1 (Ref. 77), or can be proven by steps analogous and in many cases identical to those used in the proof of Theorem 3.1 above (See Sec. 3.2). The form of the optimal control(s) can also be obtained from Neustadt's Theorem 1 (Ref. 77), or by steps analogous to those used in Section 3.4 for the minimum peak amplitude problem. The details of these proofs are omitted here. We present instead only a lemma analogous to the lemma of Sec. 3.4, which can be regarded as a generalization of Hölder's inequality to the Banach spaces considered here.

Lemma: Let $\underline{z}(t)$ and $\underline{v}(t)$ be r-component vector functions of time defined on the interval T and belonging to the spaces[48] L_p^r and L_q^r, respectively,

---

[47] See Appendix E for a discussion of this space.

[48] The vectors $\underline{v}(t)$ that we are actually concerned with in these problems are bounded in Euclidean norm, and hence belong to L_q^r. The fact that L_q^r contains other functions which are not bounded or even essentially bounded is of no importance here.

where $1 < p < \infty$ and $q = \frac{p}{p-1}$. Let T_1 be the subset of T on which $|\underline{v}(t)| \neq 0$. Then

$$\frac{1}{T} \int_T \underline{v}(t)\, \underline{z}(t)\, dt \leq \left[\frac{1}{T} \int_T |\underline{z}(t)|^p\, dt\right]^{\frac{1}{p}} \left[\frac{1}{T} \int_T |\underline{v}(t)|^q\, dt\right]^{\frac{1}{q}}$$

Furthermore, equality is taken on in this inequality if and only if $\underline{z}(t)$ satisfies

$$\underline{z}(t) = \begin{cases} K|\underline{v}(t)|^{q-2}\, \underline{v}^T(t) & \text{for almost all } t \in T_1 \\ 0 & \text{for almost all } t \in [T - T_1] \end{cases}$$

where K is some nonnegative constant.

Proof:

$$\frac{1}{T} \int_T \underline{v}\, \underline{z}\, dt \leq \frac{1}{T} \left| \int_T \underline{v}\, \underline{z}\, dt \right| \leq \frac{1}{T} \int_T |\underline{v}\, \underline{z}|\, dt \leq \frac{1}{T} \int_T |\underline{v}(t)|\, |\underline{z}(t)|\, dt$$

$$\leq \left[\frac{1}{T} \int_T |\underline{z}(t)|^p\, dt\right]^{\frac{1}{p}} \left[\frac{1}{T} \int_T |\underline{v}(t)|^q\right]^{\frac{1}{q}} = \|\underline{z}\|_p \|\underline{v}\|_q$$

The last inequality is the usual form of Hölder's inequality for integrals. The first inequality becomes an equality if and only if $\int_T \underline{v}\, \underline{z}\, dt$ is nonnegative. The second becomes an equality if and only if the real scalar quantity $\underline{v}(t)\, \underline{z}(t)$ is of the same sign almost everywhere on T (excluding points at which it is zero). The third becomed an equality if and only if $\underline{v}^T(t)$ and $\underline{z}(t)$ are colinear a. e. on T. The last inequality becomes an equality if and only if

$$|\underline{z}(t)| = K|\underline{v}(t)|^{q-1} \quad \text{for almost all } t \in T,$$

where K is a positive constant (See, for example, Appendix A of Ref. 36). These four conditions combine to give the stated result. Q. E. D.

Using this lemma and steps analogous or identical to those used in the proof of Theorem 3. 2, Section 3. 4, we conclude that a control $\underline{z}_p(t)$ of the form

$$\underline{z}_p(t) = \left\{ \begin{array}{ll} C_p |V^T \underline{c}_p|^{-1 + \frac{1}{p-1}} V^T \underline{c}_p & \text{for almost all } t \in T_1(\underline{c}_p) \\ 0 & \text{for almost all } t \in [T - T_1(\underline{c}_p)] \end{array} \right\} \tag{5.4}$$

is an optimal control for the related linear problem, where C_p is a positive constant, $\underline{c}_p$ is a constant k-component column vector of nonvanishing Euclidean length, V stands for $V(t_1, t)$, $T_1(\underline{c}_p)$ is the subset of T on which $|V^T \underline{c}_p| \neq 0$, and where C_p and $\underline{c}_p$ satisfy the boundary condition equation

$$\underline{g} = C_p \int_T |V^T \underline{c}_p|^{-1 + \frac{1}{p-1}} V V^T \underline{c}_p \, dt \tag{5.5}$$

Here the subscripted p in C_p and $\underline{c}_p$ serve as a reminder that these quantities are different, in general, for different choices of p. For convenience in later sections, we normalize $\underline{c}_p$ to unit Euclidean length, making the appropriate change in C_p in the process.

We note in passing that nothing has been said about the uniqueness of the optimal control given by Eq. 5.4. This control turns out to be unique, but this fact is not important to the use to which these results are to be put, and therefore is not investigated here.

The following theorem is needed in the next section:

Theorem 5.1: The optimal control $\underline{z}_p(t)$ for the related linear problem defined by Eqs. 5.4 and 5.5 is essentially bounded in Euclidean norm; i.e.

$$\|\underline{z}_p\| = \operatorname*{ess.\,sup.}_{t \in T} |\underline{z}(t)| < \infty.$$

Proof: Since $\underline{z}_p(t)$ is by definition in L_p^r, the quantity $\|\underline{z}_p\|_p$ is a finite number. From Eqs. 5.3 and 5.4,

$$\|\underline{z}_p\|_p = C_p[\frac{1}{T}\int_T |V^T\underline{c}_p|^{\frac{1}{p-1}}dt]^{\frac{1}{p}}$$

Since V is bounded in norm for all $t \in T$ (due to the assumption of Chapter II), since $\underline{c}_p$ is a unit vector, and since $\frac{1}{p-1}$ and $\frac{1}{p}$ are positive numbers for all p allowed here (i.e., $1 < p < \infty$), the bracketed quantity and its $(\frac{1}{p})^{th}$ power are finite numbers, which implies that C_p is finite also.

Since $\underline{z}_p(t)$ can differ from the expressions given in Eq. 5.4 on a set of measure zero, at most, it follows that

$$\|\underline{z}_p\| = \underset{t \in T}{\text{ess. sup.}} [C_p |V^T\underline{c}_p|^{\frac{1}{p-1}}].$$

As noted above, C_p and $|V^T\underline{c}_p|$ are both bounded, and $\frac{1}{p-1}$ is a finite positive number, which implies that $\|\underline{z}_p\|$ is finite also. Q. E. D.

5.3 The Limiting Process

It is shown in this section that the solution $\underline{z}_p$ of the related linear problem can be made to approach the minimum peak amplitude solution in cost to any desired degree of accuracy, by choice of p sufficiently large. Furthermore, an upper bound on the difference between the peak amplitude of $\underline{z}_p$ and the minimum peak amplitude is obtained, so that this becomes a useful computational technique for obtaining nearly-optimal controls.

Before presenting the main theorem of this section, we prove certain lemmas that are needed for the proof of the theorem.

Lemma 5.1: For all essentially bounded measurable functions $\underline{z}(t)$ defined on $T = [t_0, t_1]$, and for $1 \leq m \leq n \leq \infty$

$$\|\underline{z}\|_m \leq \|\underline{z}\|_n$$

where $\|\underline{z}\|_\infty$ is defined[49] as $[\underset{t \in T}{\text{ess. sup.}} |\underline{z}(t)|]$. (Note that m and n are

---

[49] Note that $\|\underline{z}\|_\infty$ is the same as the quantity $\|\underline{z}\|$ defined by Eq. 3.6.

not necessarily integers, and that the n used here has no connection whatsoever with the n used in Chapters I through IV to denote the number of components in the system state vector.)

Proof: First of all, we note that the quantities appearing in this lemma are defined for all m and n in the given range. Also, for m = n the lemma is trivially true.

Thus, we confine attention to cases where $m < n$. We proceed by first proving a similar result for real scalar functions v(t); namely, that

$$\text{a)}\quad \left[\frac{1}{T}\int_T |v(t)|^m dt\right]^{\frac{1}{m}} \leq \left[\frac{1}{T}\int_T |v(t)|^n dt\right]^{\frac{1}{n}} \qquad n < \infty$$

$$\text{b)}\quad \left[\frac{1}{T}\int_T |v(t)|^m dt\right]^{\frac{1}{m}} \leq \operatorname*{ess.\,sup.}_{t \in T} |v(t)|$$

where v(t) is any essentially bounded measurable real scalar function defined on T.

For part a), we apply Hölder's inequality (in the usual scalar form) to

$$\frac{1}{T}\int_T |v(t)|^m dt = \frac{1}{T}\int_T |1| \cdot |v(t)|^m dt$$

using $p = \frac{n}{n-m}$ and $q = \frac{p}{p-1} = \frac{n}{m}$. (We note that, for the given ranges of m and n, p is greater than one and less than infinity, so that Hölder's inequality is indeed applicable.) Thus

$$\frac{1}{T}\int_T |v(t)|^m dt \leq \left[\frac{1}{T}\int_T |1|^p dt\right]^{\frac{1}{p}} \left[\frac{1}{T}\int_T |v(t)|^{mq} dt\right]^{\frac{1}{q}} =$$

$$\left[\frac{1}{T}\int_T |v(t)|^n dt\right]^{\frac{m}{n}}$$

which yields

$$\left[\frac{1}{T}\int_T |v(t)|^m dt\right]^{\frac{1}{m}} \leq \left[\frac{1}{T}\int_T |v(t)|^n dt\right]^{\frac{1}{n}} \quad \text{as claimed.}$$

For part b) we proceed as follows:

By definition, $|v(t)| \leq \operatorname*{ess.\,sup.}_{t \in T} |v(t)|$ for almost all $t \in T$, so that $|v(t)|^m \leq \left[\operatorname*{ess.\,sup.}_{t \in T} |v(t)|\right]^m$ for almost all $t \in T$. Therefore

$$\frac{1}{T}\int_T |v(t)|^m\,dt \le \frac{1}{T}\int_T \left[\underset{t\,\epsilon\,T}{\text{ess. sup.}}\,|v(t)|\right]^m dt = \left[\underset{t\,\epsilon\,T}{\text{ess. sup.}}\,|v(t)|\right]^m$$

Taking the m^{th} root of both sides, we have finally that

$$\left[\frac{1}{T}\int_T |v(t)|^m\,dt\right]^{\frac{1}{m}} \le \underset{t\,\epsilon\,T}{\text{ess. sup.}}\,|v(t)| \quad \text{as claimed.}$$

Now, since the Euclidean length of a vector $\underline{z}(t)$ is just a real scalar function of t, we can replace $|v(t)|$ by $|\underline{z}(t)|$ in the above expressions, and thereby obtain the inequalities which were to be established. Q. E. D.

Comment: This inequality is not true, in general, if the definition of the norm does not include the factor $\frac{1}{T}$. (To prove this, take $v(t) = 1$, $T = 2$, $m = 1$, $n = 2$.) It was to simplify the statement of this and the following lemmas that this factor was included in the original definition of $\|\underline{z}\|_p$.

Lemma 5. 2: a) $\|\underline{z}_m\|_m \le \|\underline{z}\|_m \quad 1 < m < \infty$

b) $\|\bar{\underline{z}}\|_\infty \le \|\underline{z}\|_\infty$

where $\underline{z}_m(t)$ is the optimal control of the related linear problem with p set equal to m, $\bar{\underline{z}}(t)$ is a minimum peak amplitude control, and $\underline{z}(t)$ is any control satisfying Eq. 3. 7 (i. e., causing the system to satisfy the desired boundary conditions).

Proof: In the related problem, the cost associated with a control $\underline{z}(t)$ is $\|\underline{z}\|_m$; similarly, in the minimum peak amplitude problem the cost[50] associated with a control $\underline{z}(t)$ is $\|\underline{z}\|_\infty$. These two inequalities are thus merely statements of the facts that $\underline{z}_m(t)$ and $\bar{\underline{z}}(t)$ are optimal controls for the corresponding problems.

---

50 This statement is not precisely true, since the quantity $\|\underline{z}\|_\infty$ involves the essential supremum of $|\underline{z}(t)|$, whereas the peak amplitude was originally defined in terms of the supremum of $|\underline{z}(t)|$. However, as discussed in Section 2. 2, we here consider the controls which result from these two cost functionals to be equivalent, and hence make no distinction between the two cost functionals.

Lemma 5.3:

a) $\|\underline{z}_m\|_m \leq \|\underline{z}_n\|_m \leq \|\underline{z}_n\|_n \leq \|\underline{z}_m\|_n \quad 1 \leq m \leq n < \infty$

b) $\|\underline{z}_m\|_m \leq \|\bar{\underline{z}}\|_m \leq \|\bar{\underline{z}}\|_\infty \leq \|\underline{z}_m\|_\infty \quad 1 \leq m < \infty$

Proof: The first inequality in each line follows from Lemma 5.2, the second from Lemma 5.1, and the third from Lemma 5.2.

Lemma 5.4: $\lim_{n \to \infty} [\frac{1}{T} \int_T |V^T \underline{c}|^{\frac{n}{n-1}} dt]^{\frac{1}{n}} = 1$

where $\underline{c}_n$ is the unit vector associated with the optimal control $\underline{z}_n(t)$ of the related problem, and where the usual absolute value metric is understood in connection with the limiting process.

Proof: Define $\|V\| = \underset{t \in T}{\text{ess. sup.}} [\sup_{|\underline{c}| = 1} |V^T \underline{c}|$. (Since $\underline{c}$ is a unit vector and since every element of V is a bounded measurable function of t, $\|V\|$ is a finite number.) Either $\|V\| \geq 1$ or $\|V\| < 1$. If $\|V\| < 1$, then

$$|V^T \underline{c}_n|^{\frac{n}{n-1}} \leq \|V\|^{\frac{n}{n-1}} \leq 1$$

for all $n \geq 2$, for all $\underline{c}_n$, and for almost all $t \in T$. If $\|V\| \geq 1$, then

$$|V^T \underline{c}_n|^{\frac{n}{n-1}} \leq \|V\|^{\frac{n}{n-1}} \leq \|V\|^2$$

for all $n \geq 2$, for all $\underline{c}_n$, and for almost all $t \in T$, since for such values of n the exponent $\frac{n}{n-1}$ is less than or equal to 2. Thus

$$\frac{1}{T} \int_T |V^T \underline{c}_n|^{\frac{n}{n-1}} dt \leq \max[\frac{1}{T} \int_T dt, \frac{1}{T} \int_T \|V\|^2 dt] = \max[1, \|V\|^2] < \infty$$

On the other hand, this quantity $\frac{1}{T} \int_T |V^T \underline{c}_n|^{\frac{n}{n-1}} dt$ is bounded away from zero. We prove this as follows: The complete controllability assumption (see Appendix D) states that the matrix $[\int_T VV^T dt]$ is positive definite. Its minimum eigenvalue is thus a positive number, which we denote by η. Then for all $\underline{c}$ such that $|\underline{c}| = 1$,

$$\underline{c}^T[\int_T VV^T dt]\underline{c} = \frac{1}{T}\int_T |V^T\underline{c}|^2 dt \geq \frac{1}{T}\eta\,|\underline{c}|^2 = \frac{1}{T}\eta > 0$$

Applying Holder's inequality to this, with p = 1 and q = ∞, gives

$$\frac{1}{T}\int_T |V^T\underline{c}|^2 dt \leq [\frac{1}{T}\int_T |V^T\underline{c}|\,dt][\operatorname*{ess.\,sup}_{t\in T} |V^T\underline{c}|] \leq [\frac{1}{T}\int_T |V^T\underline{c}|\,dt]\,\|V\| = \|V^T\underline{c}\|_1\|V\|$$

so that $\|V^T\underline{c}\|_1 \geq \frac{\eta}{T\|V\|}$. Now, using this result in Lemma 5.1, and using $V^T\underline{c}$ for $\underline{z}$, 1 for m, and $\frac{n}{n-1}$ for n in this lemma, gives

$$\|V^T\underline{c}\|_{\frac{n}{n-1}} = [\frac{1}{T}\int_T |V^T\underline{c}|^{\frac{n}{n-1}} dt]^{\frac{n-1}{n}}$$

Raising both sides to the $\frac{n}{n-1}$ power gives finally

$$[\frac{1}{T}\int_T |V^T\underline{c}|^{\frac{n}{n-1}} dt] \geq [\frac{\eta}{T\|V\|}]^{\frac{n}{n-1}} > 0$$

as claimed. Therefore, since the quantity $[\frac{1}{T}\int_T |V^T\underline{c}_n|^{\frac{n}{n-1}} dt]$ is uniformly bounded and uniformly bounded away from zero for all $n \geq 2$ and for all $\underline{c}_n$, and since the exponent $\frac{1}{n}$ approached zero as n approaches infinity, the statement of the lemma follows immediately.

Theorem 5.2: The limit $\lim_{n\to\infty} \|\underline{z}_n\|$ exists and is equal to $\bar{C}$, the optimal cost for the corresponding minimum peak amplitude problem. (Here, as above, the metric involved in the limiting process is understood to be the usual absolute value metric.)

Proof: First note that the limit exists, since, from Lemma 5.3, $\|\underline{z}_n\|_n$ is a monotone nondecreasing function of n bounded above (by $\|\bar{\underline{z}}\|_\infty = \bar{C}$, to name one of the many upper bounds provided by Lemma 5.3). Let the limiting value be denoted by C_∞ for the present. From Lemma 5.3, we have that

$$\|\underline{z}_n\|_n \leq \|\bar{\underline{z}}\|_\infty$$

which can be rewritten as

$$\|\underline{z}_n\|_n = C_n[\frac{1}{T}\int_T |V^T\underline{c}_n|^{\frac{n}{n-1}} dt]^{\frac{1}{n}} \leq \bar{C}$$

Since, by definition, $\lim_{n\to\infty} \|\underline{z}_n\|_n = C_\infty$, and since, from Lemma 5.4

$$\lim_{n\to\infty} \left[\frac{1}{T} \int_T |V^T \underline{c}_n|^{\frac{n}{n-1}} dt\right]^{\frac{1}{n}} = 1$$

it follows at once that $\lim_{n\to\infty} C_n = C_\infty$ and that $C_\infty \leq \bar{C}$.

Lemma 5.3 also states that

$$\|\bar{\underline{z}}\|_\infty \leq \|\bar{\underline{z}}_n\|_\infty = C_n \left[\underset{t \in T}{\text{ess. sup.}} |V^T \underline{c}_n|^{\frac{1}{n-1}}\right]$$

It is clear from the definition of $\|V\|$ that $|V^T \underline{c}| \leq \|V\|$ for all unit vectors $\underline{c}$ and for almost all $t \in T$. Since $\underline{c}_n$ is a unit vector, it follows that, for $n \geq 2$,

$$|V^T \underline{c}_n|^{\frac{1}{n-1}} \leq \|V\|^{\frac{1}{n-1}} \qquad \text{for almost all } t \in T$$

Therefore $\left[\underset{t \in T}{\text{ess. sup.}} |V^T \underline{c}_n|^{\frac{1}{n-1}}\right] \leq \left[\underset{t \in T}{\text{ess. sup.}} \|V\|^{\frac{1}{n-1}}\right] = \|V\|^{\frac{1}{n-1}}$. Thus

$\|\bar{\underline{z}}\|_\infty = \bar{C} \leq \|\underline{z}_n\|_\infty \leq C_n \|V\|^{\frac{1}{n-1}}$. But, as n approaches ∞, C_n approaches C_∞ (from the first part of this proof), and $\|V\|^{\frac{1}{n-1}}$ approaches $\|V\|^0 = 1$, because $\|V\|$ is a bounded and nonzero number. Thus $\bar{C} \leq C_\infty$. Combining these two inequalities involving $\bar{C}$ and C_∞ gives $C_\infty \leq \bar{C} \leq C_\infty$, which of course implies that $\bar{C} = C_\infty$. Q. E. D.

With little additional effort we can also obtain the following useful result, which implies that the sequence of controls $\underline{z}_n(t)$, $n = 1, 2, \ldots$ come arbitrarily close to $\bar{C}$ in peak amplitude, and hence that this sequence is a minimizing sequence in the sense of Section 1.2.4.

Corollary: For every $\epsilon > 0$ there exists a number $N(\epsilon)$ such that $\left| \|\underline{z}_n\|_\infty - \bar{C} \right| < \epsilon$ for all $n > N(\epsilon)$. That is, we can obtain a control satisfying Eq. 3.7 which has a peak amplitude as close as we like to the minimum peak amplitude, simply by choosing n large enough and using the optimal control $\underline{z}_n(t)$ for that problem.

Proof: As stated in the proof of Theorem 5.2, $\|\underline{z}_n\|_\infty$ lies between $\bar{C}$ and $C_n \|V\|^{\frac{1}{n-1}}$. But $C_n \|V\|^{\frac{1}{n-1}}$ approaches the limit $\bar{C}$ as n approaches infinity, as is also shown in the

proof of Theorem 5. 2. Therefore $\|\underline{z}_n\|_\infty$ also approaches the limit $\bar{C}$ as n approaches infinity. The statement of this corollary is simply another way of phrasing this fact. Q. E. D.

These results, taken together, lead to the following conclusions: The minimum peak amplitude control can be approximated (in the sense that the costs are close) to any desired degree of accuracy by solutions of the related linear problem. Furthermore, for any choice of n, Lemma 5. 3 provides a bound on the amount that the actual peak amplitude of $\underline{z}_n(t)$ can exceed $\bar{C}$. To illustrate this point, suppose we pick a value of n, compute $\underline{z}_n(t)$, and determine $\|\underline{z}_n\|_n$ and $\|\underline{z}_n\|_\infty$. By Lemma 5. 3b), $\bar{C}$ must lie between $\|\underline{z}_n\|_n$ and $\|\underline{z}_n\|_\infty$. Thus, the peak amplitude for the control $\underline{z}_n(t)$ that we propose to use in place of $\bar{\underline{z}}(t)$ is $\|\underline{z}_n\|_\infty$, and we know that $\bar{C}$ can not be less than $\|\underline{z}_n\|_n$. We are thus within $[\,\|\underline{z}_n\|_\infty - \|\underline{z}_n\|_n]$ of the minimum cost $\bar{C}$. If this quantity is sufficiently small to suit us, we say that $\underline{z}_n(t)$ is a sufficiently good approximation to the optimal control.

All this does not imply that the control $\underline{z}_n(t)$ converges [in the metric defined by the norm given in Eq. 3. 6] to $\bar{\underline{z}}(t)$ as n approaches ∞. This is a point that is not resolved one way or the other by the above discussions and proofs. In fact, the examples of the next section show that, in general, one can not expect $\underline{z}_n(t)$ to converge in this metric to $\bar{\underline{z}}(t)$. This point will not be investigated further here, except to the extent that the examples of the next section give insight into the matter.

5. 4 Examples of the Limiting Process:

The limiting process discussed above will be illustrated by its application to some examples.

Example 9: Consider the same system used in Example 1, with initial conditions $\underline{x}(0) = \underline{0}$ and desired final conditions $\underline{x}(2) = [\,2, 1]^T$. Example 1 shows that this problem does not have a unique minimum peak amplitude control, and that any control satisfying

$$\left.\begin{aligned} u_1(t) &= \begin{cases} 4 & 0 \le t < 1 \\ 0 & 1 < t \le 2 \end{cases} \\ u_2(t) &= 0 \qquad 0 \le t < 1 \\ |u_2(t)| &\le 4 \qquad 1 < t \le 2 \\ &\int_1^2 (t-1)\, u_2(t)\, dt = 1 \end{aligned}\right\} \tag{5.6}$$

is a minimum peak amplitude control.

The optimal control for the related linear problem [as given by Eq. 5.4, with $\underline{u}(t) = \underline{z}(t)$] is

$$\underline{u}(t) = C_p \left| V^T(2,t)\, \underline{c}_p \right|^{-1+\frac{1}{p-1}} V^T(2,t)\, \underline{c}_p$$

Upon denoting the two components of $\underline{c}_p$ by c_{1p} and c_{2p} and making use of the expressions for $V(2,t)$ given in Example 1, this becomes

$$u_1(t) = \begin{cases} C_p \operatorname{sgn} c_{1p} \left| (1-t)\, c_{1p} \right|^{\frac{1}{p-1}} & 0 \le t \le 1 \\ 0 & 1 \le t \le 2 \end{cases}$$

$$u_2(t) = \begin{cases} 0 & 0 \le t \le 1 \\ C_p \operatorname{sgn} c_{2p} \left| (t-1)\, c_{2p} \right|^{\frac{1}{p-1}} & 1 \le t \le 2 \end{cases}$$

Use of Eq. 5.5 to evaluate C_p, c_{1p}, and c_{2p}, and normalization of $\underline{c}_p$ to unit Euclidean length yields

$$C_p = \left[2 + \frac{1}{p-1} \right] \left[|g_1|^{2p-2} + |g_2|^{2p-2} \right]^{\frac{1}{2p-2}}$$

$$c_{1p} = [\operatorname{sgn} g_1] \left[1 + \left| \frac{g_2}{g_1} \right|^{2p-2} \right]^{-\frac{1}{2p-2}}$$

$$c_{2p} = [\operatorname{sgn} g_2] \left[1 + \left| \frac{g_1}{g_2} \right|^{2p-2} \right]^{-\frac{1}{2p-2}}$$

Finally, setting $g_1 = 2$ and $g_2 = 1$ and substituting these results into the above expressions for $u_1(t)$ and $u_2(t)$, we have

$$u_1(t) = \begin{cases} 2\left[2 + \frac{1}{p-1}\right][1-t]^{\frac{1}{p-1}} & 0 \le t \le 1 \\ 0 & 1 \le t \le 2 \end{cases}$$

$$u_2(t) = \begin{cases} 0 & 0 \le t \le 1 \\ \left[2 + \frac{1}{p-1}\right][t-1]^{\frac{1}{p-1}} & 1 \le t \le 2 \end{cases}$$

Fig. 5. 1 shows plots of these functions for various values of p. As p approaches ∞, they approach the functions

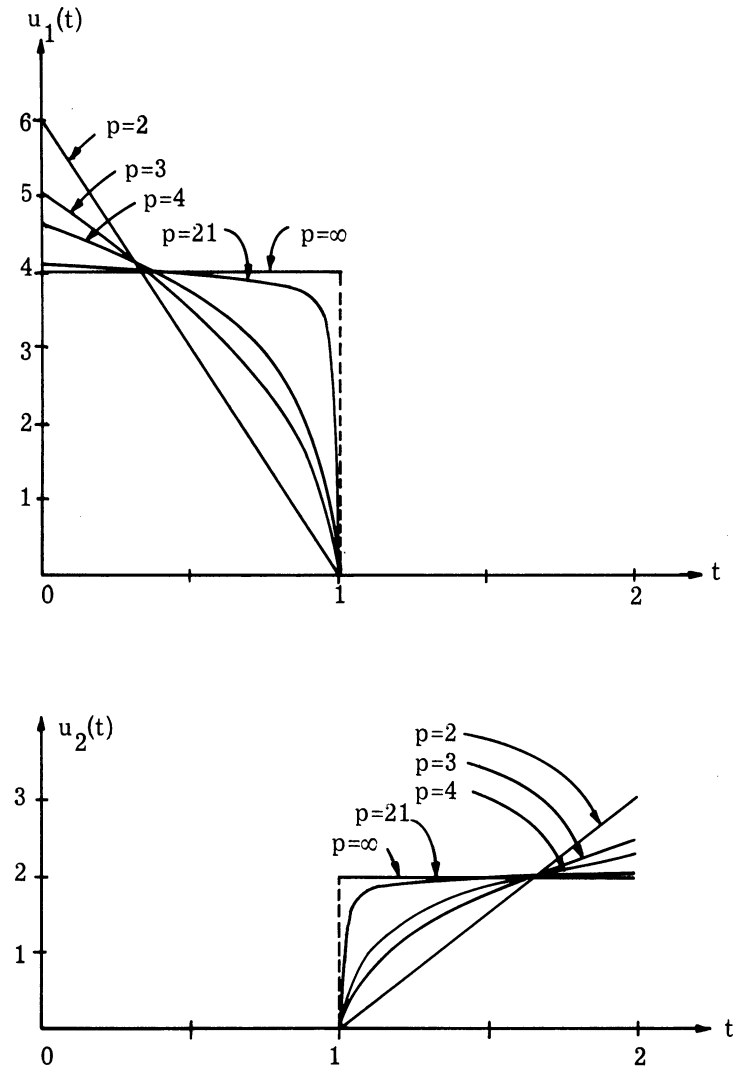

Fig. 5. 1 The limiting process illustrated for Example 9.

$$u_1(t) = \begin{cases} 4 & 0 \leq t < 1 \\ 0 & 1 < t \leq 2 \end{cases}$$

$$u_2(t) = \begin{cases} 0 & 0 \leq t < 1 \\ 2 & 1 < t \leq 2 \end{cases}$$

which can easily be shown to satisfy Eqs. 5.6. Furthermore, they approach these limiting functions (in the sense that the area between these curves and the limiting curves approaches zero, and in the sense that the peak amplitude of these curves approaches that of the limiting curve) with reasonable rapidity. For example, half of the total reduction in peak amplitude that can be achieved in going from $p = 2$ to $p = \infty$ is already accomplished for $p = 3$. Also, for $p = 21$, the control function $u_1(t)$ already looks very much like the optimal rectangular function, and the peak amplitude is within 2.5 percent of the minimum.

It is instructive to consider the bounds on the minimum peak amplitude provided by the solution of the related linear problem for various values of p, as discussed in the last section. Fig. 5.2 shows the upper bound $\|\underline{u}_p\|_\infty$ and the lower bound $\|\underline{u}_p\|_p$ for this example for various values of p. The actual value of $\bar{C}$ is $=4$.

Two other points are worthy of mention here:

1) If instead of looking at the functions $\underline{u}_p(t)$ as p increases, we look instead at the vector $\underline{c}_p$, which completely determines the shape of $\underline{u}_p(t)$ for each $p < \infty$, we see that $\underline{c}_p$ approaches $[1, 0]^T$ as p approaches ∞. Substitution of this result into Eq. 3.12 does not completely determine an optimal control. Thus, in this example the functions $\underline{u}_p(t)$ converge[51] to a particular function, but the corresponding limiting process in $\underline{c}_p$ does not uniquely determine this or any other function.

2) Any control involving $u_1(t) = 4$, $0 \leq t < 1$, $u_1(t) = 0$, $1 \leq t \leq 2$ along with any one of the functions $u_2(t)$ shown in Fig. 5.1 is an optimal control, since it has peak amplitude equal to 4 (the minimum) and it satisfies the desired boundary conditions. However, the limiting process tends toward a particular function for $u_2(t)$, which we note to be of the

---

[51] The convergence is not uniform, of course, because of the behavior of $\underline{u}_p(t)$ at $t = 1$. However, the sequence $\underline{u}_p(t)$, $p = 2, 3, \ldots$, converges to $\underline{\bar{u}}(t)$ in the metric defined by any of the norms $\| \ \|_p$, for $1 \leq p < \infty$.

same rectangular shape as the uniquely-determined control component $u_1(t)$. This is both an advantage and a disadvantage: an advantage in that the process does generate a specific and unambiguous control function; a disadvantage in that the possibility of secondary optimizations, as discussed in Sec. 3. 5, is lost.

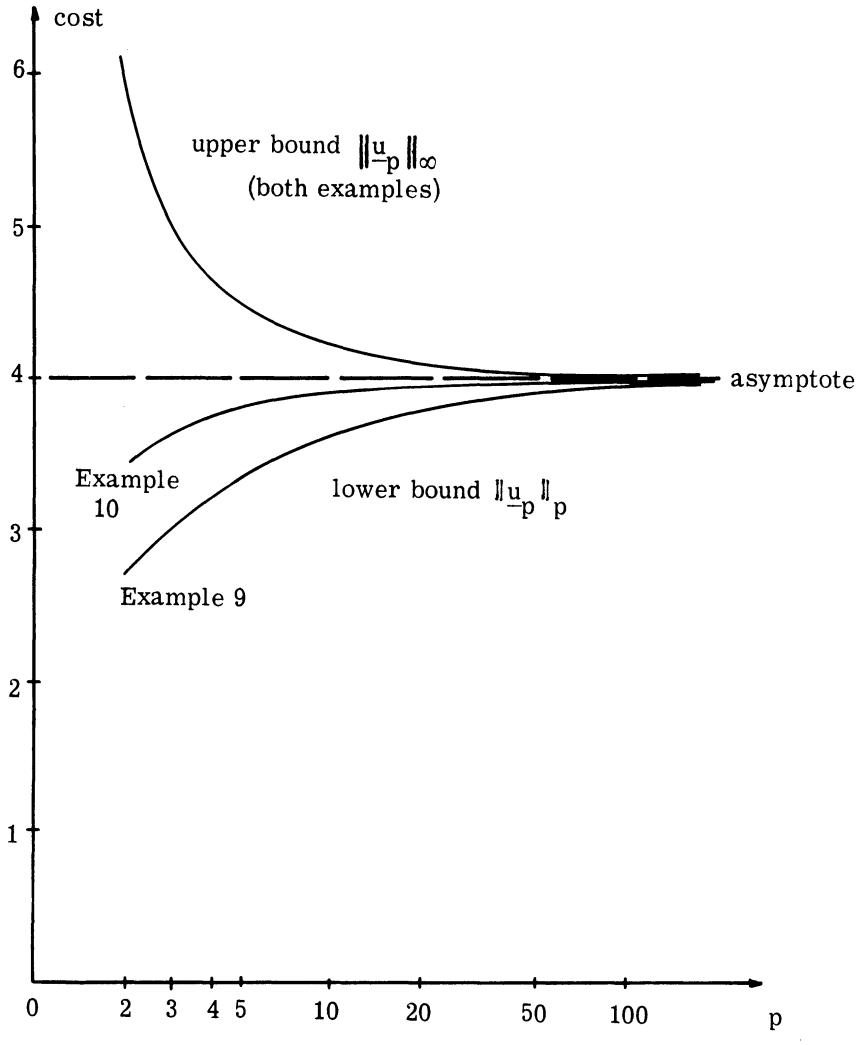

Fig. 5. 2 Upper and lower bounds on minimum peak amplitude provided by related problems for Examples 9 and 10.

Example 10: Consider the same system used in Example 3 (the double integrator) with initial state $\underline{x}(0) = \underline{0}$ and desired final state $\underline{x}(1) = [1, 0]^T$. The minimum peak amplitude control for this example is

$$\bar{u}(t) = \begin{cases} 4 & 0 \leq t < \frac{1}{2} \\ -4 & \frac{1}{2} < t \leq 1 \end{cases}$$

The solution to the related linear problem is

$$u_p(t) = 2(2 + \frac{1}{p-1}) \, |1 - 2t|^{\frac{1}{p-1}} \operatorname{sgn}[1 - 2t]$$

$$\underline{c}_p = [\sqrt{.8}, \sqrt{.2}]^T$$

$$C_p = 2(2 + \frac{1}{p-1})(.2)^{\frac{1}{2p-2}}$$

$$\|u_p\|_p = 2(2 + \frac{1}{p-1})^{1 - \frac{1}{p}}$$

$$\|u_p\|_\infty = 2(2 + \frac{1}{p-1})$$

(All of these results can be verified by means of Theorems 3. 2 and 5. 1 and the results of Example 3.)

Fig. 5. 3 shows $u_p(t)$ for various values of p. Because the bounds on the minimum peak amplitude provided by $\|u_p\|_p$ and $\|u_p\|_\infty$ are (by coincidence) the same as or similar to the corresponding results for the previous example, these bounds for this example are also plotted in Fig. 5. 2.

Note that in this example $\underline{c}_p$ is always the same, independent of p, and that the functions $u_p(t)$ resemble the minimum peak amplitude control more and more as p grows larger.

Example 11: Consider the same system used in Example 5, namely $\dot{x} = -x + u$, with fixed final time $t_1 = 1$ and with boundary conditions $x(0) = 0$, $x(1) = 1$. The minimum peak amplitude control for this problem, as derived in Example 5, is

$$\bar{u}(t) = \frac{1}{1 - e^{-1}} = \frac{e}{e - 1} = \bar{C}$$

The solution to the related linear problem is

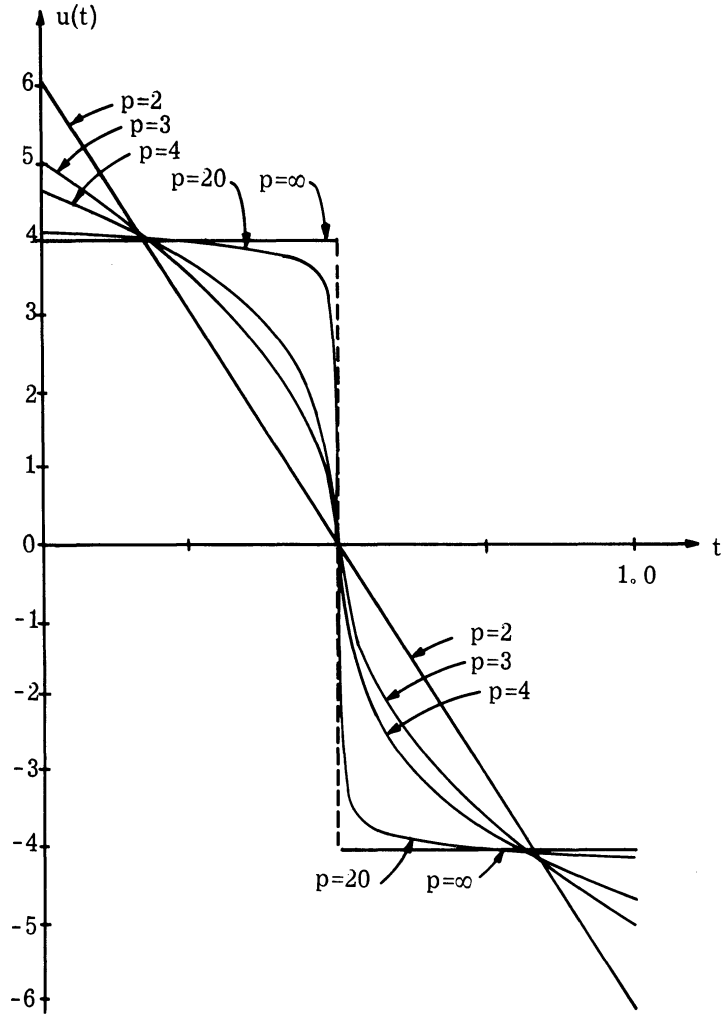

Fig. 5.3 The limiting process illustrated for Example 10.

$$u_p(t) = C_p\, e^{\frac{t-1}{p-1}}$$

$$\|u_p\|_p = C_p^{\,1-\frac{1}{p}}$$

$$\|u_p\|_\infty = C_p$$

where

$$C_p = q\frac{1}{1-e^{-q}} = q\frac{e^q}{e^q-1} \quad \text{and} \quad q = 1+\frac{1}{p-1}$$

These results follow from Eq. 5. 4, using $V(t_1, t) = e^{-t_1 + t}$. Fig. 5. 4a) shows $u_p(t)$ for various values of p, and Fig. 5. 4b) shows the behavior of the upper and lower bounds on $\bar{C}$ as a function of p. Since in this example $\bar{u}(t)$ does not involve any "switchings" (i. e., step discontinuities), the convergence of $u_p(t)$ to $\bar{u}(t)$ is uniform on the open interval (0, 1).

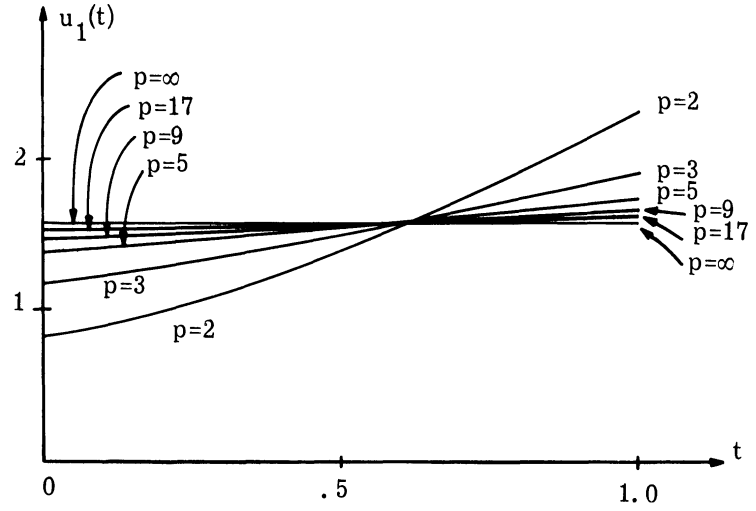

Fig. 5. 4a) $u_p(t)$ for various values of p for Example 11.

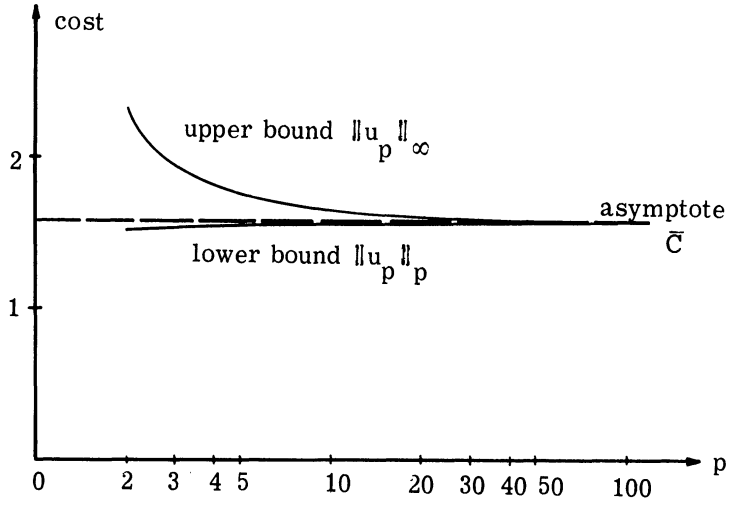

Fig. 5. 4b) Upper and lower bounds on $\bar{C}$, as a function of p.

CHAPTER VI

NONLINEAR PROBLEMS

6.1 Introduction

There is no rigorously established technique presently available for the solution of minimum peak amplitude problems for nonlinear systems. Variational techniques are not readily applicable because the cost functional is not expressible as an integral, dynamic programming fails for the same reason. Functional analysis methods for nonlinear systems are not sufficiently well developed to allow the solution of minimum peak amplitude problems.

The method to be presented, while it is also lacking in rigor, may be of engineering value in some problems, in that it provides a "likely candidate" for the optimal control, which in some cases can be shown to be optimal by other tests.

The method is essentially as outlined in the previous chapter: Set up and solve the related nonlinear problem for various values of the parameter p (using variational techniques, in this case), and then either use the optimal solution to the related problem for large p as an approximation (who knows how good an approximation) to the minimum peak amplitude control, or else take the limiting form of this solution (if it exists) as p approaches infinity as the "candidate" for the optimal control.

As has been suggested above, the difficulties that could be encountered in this approach are many. To name a few: There may be no minimum peak amplitude control (only a minimizing sequence); the related problem may not converge to the minimum, or it may converge to a local but not an absolute minimum; or the related problem may have no solution. The many powerful theorems that were available in the linear case to eliminate such possibilities are not available here. Even such basic assumptions of the linear problem as the complete controllability assumption have no convenient analytic formulation in the nonlinear case. The test of the method must therefore be in its ability to obtain results. It is not the intention of this work to investigate this problem in detail. The example given

later in this chapter serves to indicate that the method has some validity, nonetheless.

6.2 Statement of the Nonlinear Problem

Since the proposed limiting method involves the use of variational techniques, we restrict attention to a class of systems to which these variational techniques (in their present state of development; see Pontryagin, et al., Ref. 13, p. 79) are applicable; that is, we assume that the physical system to be controlled can be characterized by the system of differential equations

$$\dot{\underline{x}} = \underline{f}(\underline{x},\underline{u},t) \tag{6.1}$$

with the output $\underline{y}(t)$ being given by

$$\underline{y}(t) = D\,\underline{x}(t) \tag{6.2}$$

Here $\underline{f}(\underline{x},\underline{u},t)$ is a vector function of the vectors $\underline{x}$ and $\underline{u}$ and the scalar t, with components $f_1(\underline{x},\underline{u},t),\ldots,f_n(\underline{x},\underline{u},t)$ which are defined for all $\underline{x} \in E^n$, $\underline{u} \in E^r$, and $t \in T = [t_0,t_1]$. Furthermore, each component $f_i(\underline{x},\underline{u},t)$ of $\underline{f}(\underline{x},\underline{u},t)$ is assumed to be continuous in all its variables and continuously differentiable with respect to t and to the components $x_1,\ldots,x_n$ of $\underline{x}$. The matrix D is as defined in Section 2.1.

The problem is then to choose, from among all the bounded measurable controls $\underline{u}(t)$ which cause the system to transfer from the initial state

$$\underline{x}(t_0) = \underline{a} \tag{6.3}$$

at initial time t_0 to any state for which

$$\underline{y}(t_1) = \underline{b} \tag{6.4}$$

at final time t_1, any one for which

$$\operatorname*{ess.\,sup.}_{t \in T} \left| G(t)\,\underline{u}(t) \right|$$

is minimum. The matrix G(t) is as defined in Section 2.1.

6.3 Statement and Partial Solution of the Related Problem

In exact analogy to the related linear problem of Chapter V, we define the related nonlinear problem as the problem identical to the problem of Section 6.2 except that the functional to be minimized is

$$\|G\underline{u}\|_p = \left[\frac{1}{T}\int_T |G(t)\,\underline{u}(t)|^p\,dt\right]^{\frac{1}{p}} \qquad 1 < p < \infty$$

instead of the essential supremum of $|G(t)\,\underline{u}(t)|$. The functional $\|G\underline{u}\|_p$ is still not of a form to which the usual variational techniques apply directly, but this presents no problem, because the minimization of $\|G\underline{u}\|_p$ is completely equivalent[52] to the minimization of

$$J(\underline{u}) = \int_T |G(t)\,\underline{u}(t)|^p\,dt \qquad 1 < p < \infty$$

a functional to which variational techniques do apply.

We now apply the maximum principle of Pontryagin to obtain the following necessary condition for optimality:

> A necessary condition that a bounded measurable control $\underline{u}(t)$ and the corresponding $\underline{x}(t)$ and $\underline{y}(t)$ satisfying Eqs. 6.1, 6.2, 6.3, and 6.4 be optimal is that there exist a nonpositive constant ψ_0, a constant k-component column vector $\underline{\theta}$, and an absolutely continuous nonzero vector function $\underline{\psi}(t)$ with components $\psi_1(t),\ldots,\psi_n(t)$ satisfying

---

[52] To convince oneself of the equivalence of these two functionals, consider that a cost functional merely provides an ordering of the controls $\underline{u}(t)$. Suppose that, in the ordering provided by $\|G\underline{u}\|_p$, the control $\underline{u}_1(t)$ is lower in cost than $\underline{u}_2(t)$; i.e., $\|G\underline{u}_1\|_p < \|G\underline{u}_2\|_p$. But this implies that $[\|G\underline{u}_1\|_p]^p < [\|G\underline{u}_2\|_p]^p$ also, and hence that $J(\underline{u}_1) < J(\underline{u}_2)$. By similar arguments, one sees that $J(\underline{u}_1) < J(\underline{u}_2)$ implies $\|G\underline{u}_1\|_p < \|G\underline{u}_2\|_p$, and that $J(\underline{u}_1) = J(\underline{u}_2)$ implies and is implied by $\|G\underline{u}_1\|_p = \|G\underline{u}_2\|_p$. Thus, the ordering imposed by one cost functional is identical to the ordering imposed by the other, and the two cost functionals are equivalent, in the sense that controls which are optimal for one are optimal for the other, and vice versa.

$$\left.\begin{aligned} \dot{\psi}_i &= -\frac{\partial H(\underline{x}(t),\ \underline{\psi}(t),\ \underline{u}(t),\ t)}{\partial x_i(t)} \\ \psi_i(t_1) &= \frac{\partial Q}{\partial x_i(t_1)} \end{aligned}\right\} \qquad \begin{aligned} &t \in T \\ &i = 1,\ldots,n \end{aligned} \tag{6.5}$$

such that, for the stated $\underline{x}(t)$ and $\underline{\psi}(t)$, and for almost all $t \in T$, the quantity

$$H(\underline{x}(t),\ \underline{\psi}(t),\ \underline{v},\ t) = \psi_0 \left| G(t)\underline{v} \right|^p + \underline{\psi}^T(t)\, \underline{f}(\underline{x}(t),\ \underline{v},\ t) \tag{6.6}$$

regarded as a function of the variable $\underline{v}$, attains its maximum (over all finite $\underline{v}$) at the value $\underline{v} = \underline{u}(t)$. Here the terminal constraint function Q is defined as

$$Q = \underline{\theta}^T[D\,\underline{x}(t_1) - \underline{b}] \tag{6.7}$$

Carrying out the operations indicated in Eq. 6.5 gives

$$\begin{aligned} \dot{\psi}_i &= -\left[\frac{\partial f_1}{\partial x_i}, \ldots, \frac{\partial f_n}{\partial x_i}\right] \underline{\psi}(t) \\ \psi_i(t_1) &= \left[d_{1i}, d_{2i}, \ldots, d_{ni}\right] \underline{\theta} \end{aligned}$$

where d_{ij} is the element of the matrix D in the ith row and jth column. In vector matrix form, these equations can be written as

$$\dot{\underline{\psi}} = -J_x(\underline{x},\underline{u},t)\, \underline{\psi}(t) \tag{6.8}$$

$$\underline{\psi}(t_1) = D^T\theta \tag{6.9}$$

where $J_x(\underline{x},\underline{u},t)$ is the Jacobian matrix of $\underline{f}$ with respect to $\underline{x}$, i.e.,

$$J_x(\underline{x},\underline{u},t) = \frac{\partial(f_1, f_2, \ldots, f_n)}{\partial(x_1, x_2, \ldots, x_n)} = \begin{bmatrix} \dfrac{\partial f_1(\underline{x},\underline{u},t)}{\partial x_1} & \cdots & \dfrac{\partial f_1(\underline{x},\underline{u},t)}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f_n(\underline{x},\underline{u},t)}{\partial x_1} & \cdots & \dfrac{\partial f_n(\underline{x},\underline{u},t)}{\partial x_n} \end{bmatrix}$$

Thus far, the problem has been transformed from one involving n + k + r unknown functions of time $x_1,\ldots,x_n$, $y_1,\ldots,y_k$, $u_1,\ldots,u_r$, n differential equations (Eq. 6.1), k finite equations (Eq. 6.2), and n + k boundary condition (Eqs. 6.3 and 6.4), plus the requirement that a functional of $\underline{u}$ be minimized, into a problem involving 2n + k + r unknown functions $x_1,\ldots,x_n$, $\psi_1,\ldots,\psi_n$, $y_1,\ldots,y_k$, $u_1,\ldots,u_r$, 2n differential equations (Eqs. 6.1 and 6.8), k finite equations (Eq. 6.2), 2n + k boundary conditions (Eqs. 6.3, 6.4, and 6.9), k + 1 unknown constants ψ_0, $\theta_1,\ldots,\theta_k$, and the condition that $\underline{u}$ be chosen to maximum H at almost every $t \in T$. This last condition serves to determine r conditions that $\underline{u}$ must satisfy a. e. on T. When we take into account the fact that the equations determining $\underline{\psi}(t)$ are homogeneous in $\underline{\psi}(t)$, so that the length of $\underline{\psi}(t)$ may be arbitrarily specified at any one moment in time, we see that the transformed problem is a "well specified" problem, in that, in principle at least, there are enough differential equations, finite equations, conditions, etc., to specify all variables and constants of the problem. Whether this transformed problem does indeed have a solution is a question that can not be answered until more information about the system is given.

So that we may proceed to a more detailed discussion of the nature of the limiting process, we now restrict attention to systems for which the components of $\underline{f}(\underline{x},\underline{u},t)$ are not only continuous in $\underline{u}$ but continuously differentiable in $\underline{u}$ also. Under this assumption, the stationary points of H with respect to $\underline{u}$ can be found by setting the partial derivatives of H with respect to $\underline{u}$ equal to zero. Proceeding in this manner, we have that $\underline{u}(t)$ must satisfy

$$\frac{\partial}{\partial u_j}\left[\psi_0\left(\underline{u}^T G^T G \underline{u}\right)^{\frac{p}{2}} + \underline{\psi}^T \underline{f}(\underline{x},\underline{u},t)\right] = 0 \qquad j = 1,\ldots,r$$

or

$$\frac{p}{2}\psi_0\left(\underline{u}^T G^T G \underline{u}\right)^{\frac{p}{2}-1}\left(2[g_{1j},\ldots,g_{rj}]G\underline{u}\right) + \underline{\psi}^T \begin{bmatrix} \frac{\partial f_1(\underline{x},\underline{u},t)}{\partial u_j} \\ \vdots \\ \frac{\partial f_n(\underline{x},\underline{u},t)}{\partial u_j} \end{bmatrix} = 0 \qquad j = 1,\ldots,r$$

which can be written in vector-matrix form as

$$p\,\psi_0 \left|G\underline{u}\right|^{p-2} G^T G\underline{u} + J_u(\underline{x},\underline{u},t)\,\underline{\psi} = 0 \tag{6.10}$$

where $J_u(\underline{x},\underline{u},t)$ is the Jacobian matrix of $\underline{f}(\underline{x},\underline{u},t)$ with respect to $\underline{u}$, i.e.,

$$J_u(\underline{x},\underline{u},t) = \begin{bmatrix} \dfrac{\partial f_1(\underline{x},\underline{u},t)}{\partial u_1} & \cdots & \dfrac{\partial f_1(\underline{x},\underline{u},t)}{\partial u_r} \\ \vdots & & \vdots \\ \dfrac{\partial f_n(\underline{x},\underline{u},t)}{\partial u_1} & \cdots & \dfrac{\partial f_n(\underline{x},\underline{u},t)}{\partial u_r} \end{bmatrix}$$

By solving for $G\underline{u}$, taking the Euclidean norm of the result, and using this result to eliminate the quantity $\left|G\underline{u}\right|^{p-2}$ in Eq. 6.10, we can rewrite Eq. 6.10 as

$$G\underline{u} = \left[\frac{1}{-\psi_0 p}\right]^{\frac{1}{p-1}} \left|G^{-T} J_u^T\,\underline{\psi}\right|^{-1+\frac{1}{p-1}} G^{-T} J_u^T\,\underline{\psi} \tag{6.11}$$

a result which can be recognized as similar to the expression for the optimal control for the related linear problem, as given by Eq. 5.4. In the nonlinear case, this is not an explicit expression for $\underline{u}$, since $J_u(\underline{x},\underline{u},t)$ involves $\underline{u}$, in general. (The exception here is of course when $\underline{f}(\underline{x},\underline{u},t)$ is linear in $\underline{u}$; that is, when $\underline{f}(\underline{x},\underline{u},t)$ is expressible as $\underline{\bar{f}}(\underline{x},t) + B(t)\underline{u}$, in which case Eq. 6.11 is an explicit expression for $\underline{u}$, and resembles the results for the related linear problem even more.) Thus, in general, if $\underline{u}$ can not be solved for explicity, an iterative procedure must be used in order to determine the optimal $\underline{u}(t)$ (if it exists) from Eq. 6.11. Also, since this is only a necessary condition that H have a stationary point at $\underline{u}$, one must check to see that this $\underline{u}$ does indeed correspond to a maximum of H, and that no other $\underline{u}$ gives a higher maximum--a possibility common to all maximizations involving nonlinear functions.

6.4 Two Examples

The examples considered here are admittedly very simple ones, and the behavior of these examples can not be assumed to parallel that of nonlinear problems in their full generality. Simple examples were chosen so that the solutions could be obtained in analytic form, thus facilitating the study of their behavior under the limiting process.

Example 12:

The system for this example is the same double integrator used in Examples 3 and 10, but with a nonlinear input; namely,

$$\dot{x}_1 = x_2 \qquad \underline{x}(0) = \underline{0}$$

$$\dot{x}_2 = u^3 \qquad \underline{x}(1) = [2,0]^T$$

In the related problem we wish to choose u as a function of time so that these boundary conditions are satisfied and so that

$$\|u\|_p = \left[\frac{1}{T}\int_T |G(t)\,u(t)|^p\,dt\right]^{\frac{1}{p}} = \left[\int_0^1 |u(t)|^p\,dt\right]^{\frac{1}{p}}$$

is minimized, or equivalently, so that

$$J(u) = \int_0^1 |u(t)|^p\,dt$$

is minimized. Applying the maximum principle to this problem gives

$$H = \psi_0 |u|^p + \psi_1 x_2 + \psi_2 u^3$$

$$Q = \theta_1[x_1(1) - 2] + \theta_2 x_2(1)$$

$$\dot{\psi}_1 = 0 \qquad \psi_1(1) = \theta_1$$

$$\dot{\psi}_2 = -\psi_1 \qquad \psi_2(1) = \theta_2$$

choose u(t) to maximize H

as necessary conditions for optimality. The ψ equations are easily solved to give $\psi_1(t) = \theta_1$, $\psi_2(t) = \theta_1 + \theta_2 - \theta_1 t$, where θ_1 and θ_2 are constants. Equation 6.11 could be used to give the form of the optimal control $\underline{u}(t)$, but in this case it is just as easy to repeat the steps used to obtain Eq. 6.11, solving for $\underline{u}(t)$ explicitly in the process:

$$\frac{\partial H}{\partial u} = 0 = \psi_0 p |u|^{p-2} u + 3u^2 \psi_2$$

When $u \neq 0$, and for $p > 3$, this can be solved to give

$$u_p(t) = \left[\frac{3}{-p\psi_0}\right]^{\frac{1}{p-3}} \left|\psi_2(t)\right|^{\frac{1}{p-3}} \operatorname{sgn} \psi_2(t)$$

(The restriction of p to values greater than 3 is not a problem, since we plan to use large values of p anyway. For $p \leq 3$, different techniques must be used to find the value of $\underline{u}$ which maximizes H, but these will not be discussed here.) Letting the quantity $\left[\frac{3}{-p\psi_0}\right]^{1/p-3}$ be denoted by C_p, and making use of the above solution for $\psi_2(t)$, then gives

$$u_p(t) = C_p \left|\theta_1 + \theta_2 - \theta_1 t\right|^{\frac{1}{p-3}} \operatorname{sgn}[\theta_1 + \theta_2 - \theta_1 t]$$

Since the unforced system is linear in this case, the response of $\underline{x}$ for a given $\underline{u}(t)$ can be determined by taking the convolution of the forcing term with the fundamental matrix of the linear system;[53] namely

$$\underline{x}(t) = X(t, t_0)\,\underline{x}(t_0) + \int_{t_0}^{t} \mathbf{X}(t, s)\ [\text{forcing term}]\ ds$$

which in this case becomes

$$\underline{x}(t) = \int_0^t \begin{bmatrix} t-s \\ 1 \end{bmatrix} u^3(s) ds = C_p{}^3 \int_0^t \begin{bmatrix} t-s \\ 1 \end{bmatrix} \left|\theta_1 + \theta_2 - \theta_1 s\right|^{\frac{3}{p-3}} \operatorname{sgn}[\theta_1 + \theta_2 - \theta_1 s]\, ds$$

At $t = t_1 = 1$, this becomes

$$\underline{x}(t_1) = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = C_p{}^3 \int_0^1 \begin{bmatrix} 1-s \\ 1 \end{bmatrix} \left|\theta_1 + \theta_2 - \theta_1 s\right|^{\frac{3}{p-3}} \operatorname{sgn}[\theta_1 + \theta_2 - \theta_1 s]\, ds$$

In order to force $x_2(1)$ to be zero, it turns out that we must choose θ_1 and θ_2 so that $\theta_1 + \theta_2 - \theta_1 t$ switches sign at $t = \frac{1}{2}$, and in order that $x_1(1)$ be positive, this switching must be from positive to negative. Thus, we may set $\theta_1 = 2$, $\theta_2 = -1$, and obtain finally that

---

[53] See Coddington and Levinson, Ref. 57.

$$x_1(1) = 2 = C_p^3 \int_0^1 [1-s]\,[1-2s]^{\frac{3}{p-3}}\,\mathrm{sgn}[1-2s]\,ds$$

which can be integrated and solved for C_p to give

$$C_p = \left[4\left(2 + \frac{3}{p-3}\right)\right]^{\frac{1}{3}}$$

Using the above-stated values of θ_1 and θ_2, the following results are also obtained:

$$u_p(t) = C_p\,|1-2t|^{\frac{1}{p-3}}\,\mathrm{sgn}[1-2t]$$

$$\|u_p\|_\infty = C_p \underset{t\,\epsilon\,T}{\text{ess. sup.}} |1-2t|^{\frac{1}{p-3}} = C_p$$

$$\|u_p\|_p = C_p\left[2 + \frac{3}{p-3}\right]^{-\frac{1}{p}}$$

These results are plotted in Figs. 6.1(a) and 6.1(b), from which it can be seen that $u_p(t)$ approaches the function

$$\bar{u}(t) = \begin{cases} 2 & 0 \le t < \frac{1}{2} \\ -2 & \frac{1}{2} < t \le 1 \end{cases}$$

as p approaches infinity. That this is indeed the optimal control for the nonlinear problem can be shown by working the inverse problem (i.e., time optimal problem) using the constraint

$$|u(t)| \le 2 - \epsilon$$

and noting that for any $\epsilon > 0$, the goal $[2, 0]^T$ can not be achieved in a time $t_1 \le 1$. We have thus shown a control with peak amplitude 2 which satisfies the desired conditions, and have shown that no control with smaller peak amplitude can accomplish this, which constitutes a proof of the optimality of $\bar{u}(t)$.

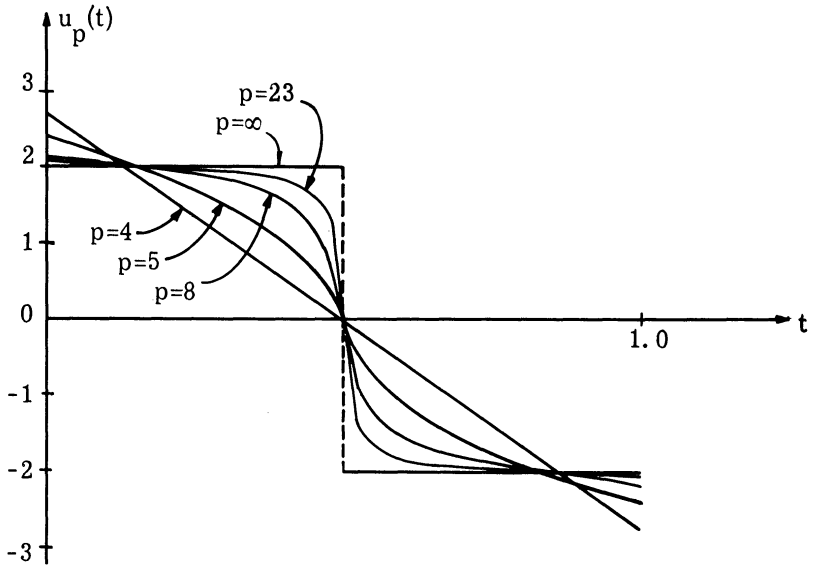

Fig. 6.1(a). The optimal control $u_p(t)$ for the related nonlinear problem for various values of p, for Example 12.

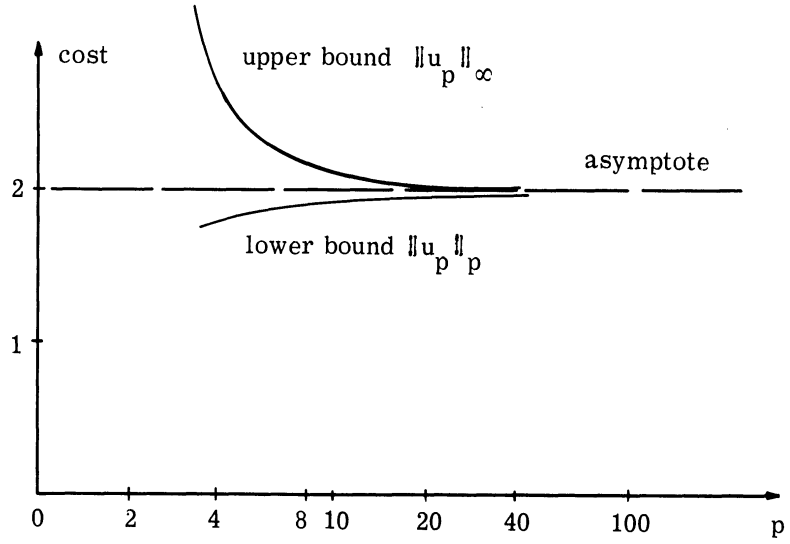

Fig. 6.1(b). Bounds on the optimal peak amplitude provided by $\|u_p\|_\infty$ and $\|u_p\|_p$ for Example 12.

Note that this use of the inverse (i.e., time-optimal) problem to prove optimality may or may not work in a given case. For example, if the final time, t_1, had been

specified at a time for which the minimum peak amplitude problem had no minimum time inverse [i. e. , a time such as t_{1d} or t_{1e} or t_{1f} in Fig. 4. 1(a)] then the minimum time problem would indicate that the desired result could be achieved with the given limit on $|u|$, and in fact at a final time considerably less than the given final time. This would not disprove the optimality of the limiting $\underline{u}(t)$; it would merely be an inconclusive test.

Example 13:

In this example also, a system has been chosen for which the state $\underline{x}(t)$ can be written out as a known functional of $\underline{u}(t)$. The system in this case is

$$\dot{x} = -1 + e^{-x}u \qquad x(0) = 0 \qquad x(1) = \log_e 2$$

for which it can be shown that, for the given initial value,

$$x(t) = \log_e \left[e^{-t} + \int_0^t e^{-t+s}\, u(s)ds \right] \tag{6.12}$$

so long as the expression in parentheses remains positive. Proceeding as above with the related nonlinear problem, we have as necessary conditions for optimality that u(t), x(t), and $\psi(t)$ must satisfy

$$\dot{x} = -1 + e^{-x}u \qquad x(0) = 0 \qquad x(1) = \log_e 2$$

$$\dot{\psi} = \psi\, e^{-x}u$$

$$u(t) \text{ must maximize } H = \psi_0 |u|^p - \psi + \psi\, e^{-x}u$$

where ψ_0 is a nonpositive constant and ψ must be an absolutely continuous function of t which is never equal to zero on the interval $[0,1]$. (This is merely a statement of the necessary conditions for optimality given in Section 6. 3, appropriately specialized to this problem.) The maximization of H with respect to u can be carried out by setting $\frac{\partial H}{\partial u} = 0$, and then solving for u, which gives

$$u(t) = \left[\frac{1}{-\psi_0 p} \right]^{\frac{1}{p-1}} \left| \psi(t)\, e^{-x(t)} \right|^{\frac{1}{p-1}} \operatorname{sgn} \psi(t)$$

$$= C_p \left| \psi(t)\, e^{-x(t)} \right|^{\frac{1}{p-1}} \operatorname{sgn} \psi(t)$$

All of these conditions are satisfied by

$$u_p(t) = C_p\, e^{\frac{t-1}{p-1}} \qquad C_p = \frac{q(2 - e^{-1})}{1 - e^{-q}}$$

$$x(t) = \log_e \left[e^{-t} + \frac{1}{q} C_p\, e^{-t-\frac{1}{p-1}} (e^{tq} - 1) \right]$$

$$\psi(t) = e^{-1} + (2 - e^{-1}) \left(\frac{e^{tq} - 1}{e^{q} - 1} \right)$$

where, as always, $q = \frac{p}{p-1}$. For this control $u_p(t)$, it is also easily shown that

$$\|u_p\|_\infty = C_p \qquad \|u_p\|_p = C_p \left[\frac{1 - e^{-q}}{q} \right]^{\frac{1}{p}} .$$

These results are plotted in Figs. 6.2(a) and 6.2(b). As p approaches ∞, $u_p(t)$ approaches the function

$$\bar{u}(t) = \frac{2 - e^{-1}}{1 - e^{-1}} = \frac{2e - 1}{e - 1} \qquad 0 \leq t \leq 1$$

which, upon substitution into Eq. 6.12, is easily shown to satisfy the desired final condition. Furthermore, it can be shown by means of the inverse problem (see Chapter IV) that this $\bar{u}(t)$ is indeed the minimum peak amplitude control, by the same reasoning as was used in Example 12. Alternatively, it could be argued that since the function $v = \log_e(w)$ has a unique inverse $w = e^v$ over the ranges involved here, the value $x(1) = \log_e 2$ can be achieved only if the bracketed expression in Eq. 6.12 equals 2 at $t = 1$. Linear systems optimization techniques can then be applied to the bracketed expression to show that $\bar{u}(t)$ is indeed optimal.

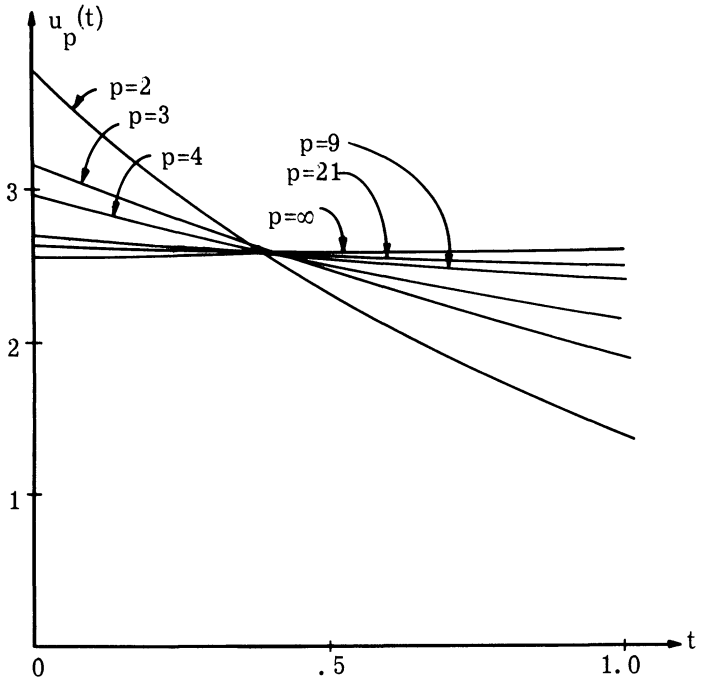

Fig. 6.2(a). The optimal control $u_p(t)$ for the related nonlinear problem for various values of p, for Example 12.

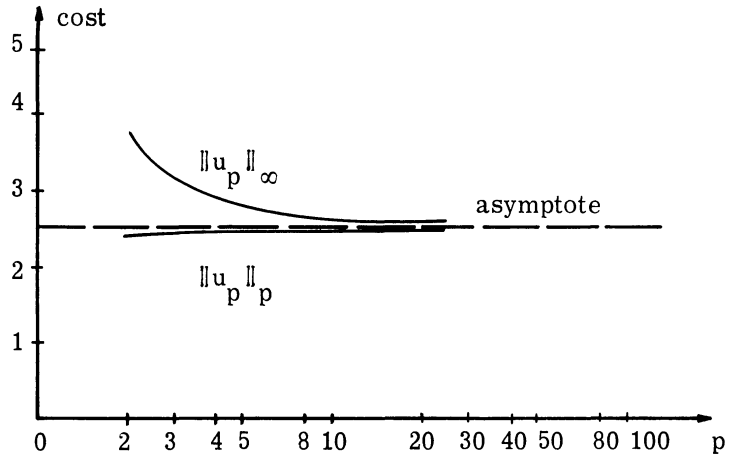

Fig. 6.2(b). Bounds on the optimal peak amplitude provided by $\|u_p\|_\infty$ and $\|u_p\|_p$ for Example 13.

CHAPTER VII

COMPUTATIONAL ALGORITHMS AND EXAMPLES

7.1 Preliminary Discussion

For proper linear systems (defined in Section 3.6) the problem of actually determining the numerical values of the minimum peak amplitude control in a given case is greatly simplified by Theorem 3.2, since, according to this theorem, we need determine only the direction of the constant k-component vector $\bar{\underline{c}}$ and the value of the cost $\bar{C}$ (which is easily determined once $\bar{\underline{c}}$ is known) in order to specify completely the optimal control almost everywhere on T, by means of Eq. 3.12.

The first computational algorithm discussed here involves the determination of $\bar{\underline{c}}$ and $\bar{C}$ for proper linear systems by steepest-descent techniques in k-dimensional Euclidean space. Before discussing this algorithm in detail, we make some preliminary observations and definitions to lay the theoretical groundwork for the algorithm: Let

$$\underline{h}(\underline{c}) = \int_T V(t_1, t) \frac{V^T(t_1, t)\,\underline{c}}{|V^T(t_1, t)\,\underline{c}|}\, dt \tag{7.1}$$

$$\cos\theta = \frac{(\underline{g}, \underline{h}(\underline{c}))}{|\underline{g}|\ |\underline{h}(\underline{c})|} \tag{7.2}$$

$$E = \frac{1}{|\underline{g}|}\, |\underline{g} - C(\underline{c})\, \underline{h}(\underline{c})| \tag{7.3}$$

$$\text{grad } E = \left[\frac{\partial E}{\partial c_1}, \ldots, \frac{\partial E}{\partial c_k}\right]^T \tag{7.4}$$

where $C(\underline{c})$ is chosen so as to minimize the error E for the given $\underline{g}$ and $\underline{h}(\underline{c})$. [This turns out to be

$$C(\underline{c}) = \frac{(\underline{g}, \underline{h}(\underline{c}))}{|\underline{h}(\underline{c})|^2} \tag{7.5}$$

as can be shown by setting $\frac{\partial E}{\partial C} = 0$ and solving for $C(\underline{c})$.] Using this value of $C(\underline{c})$, Eq. 7. 3 can then be reduced to

$$E = [\, 1 - \cos^2 \theta \,]^{\frac{1}{2}} \tag{7.6}$$

It is assumed throughout this section that $|\underline{g}| > 0$.

The set of points obtained by evaluating Eq. 7. 1 for all possible unit vectors $\underline{c}$ in E^k is the boundary ∂S_1 of the set S_1 defined in the proof of Theorem 3. 1, since these points are precisely those obtained by substituting into Eq. 3. 7 all controls of the form given by Eq. 3. 12 with $\bar{C}$ set equal to unity. (We need consider only unit vectors $\underline{c}$ because $\underline{h}(\underline{c})$ is independent of the magnitude of $\underline{c}$, so long as $\underline{c} \neq \underline{0}$.)

We now state certain results that are useful in the discussion of the convergence of the computational algorithms: The set of points ∂S_1 with the k - dimensional Euclidean metric, and the surface ∂U_k of the unit hypersphere (centered on the origin) of E^k with the same metric, are metric spaces. Concerning these spaces we have the following

Lemma: Eq. 7. 1 defines a continuous mapping of ∂U_k onto ∂S_1 for systems which are proper on the interval T.

Proof: The mapping is onto because every point in ∂S_1 is attained by means of an optimal control with cost $\bar{C}$ equal to unity [see part b) of the proof of Theorem 3. 1], and for proper systems every optimal control corresponds to a unit vector $\underline{c}$ (see Theorem 3. 3). The continuity of this mapping will have been shown if we can show that for every $\underline{c}$ and the corresponding $\underline{h}(\underline{c})$, and for every neighborhood $N_\epsilon = [\underline{h}: \underline{h} \in \partial S_1, \ |\underline{h} - \underline{h}(\underline{c})| < \epsilon \,]$ of $\underline{h}(\underline{c})$, there is a neighborhood $N_\delta = [\underline{d}: \underline{d} \in \partial U_k, \ |\underline{d} - \underline{c}| < \delta \,]$ of $\underline{c}$ such that every point of N_δ maps into N_ϵ. That this is so can best be shown graphically. Consider a typical set S_1 for a two-dimensional problem, as shown in Fig. 7. 1a). Consider the unit vector $\underline{c}$ and its image $\underline{h}(\underline{c})$. By Theorem 3. 2, $\underline{c}$ is an outward normal to ∂S_1 at $\underline{h}(\underline{c})$. The neighborhood N_ϵ of $\underline{h}(\underline{c})$ contains other points $\underline{h}$ of ∂S_1, each with a corresponding normal direction. Because the system is proper on the interval T, the set S_1 has no "flats" (see Theorem 3. 3). Thus these normals are not colinear, but lie in some (nondegenerate) cone, the limits of which are shown by $\underline{c}'$ and $\underline{c}''$ in Fig. 7. 1a). To find a satisfactory neighborhood N_δ of $\underline{c}$ we therefore need only choose a δ small enough so

that N_δ lies completely in the interior of the stated cone; i. e., such that neither $\underline{c}'$ nor $\underline{c}''$ lie in N_δ. Such a neighborhood will always map inside N_ϵ. This argument goes through whether or not the point $\underline{h}(\underline{c})$ is at or near a "corner" of S_1, as is shown by Fig. 7. 1b). Furthermore, the same arguments can be generalized to k- dimensional space.

Q. E. D.

Because $\underline{h}(\underline{c})$ is continuous in $\underline{c}$, and because $|\underline{h}(\underline{c})| > 0$ for all $\underline{c}$ [see part b) of the proof of Theorem 3. 1], the scalar functions $\cos\theta$, $C(\underline{c})$, and E are also continuous functions of $\underline{c}$ (with the appropriate Euclidean metric being understood in each case). This follows because continuous functions of continuous functions are continuous.

It can readily be seen from Eq. 7. 1 that for $\underline{c} = \overline{\underline{c}}$, where $\overline{\underline{c}}$ is the unit normal corresponding to the goal $\underline{g}$, $\underline{h}(\underline{c})$ will be colinear with $\underline{g}$, which implies that $\cos\theta = 1$, $E = 0$, and

$$C(\overline{\underline{c}}) = \frac{|\underline{g}|}{|\underline{h}(\overline{\underline{c}})|} = \overline{C}$$

On the basis of all this we can conclude that if the computer program generates a sequence of unit vectors $\underline{c}_i$ which converge to the unit vector $\overline{\underline{c}}$ (in the Euclidean metric on E^k), then the corresponding sequences in $C(\underline{c})\underline{h}(\underline{c})$, $\cos\theta$, $C(\underline{c})$, and E will converge to $\underline{g}$, 1, $\overline{C}$, and zero, respectively (with the Euclidean metric understood in each case). The convergence of all these sequences follows from the continuity of the various mappings involved (see Theorem 13. B, Simmons, Ref. 80, p. 76).

Conversely, it is easy to show, by arguments similar to those used above, that if we have a sequence of values of E which converges to zero, the corresponding $C(\underline{c})$, $\cos\theta$, and $C(\underline{c})\underline{h}(\underline{c})$ converge to $\overline{C}$, 1, and $\underline{g}$ respectively. This fact underlies the computer algorithms to be discussed, which approximate the optimal control by choosing a sequence of vectors $\underline{c}_i$ such that the error E is successively reduced at each step. We note in passing that an alternative approach would be to implement Eq. 3. 13 by seeking to maximize the expression shown there. This approach has much in common with the one used here, when the details of the two methods are compared, but it has

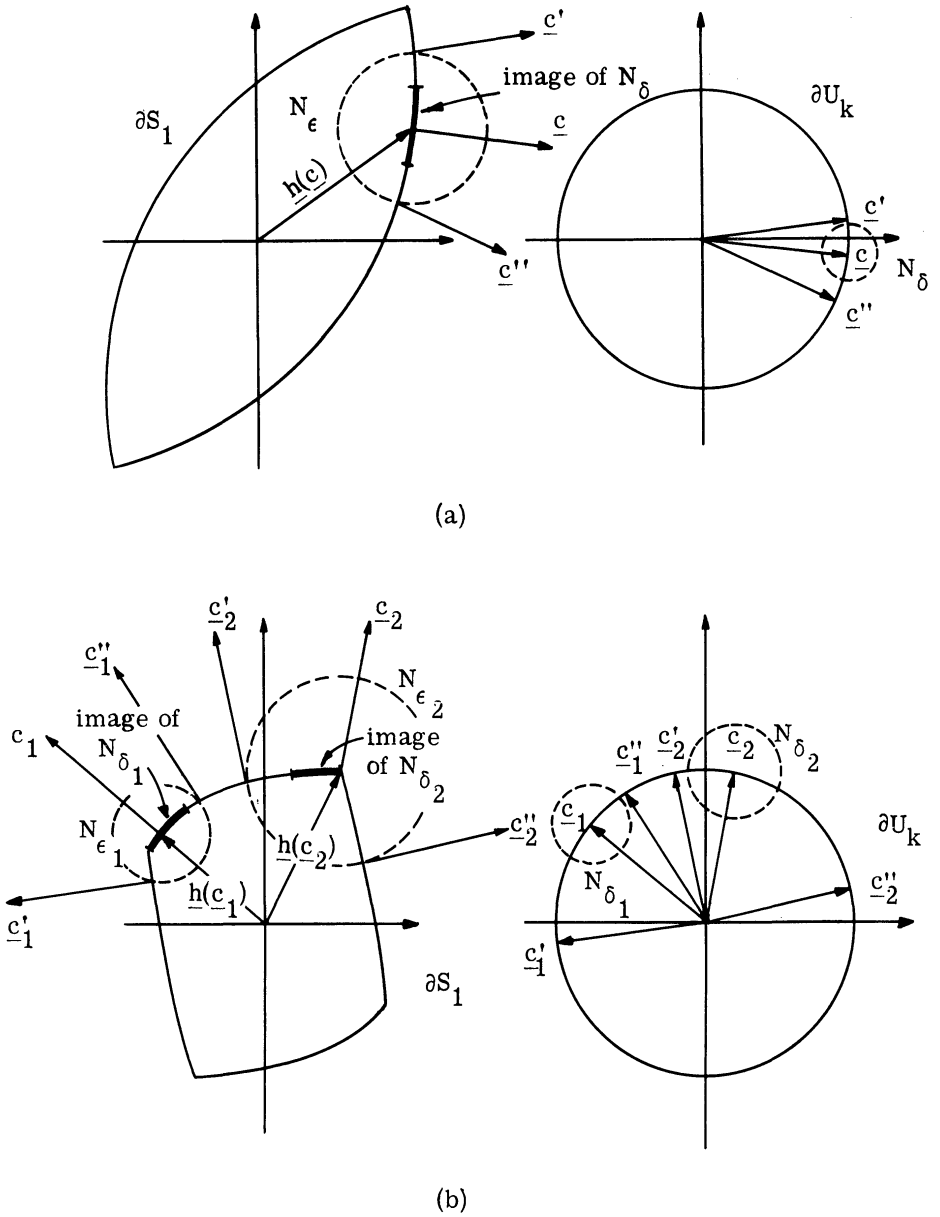

Fig. 7.1 Constructions to show the continuity of the mapping of ∂U_k onto ∂S_1.

the disadvantage that the maximum value of the expression is not known, so that there is no convenient way to tell how close one is to the maximum at any given stage of the process. In contrast, when we seek to minimize E, we of course know the absolute minimum value to be zero, so that we can readily supply an error criterion to the computer to terminate the iteration when E is sufficiently small.

However, this advantage of the method used here is more apparent than real, since there is no simple relationship between the size of E and the amount that $C(\underline{c})$ deviates from $\bar{C}$. If $\underline{g}$ happened to be in a direction such that in the vicinity of this direction a small change in the direction of $\underline{h}(\underline{c})$ produced a relatively large change in the magnitude of $\underline{h}(\underline{c})$, it could happen that an error of, for example, 2 percent could result in a $C(\underline{c})$ which differed from $\bar{C}$ by 5 percent, or even more, in either direction. To illustrate this point, we note that $C(\underline{c})$, as determined from Eq. 7.5, is the ratio of the magnitude of projection of $\underline{g}$ in the direction of $\underline{h}(\underline{c})$ to the magnitude of $\underline{h}(\underline{c})$. A small change in the direction of $\underline{h}(\underline{c})$ will not change the projection very much, but if it changes the magnitude of $\underline{h}(\underline{c})$ considerably, the stated result will take place. This effect comes about because of the departure of the set S_1 from a perfect hypersphere, since if S_1 were a hypersphere there would be no change in the magnitude of $\underline{h}(\underline{c})$ as its direction changed. Unless the shape of ∂S_1 is known in the vicinity of $\underline{g}$ (and it usually is not, in a typical problem) it is not even possible to estimate the "error amplification factor" that can result. There is no simple way to avoid this difficulty (whatever algorithm is used), and the usual approach is to set the allowable error limit rather small, as a "safety factor" against such occurrences.

7.2 An Algorithm for Proper Systems

We wish to approximate, by an iterative process, that value $\bar{\underline{c}}$ of $\underline{c}$ (or, more precisely, any value $\bar{\underline{c}}$ of $\underline{c}$) for which $\underline{h}(\underline{c})$, as defined by Eq. 7.1, is colinear with the goal $\underline{g}$. Since for all $\underline{c}$

$$(\underline{c}, \underline{h}(\underline{c})) = \int_T |V^T(t_1, t)\,\underline{c}|\,dt > 0$$

we know that $\bar{\underline{c}}$ must have a positive inner product with $\underline{g}$, and hence must lie in the hyperhemisphere of ∂U_k of vectors having positive inner products with $\underline{g}$. A good first guess for $\bar{\underline{c}}$ is simply the unit vector colinear with $\underline{g}$; i.e., $\frac{\underline{g}}{|\underline{g}|}$. This is the initial value used in the algorithm.

Having made a choice of $\underline{c}$, we compute $\underline{h}(\underline{c})$ from Eq. 7.1 and then determine $\cos\theta$, $C(\underline{c})$, and E from Eqs. 7.2, 7.5, and 7.6. If E is sufficiently small (i.e., less than the preset limit), computation stops and the results are printed out. Otherwise,

the value of $\underline{c}$ is changed to a new value $\underline{c}'$ and the whole process is repeated. The key issue is how to select the new value $\underline{c}'$. This decision can be broken down into two steps:

a) determine the angular direction in which $\underline{c}$ should be shifted, and

b) decide how large an angular shift to carry out.

This algorithm shifts $\underline{c}$ in the direction which tends to reduce E the fastest, as indicated by the gradient of E evaluated at $\underline{h}(\underline{c})$. Where this gradient is zero (such as at a "corner" of S_1, where a change in the normal vector $\underline{c}$ may result in no change in $\underline{h}(\underline{c})$, because of the nonuniqueness of the normal at such points), the angular shift in $\underline{c}$ is carried out in that direction which, if carried out on $\underline{h}(\underline{c})$, would rotate $\underline{h}(\underline{c})$ into $\underline{g}$. (This, by the way, is the direction which would be called for if the process implied by Eq. 3. 13 were implemented using a gradient technique.)

As to the amount of the shift, there is no simple best answer. Too small a shift may result in an excessive number of iterations being required, while too large a shift may cause oscillatory behavior and nonconvergence of the algorithm. The approach used here is as follows: The initial angular shift is chosen as $K\theta$, where K is a constant set initially by the user of the algorithm and θ is defined from Eq. 7. 2 as

$$\theta = \cos^{-1}(\cos\theta); \quad 0 \leq \theta \leq \pi$$

The $\underline{h}(\underline{c}')$ corresponding to this new $\underline{c} = \underline{c}'$ is computed, and from it the new value of E. If the new value of E is less than the previous one, then computation proceeds as indicated above. If it is <u>not</u>, then this $\underline{h}(\underline{c}')$ and $\underline{c}'$ are discarded and a $\underline{c}'$ with half of the rotation used previously is used instead. Also, the value K is cut in half, for use in succeeding iterations. If the resulting error is less than the original error, then the next value of $\underline{c}$ is chosen in the negative gradient direction (or, for systems with "corners", in the direction required to shift $\underline{h}(\underline{c})$ into $\underline{g}$), as indicated above. Otherwise, the angular shift and the value of K are cut in half again, and so on. Eventually, of course, the process of successively halving the angular shift must end up with a shift small enough that the gradient dominates the higher order curvatures, so that the process of successive halving eventually terminates and normal computation resumes.

In this algorithm there is no guarantee that there will be convergence to within an acceptable error in any given number of steps, but the results on all problems tried thus far have been good, convergence to within an error E of 0.1 percent having been obtained in 10 iterations or less (often much less).

A simplified diagram of this algorithm is shown in Fig. 7.2, and a complete listing in the MAD language is given on pages 120-123. We offer a few notes of explanation in connection with this algorithm: It is possible to derive analytic expressions for grad E, but these expressions involve improper integrals (unbounded integrands) if the set S_1 has "corners". Therefore, the approach used here is to approximate grad E by

$$100 \begin{bmatrix} E_1 - E \\ \\ E_k - E \end{bmatrix}$$

where E is the error corresponding to the unit vector $\underline{c} = [c_1, .., c_{i-1}, c_i, c_{i+1}, \ldots, c_k]^T$ and E_i, i = 1,..,k, is the error corresponding to the unit vector $\underline{c}_i [c_1, \ldots, c_{i-1}, c_i + .01, c_{i+1}, \ldots, c_k]^T$, normalized to unit length.

External functions are used to compute the matrices $\mathbf{X}(t_1, t)$, B(t), and G(t). This allows the user great flexibility in the use of the program. Depending on the problem, these external functions may be used to insert constant matrices, compute the matrices from given functions, or, in the case of $\mathbf{X}(t_1, t)$, generate them by solving a system of differential equations. Some of the subroutines that have been used store all the values computed during the first entry to the subroutine, so that on succeeding iterations no computations are needed. In this algorithm, the matrix D is required to be of the form

$$\text{k rows} \left\{ \begin{bmatrix} 1 & 0 & 0 \ldots 0 & 0 \ldots 0 \\ 0 & 1 & 0 \ldots 0 & 0 \ldots 0 \\ 0 & 0 & 1 \ldots 0 & 0 \ldots 0 \\ \\ 0 & 0 & 0 \ldots 0 & 0 \ldots 0 \end{bmatrix} \right.$$

k columns; n - k columns

```
                    PROGRAM FOR PROPER SYSTEMS
$   COMPILE MAD, EXECUTE, DUMP, PUNCH OBJECT, PRINT OBJECT
    START  READ DATA
           PRINT FORMAT HEADA
           VECTOR VALUES HEADA = $
          1 44H1PROBLEM. CHOOSE THE INPUTS U(1)...U(R) AS F,
          2 44HUNCTIONS OF TIME SO AS TO FORCE THE GIVEN SY,
          3 44HSTEM FROM ITS INITIAL STATE X(1)...X(N) =   /
          4 44H ALPHA(1)...ALPHA(N) AT INITIAL TIME TZERO T,
          5 44HO THE DESIRED GOAL X(1)...X(K) = BETA(1)...B,
          6 44HETA(K) AT TIME T1 AND SO THAT THE P-NORM OF /
          7 44H THE AMPLITUDE OF THE VECTOR GU IS MINIMUM. ,
          8 44HG IS ANY INVERTIBLE MATRIX FUNCTION OF TIME./*$
           PRINT FORMAT HEADB
           VECTOR VALUES HEADB = $
          1 44HOINPUT DATA REQUIRED. SYSTEM ORDER N, NUMBER,
          2 44H OF STATE VARIABLES TO BE CONTROLLED K, NUMB,
          3 44HER OF CONTROL INPUTS R, NUMBER OF INTEGRA-  /
          4 44H TION INTERVALS S, INITIAL TIME TZERO, FINAL,
          5 44H TIME T1, ERROR CRITERION ERB, ITERATION GAI,
          6 44HN CONSTANT CAPK, MAXIMUM NUMBER OF ITER-    /
          7 44H ATIONS ITR, PRINTU (0 RESULTS IN NO PRINTOU,
          8 44HT OF U. J RESULTS IN U(1)...U(R) BEING PRINT,
          9 44HED OUT EVERY J-TH INTEGRATION INTERVAL.),   /*$
           PRINT FORMAT HEADC
           VECTOR VALUES HEADC = $
          1 44H INITIAL STATE ALPHA(1)...ALPHA(N), GOAL BET,
          2 44HA(1)...BETA(K), AND NORM EXPONENT P.  P .L. ,
          3 44H2 IMPLIES P = INFINITY. EXTERNAL FUNCTIONS  /
          4 44H FND., CON., AND WGT.  MUST ALSO BE PROVIDED,
          5 44H FOR COMPUTATION OF THE FUNDAMENTAL, CONTROL,
          6 44H, AND WEIGHTING MATRICES.                   /*$
           PRINT FORMAT HEADD
           VECTOR VALUES HEADD = $
          1 44HOINTERPRETATION OF OUTPUT DATA. A, BEE, G, A,
          2 44HND GINV ARE THE SYSTEM, CONTROL, WEIGHTING, ,
          3 44HAND INVERSE WEIGHTING MATRICES, RESPECTIVELY/
          4 44H IF U IS PRINTED OUT, THE COLUMNS BELOW INDI,
          5 44HCATE THE INTEGRATION INDEX, THE CORRESPONDIN,
          6 44HG TIME, AND THE ELEMENTS OF U/CAPC, RESP.   /
          7 44H CAPC = PEAK AMPLITUDE. IC = NO. OF ITERATIO,
          8 44HNS. FLAG = 1, CONVERGENCE. FLAG = 2, TOO MAN,
          9 44HY ITERATIONS. FLAG = 3, BOTH.               /*$
           WHENEVER P .GE. 2.0
           PWR =-(P - 2.)/(2.*P-2.)
           OTHERWISE
           PWR =-.5
           END OF CONDITIONAL
           NN=N*N
           KK=K*K
           KN = K*N
           KR=K*R
           KRS = KR*S
           DT=(T1-TZERO)/S
           W0=DT*(2.+25.*DT)/48.
           W1=DT*(70.-14.*DT)/48.
           W2=DT*(48.-11.*DT)/48.
           INDEX=PRINTU
           IC = 0
           FLAG=0
           THROUGH BEN, FOR Q=0,1,Q.G.S
           T=T1-Q*DT
           EXECUTE FND.(T1,T,DT,PHI,A,N,Q)
           EXECUTE CON.(T,BEE,N,R,Q,BN)
           EXECUTE WGT.(T,G,GINV,R,IC,Q,GN,S,FLAG,PRINTU)
           WHENEVER FLAG .G. 0, TRANSFER TO BEN
           WHENEVER BN .G. 0
           THROUGH DOLLY, FOR I=1,1,I.G.K
           M = I*N - N
           THROUGH DOLLY, FOR J=1,1,J.G.R
           Y = KRQ + M + J
           PSI(Y) = 0.
           THROUGH DOLLY, FOR L = 1, 1, L .G. N
DOLLY      PSI(Y) = PSI(Y) + PHI(M+L)*BEE(L*R-R+J)
           OTHERWISE
           THROUGH EVE, FOR I = 1, 1, I .G. K
           M = I*N - N
           Y = KRQ + I*R - R
           THROUGH EVE, FOR J=1,1,J.G.R
EVE        PSI(Y+J) = PHI(M+J)
           END OF CONDITIONAL
           WHENEVER GN .G. 0
           THROUGH ADA, FOR I=1, 1, I .G. K
           Y = KRQ + I*R - R
           THROUGH FRAN, FOR J=1, 1, J .G. R
           STOR(J)=0.
           THROUGH FRAN, FOR L=1, 1, L .G. R
FRAN       STOR(J) = STOR(J) + PSI(Y+L)*GINV(L*R-R+J)
           THROUGH ADA, FOR J=1, 1, J .G. R
ADA        PSI(Y+J) = STOR(J)
           END OF CONDITIONAL
BEN        CONTINUE
           WHENEVER FLAG .G. 0, TRANSFER TO START
           ALPHA(0) = 0.
           BETA(0)=0.
           GAMMA(0)=0.
           THROUGH ALEX, FOR I=1,1,I.G.N
ALEX       ALPHA(0) = ALPHA(0) + ALPHA(I)*ALPHA(I)
           THROUGH GREG, FOR I=1,1,I.G.K
           GAMMA(I)=BETA(I)
           M=I*N-N
           THROUGH HARRY, FOR J = 1, 1, J.G.N
HARRY      GAMMA(I)=GAMMA(I)-PHI(M+J)*ALPHA(J)
           BETA(0)=BETA(0)+BETA(I)*BETA(I)
GREG       GAMMA(0)=GAMMA(0)+GAMMA(I)*GAMMA(I)
           ALPHA(0) = SQRT.(ALPHA(0))
           BETA(0)  = SQRT.(BETA(0))
           GAMMA(0) = SQRT.(GAMMA(0))
```

```
        PRINT RESULTS ALPHA(0),BETA(0),GAMMA(0),ALPHA(1)...
       1ALPHA(N),BETA(1)...BETA(K),GAMMA(1)...GAMMA(K)
        WHENEVER (GAMMA(0).L..0001*BETA(0)).OR.(GAMMA(0).L.
       1.0001*ALPHA(0))
        PRINT COMMENT $ DEGENERATE PROBLEM$
        TRANSFER TO START
        END OF CONDITIONAL
        THROUGH CLARA, FOR I=1, 1, I .G. K
CLARA   C(I) = GAMMA(I)
REPEAT  EXECUTE ZERO.(H(0)...H(K),PARTL(1)...PARTL(KK),
       1STOR(1)...STOR(K))
        THROUGH WALLY, FOR I = 1, 1, I .G. K
WALLY   C(0)=C(0)+C(I)*C(I)
        C(0)=SQRT.(C(0))
        THROUGH LARRY, FOR I=1,1,I.G.K
LARRY   C(I)=C(I)/C(0)
        THROUGH AL, FOR Q=0, 1, Q .G. S
        W=DT
        WHENEVER (Q.E.S).OR.(Q.E.0), W=W0
        WHENEVER (Q.E.S-1).OR.(Q.E.1), W=W1
        WHENEVER (Q.E.S-2).OR.(Q.E.2), W=W2
        T = T1 - Q*DT
        KRQ = KR*Q
        U(0) = 0.
        THROUGH BOB, FOR I = 1, 1, I .G. R
        U(I) = 0.
        M = KRQ + I - R
        THROUGH CAL, FOR J = 1, 1, J .G. K
CAL     U(I) = U(I) + C(J)*PSI(M+J*R)
BOB     U(0) = U(0) + U(I)*U(I)
        THROUGH DAN, FOR I=1, 1, I .G. K
        M = I*R - R
        Y= KRQ + M
        THETA(I) = 0.
        THROUGH ED, FOR J=1, 1, J .G. R
        USTOR(M+J) = U(J) + .01*PSI(Y+J)
ED      THETA(I) = THETA(I) + USTOR(M+J)*USTOR(M+J)
        THETA(I) = THETA(I).P.PWR
        THROUGH DAN, FOR J=1, 1, J .G. R
DAN     USTOR(M+J) = THETA(I)*USTOR(M+J)
        U(0) = U(0).P.PWR
        THROUGH FRANK, FOR I=1, 1, I .G. R
FRANK   U(I) = U(0)*U(I)
        WHENEVER FLAG.G.0 .AND.PRINTU.G.0
        WHENEVER INDEX .E. PRINTU
        INDEX=1
        WHENEVER GN .GE. 1
        EXECUTE WGT.(T,G,GINV,R,IC,Q,GN,S,FLAG,PRINTU)
        THROUGH CARL, FOR I=1,1,I.G.R
        STOR(I)=0.
        M=I*R-R
        THROUGH CARL, FOR J=1,1,J.G.R
CARL    STOR(I)=STOR(I)+GINV(M+J)*U(J)
        THROUGH DOUG, FOR I=1,1,I.G.R
DOUG    U(I)=STOR(I)
        END OF CONDITIONAL
        PRINT FORMAT OUTPUT,Q,T,U(1)...U(R)
        OTHERWISE
        INDEX=INDEX+1
        END OF CONDITIONAL
        END OF CONDITIONAL
        H(0) = 0.
        THROUGH IRA, FOR I=1, 1, I .G. K
        M = KRQ + I*R -R
        THETA(I) = 0.
        THROUGH IRA, FOR J=1, 1,  J .G. R
IRA     THETA(I) = THETA(I) + PSI(M+J)*U(J)
        THROUGH JIM, FOR I=1, 1, I .G. K
        H(I) = H(I) + W*THETA(I)
JIM     H(0) = H(0) + H(I)*H(I)
        H(0) = SQRT.(H(0))
        THROUGH JACK, FOR I=1, 1, I .G. K
        Y = I*K - K
        Z = I*R - R
        THROUGH KEN, FOR J=1, 1, J .G. K
        M = KRQ + J*R - R
        THETA(J) = 0.
        THROUGH KEN, FOR L=1, 1, L .G. R
KEN     THETA(J) = THETA(J) + PSI(M+L)*USTOR(Z+L)
        THROUGH JACK, FOR J=1, 1, J .G. K
JACK    PARTL(Y+J) = PARTL(Y+J) + W*THETA(J)
        THROUGH AL, FOR I=1, 1, I .G. K
        M = I*K - K
        THROUGH LEN, FOR J=1, 1, J .G. K
        Y = M+J
LEN     STOR(I) = STOR(I) + PARTL(Y)*PARTL(Y)
AL      STOR(I) = SQRT.(STOR(I))
        EXECUTE ZERO.(GAMH,THETA(1)...THETA(K))
        THROUGH MAX, FOR I=1, 1, I .G. K
        M = I*R - R
        THROUGH MAX, FOR J=1, 1,  J .G. K
        GAMH = GAMH + GAMMA(I)*H(I)
        CAPC = GAMH/(H(0)*H(0))
        HCOS = GAMH/(GAMMA(0)*H(0))
        ERR = SQRT.(1. - HCOS*HCOS)
MAX     THETA(I) = THETA(I) + GAMMA(J)*PARTL(M+J)
        THROUGH OTTO, FOR I=1, 1, I .G. K
        COS = THETA(I)/(GAMMA(0)*STOR(I))
OTTO    GRAD(I) = 100.*(SQRT.(1. - COS*COS) - ERR)
        GRAD(0) = 0.
        THROUGH PAUL, FOR I=1, 1, I .G. K
PAUL    GRAD(0) = GRAD(0) + GRAD(I)*GRAD(I)
        IC = IC + 1
        WHENEVER FLAG.G.0
        PRINT RESULTS CAPC, ERR, IC, FLAG, CAPK
        PRINT FORMAT GAMM,GAMMA(1)...GAMMA(K)
        PRINT FORMAT FINX,H(1)...H(K)
        PRINT FORMAT FINC,C(1)...C(K)
```

```
      TRANSFER TO START
      END OF CONDITIONAL
      WHENEVER ERR.LE.ERB,FLAG=1
      WHENEVER IC .G. ITR, FLAG = FLAG + 2
      WHENEVER IC = 1
      THROUGH BETSY, FOR I=1, 1, I.G. K
BETSY CP(I) = C(I)
      EP = ERR
      TRANSFER TO JUMP
      END OF CONDITIONAL
      WHENEVER ERR .G. EP
      THROUGH CHRIS, FOR I=1, 1, I.G. K
CHRIS C(I) = C(I) + CP(I)
      CAPK = 0.5* CAPK
      TRANSFER TO REPEAT
      END OF CONDITIONAL
      EP = ERR
      THROUGH DOTTIE, FOR I=1, 1, I.G.K
DOTTIE CP(I) = C(I)
JUMP  WHENEVER GRAD(0) .LE. .001
      CG = 0.
      CH = 0.
      THROUGH EDIE, FOR I=1, 1, I.G.K
      CG = CG + GAMMA(I)*C(I)
EDIE  CH = CH + GAMMA(I)*H(I)
      HCON = CG/CH
      D(0) = 0.
      THROUGH FERN, FOR I=1, 1, I.G.K
      D(I) = GAMMA(I) - HCON*H(I)
FERN  D(0) = D(0) + D(I)*D(I)
      HCON = CAPK*ERR/(HCOS*SQRT.(D(0)))
      THROUGH GAIL, FOR I=1, 1, I.G.K
GAIL  C(I) = C(I) + HCON*D(I)
      OTHERWISE
      HCON = CAPK*ERR/(HCOS*SQRT.(GRAD(0)))
      THROUGH JOAN, FOR I=1,1,I .G. K
JOAN  C(I) = C(I) - HCON*GRAD(I)
      END OF CONDITIONAL
      TRANSFER TO REPEAT
      DIMENSION ALPHA(10),BETA(10),GAMMA(10),PSI(10100),
     1H(10),PHI(100),THETA(10),BEE(100),U(10),PARTL(100),
     2GRAD(10),GINV(100),STOR(10),G(100),A(100),C(10),
     4CP(10),USTOR(100), D(10)
      EQUIVALENCE (PHI,PARTL), (BEE,USTOR)
      INTEGER K,N,R,I,J,L,M,Q,  Y,Z,FLAG,ITR,BN,KR,S, NN,
     1PRINTU,INDEX,IC ,KK,KN,GN,KRQ,KRS
      VECTOR VALUES GAMM=$7H0GAMMA ,10E12.4/(S7,10E12.4)*$
      VECTOR VALUES FINX=$7H0H      ,10E12.4/(S7,10E12.4)*$
      VECTOR VALUES FINC=$7H0C      ,10E12.4/(S7,10E12.4)*$
      VECTOR VALUES OUTPUT=$1H ,I6,10E12.4/(S19,9E12.4)*$
      END OF PROGRAM
```

TYPICAL DATA SET

```
$   DATA
    N=10,K=4,R=3,S=100,TZERO=0.,T1=1.,ERB=.001,CAPK=2.,
    ITR=10,G(1)=1.,0.,0.,0.,1.,-1.,0.,1.,2.,BEE(1)=0.,0.,0.,
    0.,0.,0.,1.,0.,0.,0.,0.,0.,0.,0.,0.,0.,0.,1.,0.,0.,0.,0.,
    0.,0.,0.,0.,0.,1.,1.,1.,1., P=1, PRINTU=10,ALPHA(1)=0.,
    0.,0.,0.,0.,0.,1.,1.,1.,1., P=1, PRINTU=10,ALPHA(1)=0.,2.,
    2.,0.,1.,1.,0.,0.,4.,0.,0.,25.,BETA(1)=10.,10.,20.,0.*
```

TYPICAL EXTERNAL FUNCTION FOR COMPUTING PHI

```
$   COMPILE MAD, PUNCH OBJECT, PRINT OBJECT
        EXTERNAL FUNCTION (T1X,TX,DTX,PHIX,AX,NX,QX)
        ENTRY TO FND.
        INTEGER IX,JX,MX,LX,YX,ZX,QX,NNX
        DIMENSION EXPA(100),STORX(10)
        NNX = NX*NX
        WHENEVER QX .E. 0
        PRINT COMMENT $0PHI IS COMPUTED FROM THE DIFFERENTIAL
       1 EQUATION USING FND&1001 AND  $
        PRINT RESULTS AX(1)...AX(NNX)
        EXECUTE ZERO.(EXPA(1)...EXPA(NNX),STORX(1)...
       1STORX(NX),PHIX(1)...PHIX(NNX))
        THROUGH FIDO, FOR  IX = 1, NX+1, IX .G. NNX
        PHIX(IX) = 1.
FIDO    EXPA(IX) = 1.
        THROUGH SPOT, FOR IX=1,1, IX .G. 12
        WX =-DTX/IX
        THROUGH LASSY, FOR JX=1, 1, JX .G. NX
        YX = JX*NX - NX
        THROUGH LAD, FOR MX =1, 1, MX .G. NX
        ZX = MX*NX - NX
        THROUGH LAD, FOR LX=1, 1, LX .G. NX
LAD     STORX(MX)= STORX(MX)+WX*AX(ZX+LX)*PHIX(YX+LX)
        THROUGH LASSY, FOR MX=1,1, MX .G. NX
        PHIX(YX+MX)=STORX(MX)
LASSY   STORX(MX) = 0.
        THROUGH SPOT, FOR JX = 1, 1, JX .G. NNX
SPOT    EXPA(JX)= EXPA(JX) + PHIX(JX)
        EXECUTE ZERO.(PHIX(1)...PHIX(NNX))
        THROUGH SNOOPY, FOR IX=1, NX+1, IX .G. NNX
SNOOPY  PHIX(IX) = 1.
        OTHERWISE
        THROUGH PLUTO, FOR IX=1, 1, IX .G. NX
        YX= IX*NX - NX
        THROUGH ROVER, FOR JX= 1, 1, JX .G. NX
        STORX(JX) = 0.
        THROUGH ROVER, FOR LX=1,1, LX .G. NX
ROVER   STORX(JX) = STORX(JX)+PHIX(YX+LX)*EXPA(LX*NX-NX+JX)
        THROUGH PLUTO, FOR JX = 1, 1, JX .G. NX
PLUTO   PHIX(YX+JX) = STORX(JX)
        END OF CONDITIONAL
        FUNCTION RETURN
        END OF FUNCTION
```

```
                TYPICAL EXTERNAL FUNCTION FOR COMPUTING G
             (IN THIS CASE THE CONSTANTS OF G ARE FED IN AS DATA)
$    COMPILE MAD, PUNCH OBJECT, PRINT OBJECT
           EXTERNAL FUNCTION (TZ,GZ,GINVZ,RZ,ICZ,QZ,GNZ,SZ,
          1  FLAZ,PRINTU)
           ENTRY TO WGT.
           INTEGER RZ,QZ,ICZ,GNZ,SZ,RRZ,IZ,JZ,LZ,FLAZ,PRINTU,
          1  MZ,YZ
           WHENEVER ICZ = 0
           WHENEVER QZ = 0
           RRZ = RZ * RZ
           THROUGH TABBY,FOR IZ=1, 1, IZ .G. RZ
           YZ = IZ * RZ - RZ
           THROUGH TABBY,FOR JZ=1, 1, JZ .G. RZ
    TABBY  GINVZ(YZ+JZ)= (GZ(YZ+JZ) + GZ(JZ*RZ - RZ + IZ))/2.
           MZ = BORDS.(RZ,RZ,GINVZ(1),DZ)
           PRINT RESULTS GZ(1)...GZ(RRZ)
            WHENEVER MZ .L. 1
           PRINT COMMENT $0G IS NOT INVERTIBLE. $
           QZ = SZ + 1
           FLAZ = 4
           OTHERWISE
           PRINT COMMENT $0G-INVERSE, COMPUTED USING WGT40000,
          1 IS THE CONSTANT R X R MATRIX $
           PRINT RESULTS GINVZ(1)...GINVZ(RRZ)
           GNZ= 4
           END OF CONDITIONAL
           END OF CONDITIONAL
           END OF CONDITIONAL
           FUNCTION RETURN
           END OF FUNCTION
```

```
                TYPICAL EXTERNAL FUNCTION FOR COMPUTING B
             (IN THIS CASE THE CONSTANTS OF B ARE FED IN AS DATA)
$    COMPILE MAD, PUNCH OBJECT, PRINT OBJECT
           EXTERNAL FUNCTION (TY,BEEY,NY,RY,QY,BNY)
           ENTRY TO CON.
           INTEGER NY,RY,QY,BNY,NRY
           WHENEVER QY .E. 0
           BNY = 3
           NRY = NY*RY
           PRINT COMMENT $0BEE IS THE CONSTANT N X R MATRIX  $
           PRINT RESULTS BEEY(1)...BEEY(NRY)
           END OF CONDITIONAL
           FUNCTION RETURN
           END OF FUNCTION
```

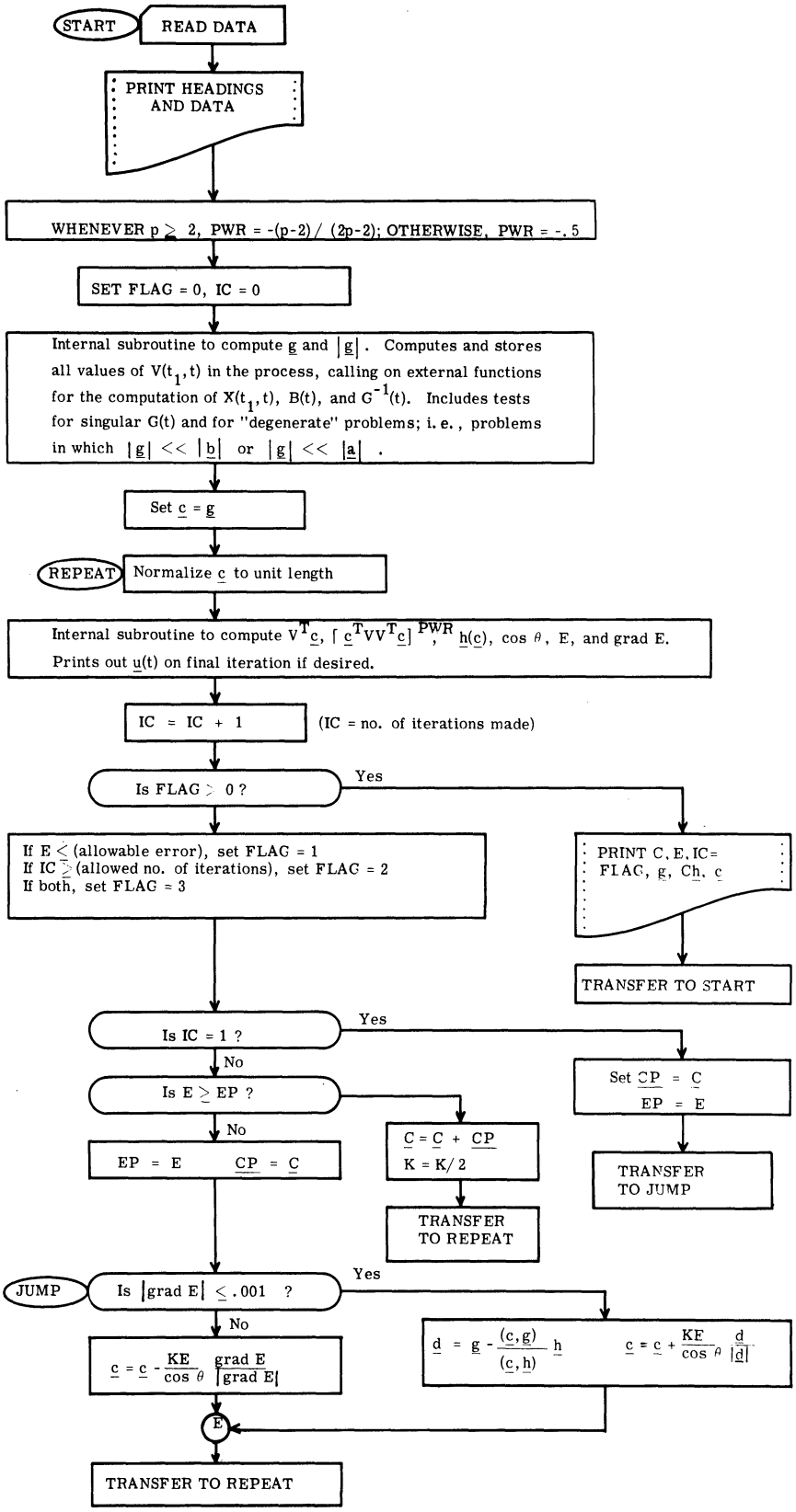

Fig. 7.2 Simplified diagram of computer program for proper systems.

That is, the first k components of the state vector are taken as the output vector, where k may take any value from 1 to n, inclusive.

The time interval T is divided into 100 subintervals for purposes of numerical integration, and a modified version of Simpson's rule is used to increase the accuracy of the integrations.

Because it required the addition of only one step for the program, the capability to solve the related linear problem (see Chapter V) was written into the program. Any value of $p \geq 2$ may be used.

Finally, we note that in the IBM 7090, for which the program was written, there is no provision for distinguishing between upper and lower case letters. This has necessitated certain changes in notation. A listing of the notation used throughout this work and the corresponding notation used in the program listing follows:

$\underline{a}$ ..... ALPHA	$\underline{g}$ ..... GAMMA	K ..... CAPK
$\underline{b}$ ..... BETA	$(\underline{g},\underline{h})$ ..... GAMX	$\underline{u}(t)$ ..... U
B(t) ..... BEE	G(t) ..... G	$V(t_1,t)$ .... PSI
$\underline{c}$ C	$G^{-1}(t)$ .... GINV	$\underline{x}(t)$ X
$C(\underline{c})$ ..... CAPC	grad E ... GRAD	$X(t_1,t)$ .... PHI
E ..... ERR	$\underline{h}(c)$ ..... H	

Fig. 7.3 shows an example of the convergence of this algorithm in a two-dimensional problem, with the initial value of K chosen as 2. From this figure we see that the first two iterations, while producing no change in $\underline{h}(\underline{c})$, nonetheless shift $\underline{c}$ toward the edge of the cone of normals corresponding to the corner. The third iteration not only shifts $\underline{c}$ out of this cone, but goes well beyond the goal --- so far beyond that the error is actually increased over that associated with $\underline{h}(\underline{c}_2)$. Thus, the value of K is cut in half and the normal $\underline{c}_4$ results. This gives an $\underline{h}(\underline{c})$ rather close to $\underline{g}$ in direction. The next iteration (with K now equal to 1) gives an even better result. Further iterations bring $\underline{h}(\underline{c})$ even closer to the direction of $\underline{g}$, but these can not readily be shown graphically. In this example, $C(\underline{c})$ is within 2.5 percent of $\overline{C}$ after the fifth iteration, whereas the initial error in $C(\underline{c})$ was 50 percent [both expressed as a percentage of $C(\underline{c})$]. The value of E after the fifth iteration is about 0.015.

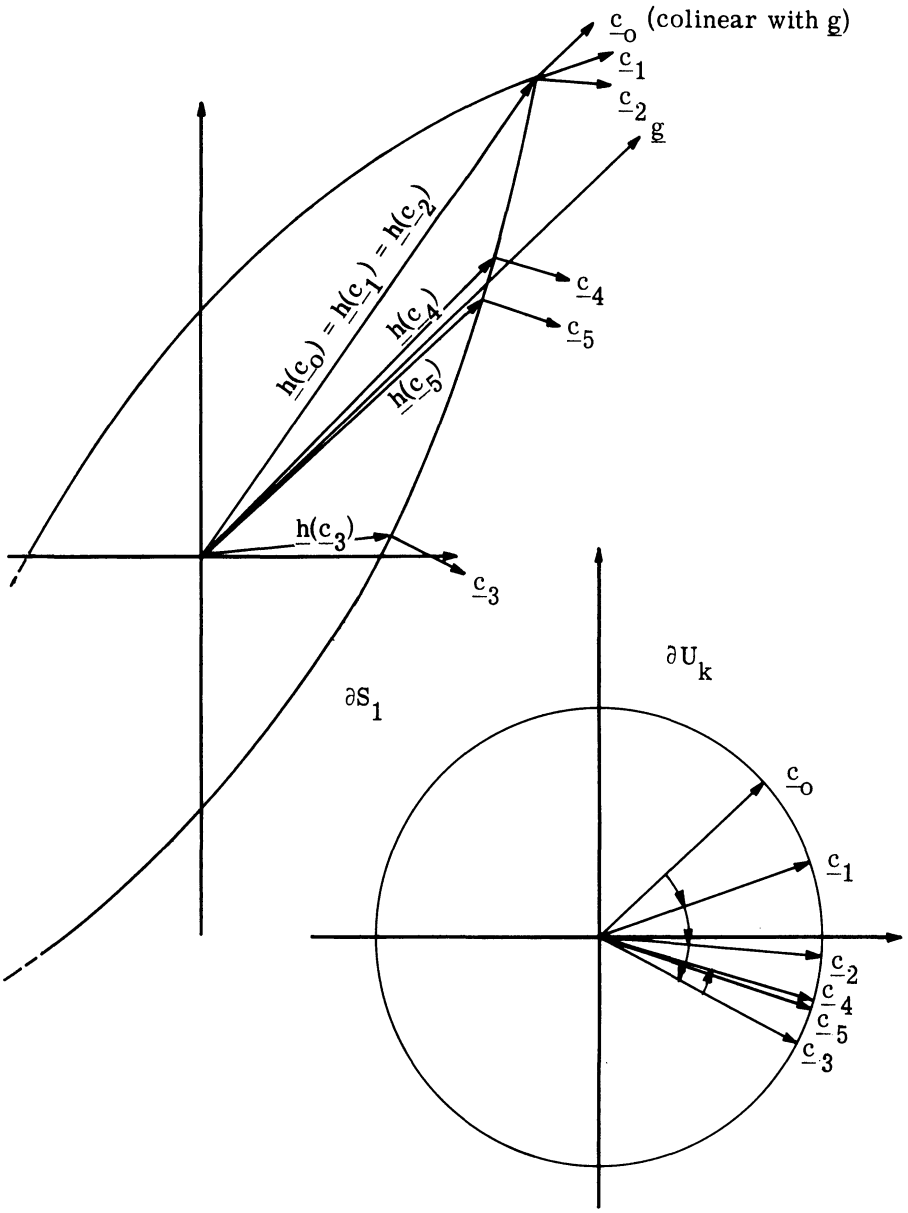

Fig. 7.3 Successive iterations of the computational algorithm for a 2-dimensional example.

7.3 Another Algorithm

A second algorithm, a simplified diagram of which is shown in Fig. 7.4, has been used with good results on systems for which the matrix

$$W(\underline{c}) = \int_T \frac{V(t_1, t)\, V^T(t_1, t)}{|V^T(t_1, t)\, \underline{c}|}\, dt$$

is bounded in norm[54] (uniformly in $\underline{c}$) and is invertible for all unit vectors $\underline{c} \in E^k$.

Lemma: A sufficient condition that $W(\underline{c})$ be bounded in norm and invertible is that there exist an $\epsilon > 0$ such that $|V^T(t_1, t)\underline{c}| \geq \epsilon$ for all unit vectors $\underline{c} \in E^k$ and for almost all $t \in T$. Here $V(t_1, t)$ must satisfy the boundedness conditions imposed in Chapters II and III.

Proof: That this is a sufficient condition, as claimed, is easily shown by noting that under this condition and the boundedness of $V(t_1, t)$, the quadratic form

$$\underline{d}^T W(\underline{c})\underline{d} = \int_T \frac{|V^T(t_1, t)\underline{d}|^2}{|V^T(t_1, t)\underline{c}|} dt$$

is positive and bounded for all bounded nonzero $\underline{d}$, which implies that $W(\underline{c})$ is not only invertible but positive definite. Q. E. D.

This lemma points directly to the relaxed sufficient condition given in the following:

Lemma: The matrix $W(\underline{c})$ is bounded in norm and invertible if a) $|V^T(t_1, t)\underline{c}| > 0$ for all unit vectors $\underline{c} \in E^K$ and for almost all $t \in T$, and b) the integral

$$\int_T \frac{dt}{|V^T(t_1, t)\underline{c}|}$$

exists for all such $\underline{c}$. (Here, as above, $V(t_1, t)$ is assumed to satisfy the conditions of Chapters II and III.)

Proof: The proof follows that the preceeding lemma in every respect except that in order to show that $W(\underline{c})$ is bounded in norm we must make use of the condition in part b) of this lemma to show that the quadratic form is bounded. Q. E. D.

---

[54] We require boundedness because the computer can not operate with infinitely-large numbers. The norm of a matrix is defined as in Section 3. 2 (footnote).

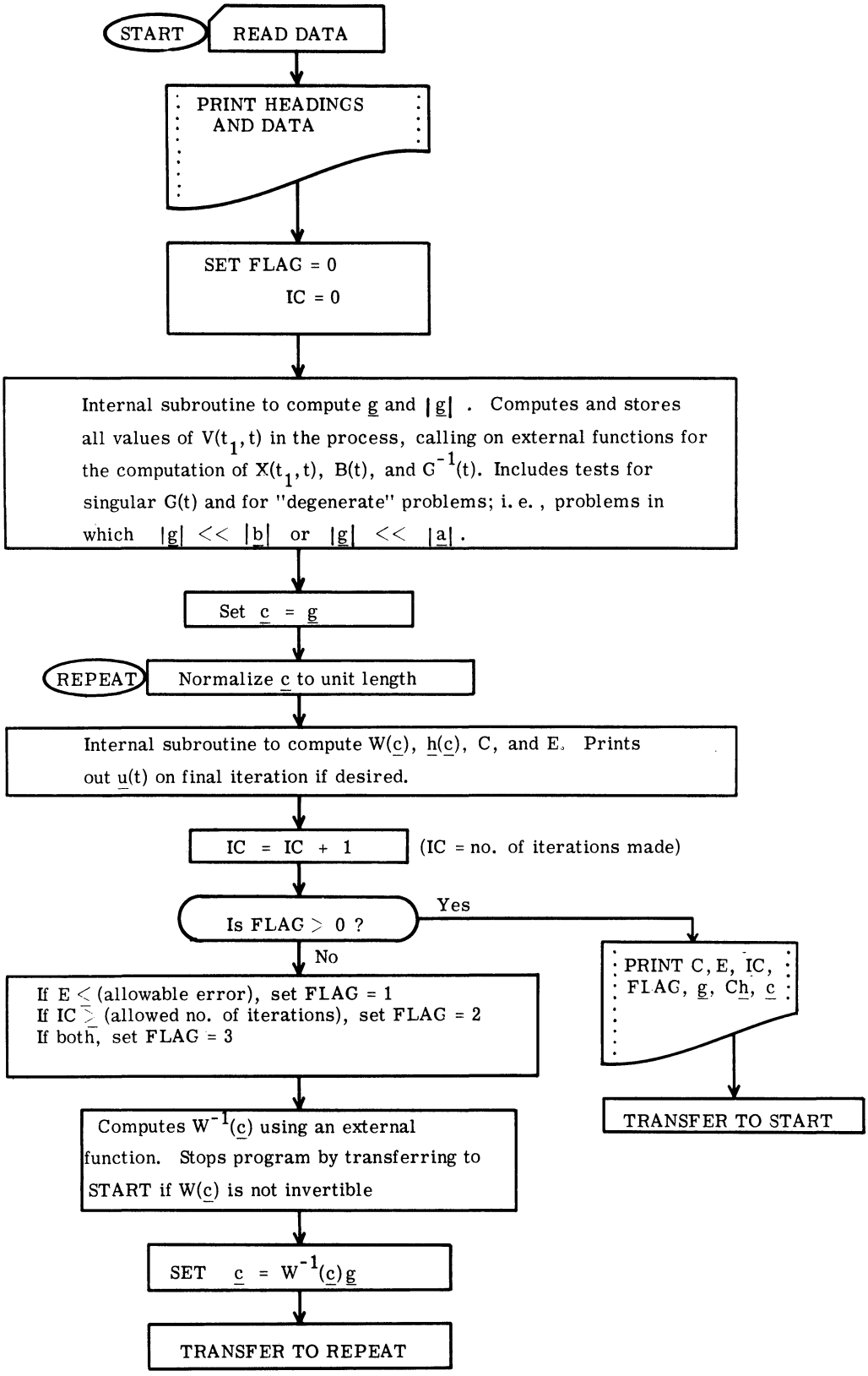

Fig. 7. 4 Simplified diagram for second algorithm.

The basic idea of this algorithm is very simple, and follows the pattern used in contraction mappings[55]: We choose a unit vector $\underline{c}$ (the initial choice, as before, is $\frac{\underline{g}}{|\underline{g}|}$), and then compute the next value of $\underline{c}$ as the unit vector

$$\underline{c}' = \frac{W^{-1}(\underline{c})\underline{g}}{|W^{-1}(\underline{c})\underline{g}|} \tag{7.7}$$

Using this value $\underline{c}'$, we compute $W(\underline{c}')$ and $\underline{h}(\underline{c}') = W(\underline{c}')\underline{c}'$. The error E is evaluated as before, and if it is sufficiently small, the results are printed out and iteration ceases. Otherwise, a new value of $\underline{c}$ is computed by the same rule, and so on.

Although the convergence of this algorithm has not been proved theoretically, the algorithm has in fact converged for the half-dozen or so cases to which it was applied. Indeed, in the second-and third-order problems for which comparisons were made, it has proven to be somewhat faster than the first algorithm. Any advantage it has over the other might be expected to diminish or disappear for higher order systems, however, because of the difficulties involved in the inversion of large matrices.

Fig. 7.5 illustrates the convergence of the algorithm for the simple system used in Example 4, Section 3.7, with boundary conditions $\underline{x}(0) = \underline{0}$, $\underline{x}(1) = [1, 1]^T$. For this system

$$W(\underline{c}) = \begin{bmatrix} \frac{1}{2c_1^2}\left(1 - \frac{c_2^2}{|c_1|}\log_e\frac{1+|c_1|}{|c_2|}\right) & 0 \\ 0 & \frac{1}{|c_1|}\log_e\frac{1+|c_1|}{|c_2|} \end{bmatrix}$$

for $c_1 \neq 0$ and

$$W(\underline{c}) = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix} \quad \text{for } c_1 = 0.$$

---

[55] It has not been proven to be a contraction mapping, however.

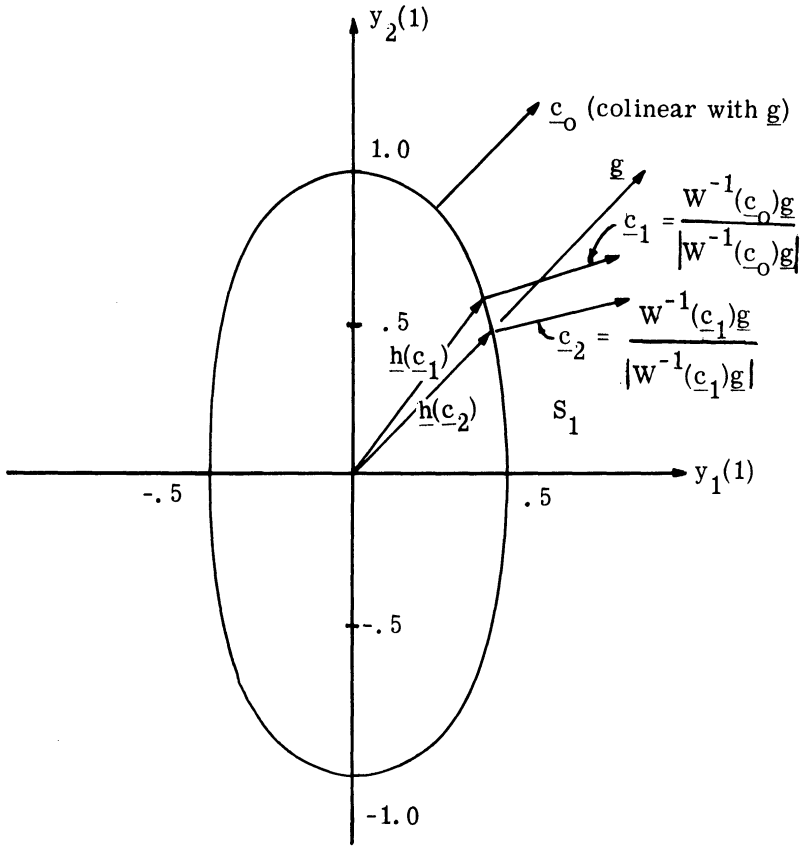

Fig. 7.5 Convergence of the second algorithm for a 2-dimensional problem.

This figure shows that after only one iteration the error has been greatly reduced and after two iterations the error is too small to be shown graphically to the scale used in Fig. 7.5.

7.4 Some Additional Examples

In this section we present some additional examples, solved with the aid of one of the other of the computational algorithms mentioned above. These examples have been chosen primarily to illustrate the various kinds of behavior that can result in minimum peak amplitude problems.

Example 14: An undamped oscillatory system.

System Characterization:

$$\dot{x}_1 = x_2 + u_1$$

$$\dot{x}_2 = -x_1 + u_2$$

Boundary Conditions:

$$\underline{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \underline{x}(t_1) = \begin{bmatrix} 0 \\ b_2 \end{bmatrix}$$

Curves of $\bar{C}$ vs t_1 for various values of b_2 are shown in Fig. 7.6. We note that each of these curves exhibits some ripple (except the one with $b_2 = 0$), and hence that (except for the case $b_2 = 0$) the inverse relationship is not complete for these cases (see Sections 4.3 and 4.5).

Example 15: A problem having an absolute minimum cost.

System Characterization:

$$\dot{x}_1 = -x_1 + x_2 + u_1$$

$$\dot{x}_2 = -2x_2 + u_2$$

Boundary Conditions:

$$\underline{x}(0) = \begin{bmatrix} 0 \\ 4 \end{bmatrix} \qquad x_1(t_1) = 1.1477$$

$$x_2(t_1) \text{ unspecified}$$

The curve of $\bar{C}$ vs. t_1 is shown for this example in Fig. 7.7 from which we see that $\bar{C}$ takes on its absolute minimum value at a value of t_1 near 1.0.

Example 16: A system with a stable focus.

System Characterization:

$$\dot{x}_1 = -x_1 + 2\pi x_2 + u_1$$

$$\dot{x}_2 = 2\pi x_1 - x_2 + u_2$$

Boundary Conditions:

$$\underline{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad x_1(t_1) = b_1$$

$$x_2(t_1) \text{ unspecified}$$

The resulting $\bar{C}$ vs. t_1 curve is shown in Fig. 7.8. For the curve with $b_1 = 1.0$, note the succession of minima, all having the same value.

Example 17: A time-varying system.

System Characterization:

$$\dot{\underline{x}} = \begin{bmatrix} 0 & 1 \\ -\frac{5}{t^2} & -\frac{7}{t} \end{bmatrix} \underline{x} + \underline{u}$$

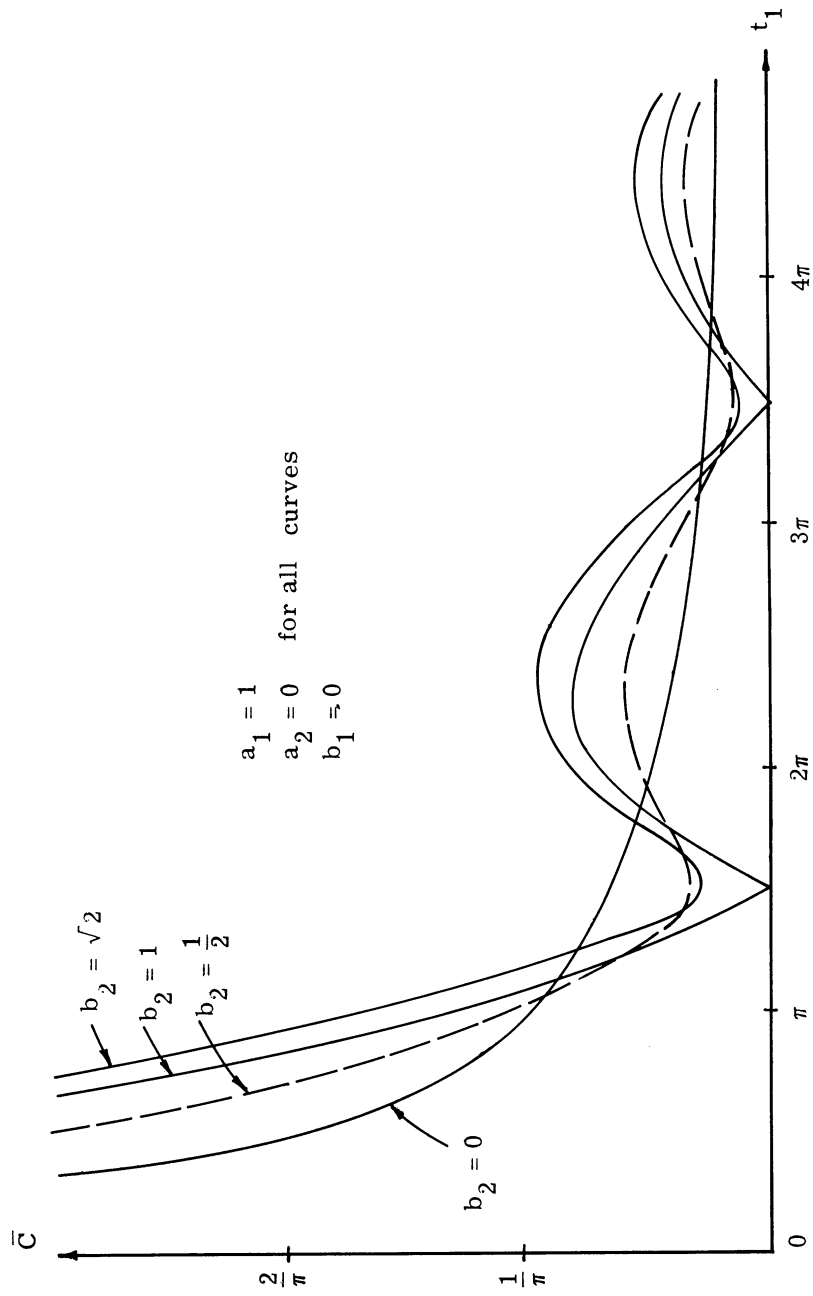

Fig. 7.6 $\bar{C}$ - vs - t_1 for an undamped oscillatory system, for various final conditions.

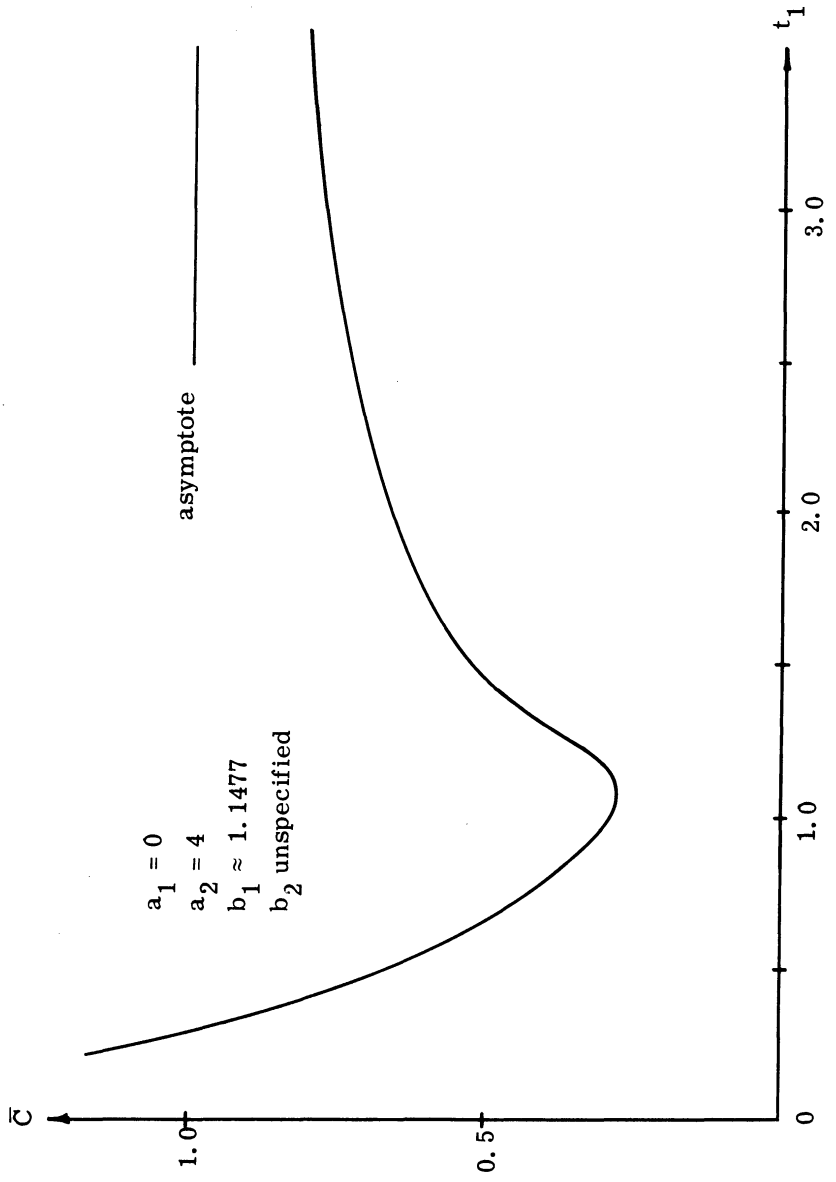

Fig. 7.7 $\bar{C}$ -vs- t_1 for a system with a stable node, for a particular set of boundary conditions.

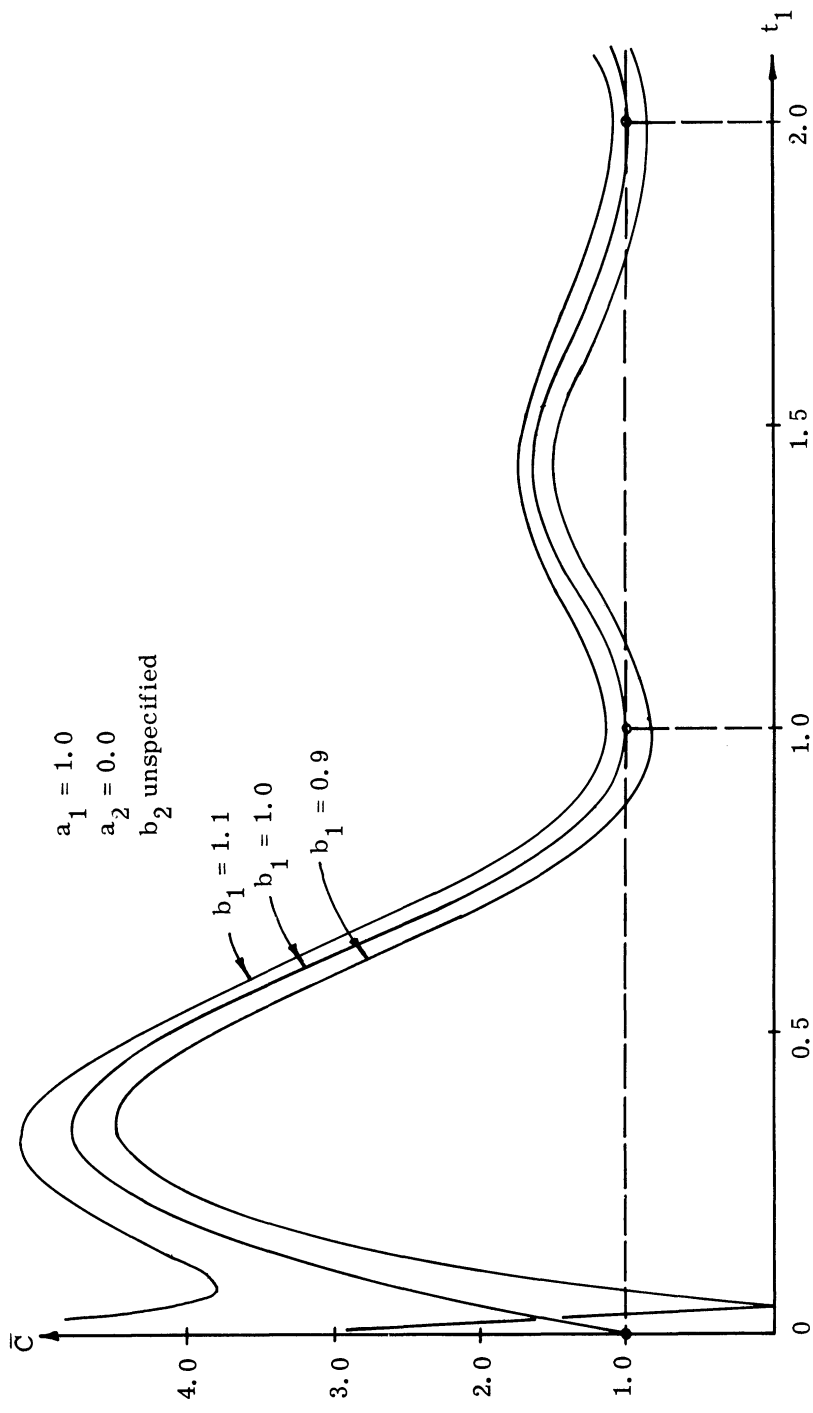

Fig. 7.8 $\bar{C}$ - vs - t_1 for a system with a stable focus, for various final conditions.

Boundary Conditions:

$$\underline{x}(t_0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \underline{x}(t_1) = \begin{bmatrix} .14 \\ -.45 \end{bmatrix}$$

$$t_0 = 1.0 \qquad t_1 > t_0$$

The first computational algorithm was used solve this problem for a series of values of t_1 from 1.2 to 2.0. The values of $\bar{C}$ obtained are plotted vs. t_1 in Fig. 7.9a). Fig. 7.9b) shows the shape of the control functions $u_1(t)$ and $u_2(t)$ for the value $t_1 = 1.8$. (These results are typical of all the various values of t_1 investigated.)

The number of iterations required for convergence to within an error E of .001 ranged from 10 for $t_1 = 1.2$ to 2 for $t_1 = 2.0$, with the average number of iterations for the twelve chosen values of t_1 being slightly over three and the most common number being two. Execution time on an IBM 7090 was 41.5 seconds for all twelve problems (i.e., all twelve values of t_1).

Example 18: A sixth-order time-varying system. This example involves an object moving in three-dimensional Euclidean space subject to velocity damping. The mass of the object decreases linearly with time. Initial position and velocity are specified, and the control is to be chosen so as to force the object to a given destination at a given time while requiring the minimum peak amplitude. As noted in Chapter I, this problem can be identified with that of determining the minimum-thrust steerable rocket engine that will accomplish the given task.

System Characterization:

$$\dot{\underline{x}} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{K_d}{M(t)} & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{K_d}{M(t)} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{K_d}{M(t)} \end{bmatrix} \underline{x} + \frac{1}{M(t)} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \underline{u}$$

$$\underline{y} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \underline{x}$$

where $M(t) = M_p + K_m(t_1 - t)$

Boundary Conditions:

$$\underline{x}(0) = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} \qquad \underline{y}(t_1) = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

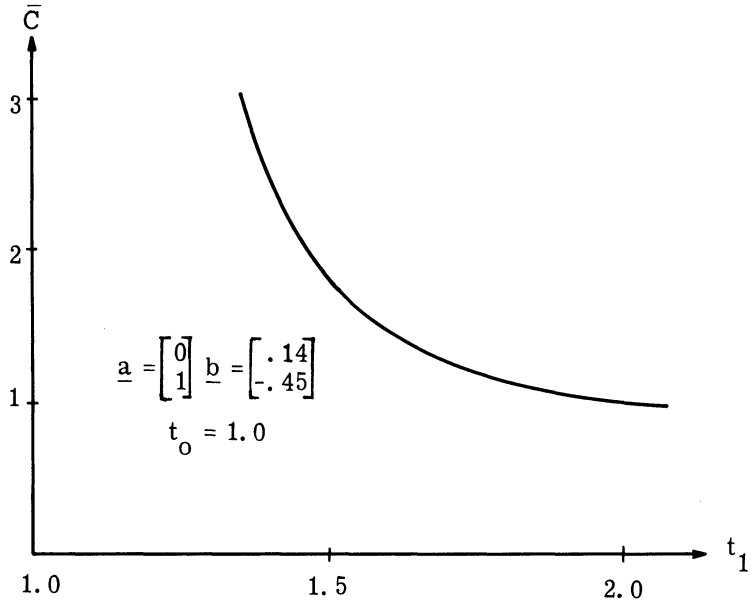

Fig. 7.9(a) $\bar{C}$ vs. t_1 for a time-varying system.

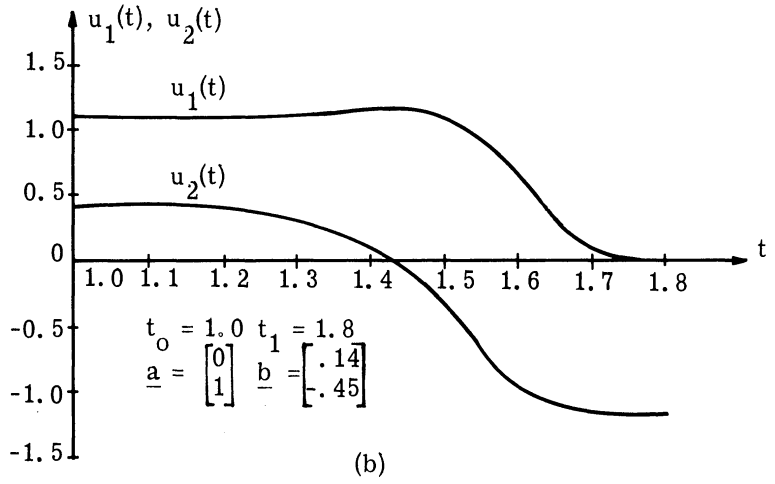

Fig. 7.9(b) The optimal control for a particular choice of t_1.

For this system, it can be shown that

$$\mathbf{X}(t,s) = \begin{bmatrix} 1 & 0 & 0 & f_1(t,s) & 0 & 0 \\ 0 & 1 & 0 & 0 & f_1(t,s) & 0 \\ 0 & 0 & 1 & 0 & 0 & f_1(t,s) \\ 0 & 0 & 0 & f_2(t,s) & 0 & 0 \\ 0 & 0 & 0 & 0 & f_2(t,s) & 0 \\ 0 & 0 & 0 & 0 & 0 & f_2(t,s) \end{bmatrix}$$

where

$$f_1(t,s) = \frac{M(s)}{K_d + K_m} \left[1 - \left(\frac{M(t)}{M(s)} \right)^{\frac{K_d}{K_m} + 1} \right]$$

and

$$f_2(t,s) = \left[\frac{M(t)}{M(s)} \right]^{\frac{K_d}{K_m}}$$

A series of problems of this type were run on the computer, using the second algorithm. Numerical values used were

$$M_p = 1000 \quad K_d = 5 \times 10^{-4} \quad t_1 = 5 \times 10^5$$

$$\underline{a} = [0,\ 0,\ 0,\ 10000,\ 500,\ 1000]^T$$

$$\underline{b} = [10^{10},\ 0,\ 0]^T$$

and K_m ranging from 0. 002 to 0. 02 in 9 steps. (Note that variations in K_m correspond to variations in initial mass, since the final mass is fixed at 1000.) The resulting values of $\bar{C}$ are plotted vs. K_m in Fig. 7. 10. The algorithm required two iterations in each case to converge to within an error E of 0. 001, and the total execution time for the 9 problems on an IBM 7090 computer was 44. 1 seconds.

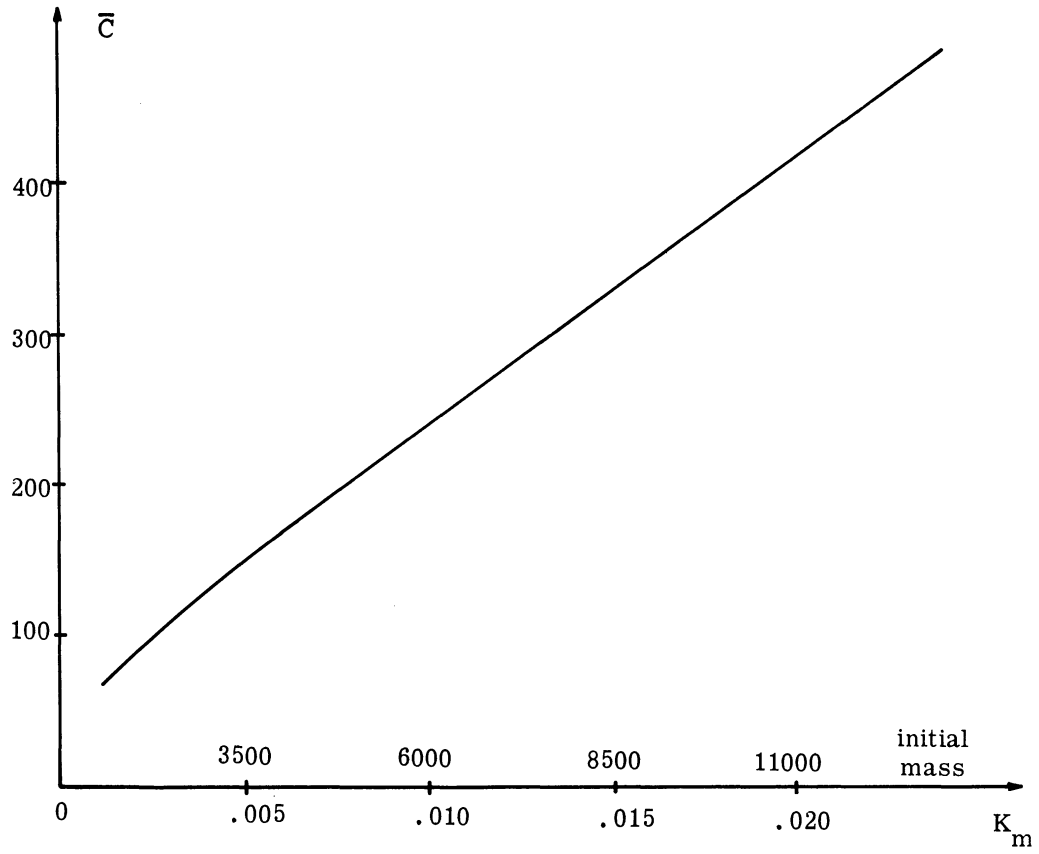

Fig. 7.10 The variation of the minimum peak amplitude as a function of the constant K_m (and the initial mass) for Example 18.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

8.1 Introduction

Certain optimal control problems, designated here as minimum peak amplitude control problems, are studied in detail in this work. Stated briefly, these problems involve the determination of a set of control variables (i. e. , the control vector) as a function of time so as to accomplish a specified goal with the minimum cost, where the cost is defined as the essential supremum over the given time interval of the Euclidean length of the control vector, considered as a vector function of time in Euclidean space. The significant feature of these problems is that this cost functional can not be expressed as an integral over time of some function of the control and state variables, so that classical variational techniques are not applicable. This feature is responsible for both the interest and the difficulties held by this class of problems.

8.2 Conclusions

After a brief statement of the problem and an identification of problems of this type with certain system design problems in electronics and rocketry, Chapter I provides a discussion of the history and present status of optimal control theory. Four major approaches to optimization problems are discussed in some detail:

1) Classical calculus of variations and its modern extensions, particularly Pontryagin's maximum principle: This approach offers rigorous proofs of optimality, but is not very well adapted to numerical computations (because of the two-point boundary value problems involved). These techniques apply equally well to linear and nonlinear problems, but in practice the nonlinear problems are, of course, much more difficult to solve, in general.

2) Dynamic programming: This approach does not have the mathematical rigor of the variational methods, but it does not involve two-point boundary value problems and it can be used to obtain answers (without assurance of optimality, unfortunately) for a larger class of problems than any of the other techniques discussed. This approach also applies to both linear and nonlinear problems.

3) Functional analysis: The considerable body of theory available in functional analysis has been applied to certain linear optimal control problems, often resulting in proofs, derivations, and solutions which are neat and concise, as compared to those involved in the other techniques discussed. Of particular interest are the existence theorems that this approach provides. Unfortunately, these results can not be extended to nonlinear problems, in general.

4) Direct methods: These methods involve the approximation of the optimal control by an iterative process operating directly on the cost in some way (rather than the solution of a set of auxiliary equations). Drawbacks include the need for proving that the iterative process converges to the optimum.

Recent efforts to render the various optimization techniques of more engineering significance are also discussed briefly; to name a few of these: improved computational techniques, optimal control laws, vector-valued cost functions, the inverse problem, and multilevel control.

Chapter I concludes with a discussion of the various approaches already discussed as they apply (or fail to apply) to minimum peak amplitude control problems. The applicability of the functional analysis approach is noted.

The problem to be considered is formulated in detail in Chapter II, and the various limitations and assumptions are stated and discussed. Various generalizations of the problem are also noted and discussed.

The main results of this work are contained in Chapter III. These are:

1) An existence theorem for optimal controls;

2) A theorem given conditions on the form of the optimal control;

3) A uniqueness theorem applicable to a restricted class of systems (here called proper systems);

4) Extensive discussions concerning the relationship between the shape of the reachable set and the form of the optimal controls;

5) Various generalizations of the problem,

6) A new approach to the selection of cost functionals for optimal control problems, called the "hierarchical" approach.

Theorems and ideas from modern analysis and topology are used in the proofs of most of the mathematical results of this chapter.

The relationship between the original minimum peak amplitude problem and a certain time-optimal problem, here called the inverse problem, is investigated in Chapter IV. A close connection between the two problems is established; namely that under certain conditions (discussed in considerable detail) the control which is optimal for one is optimal for the other. However, it is shown that not all minimum peak amplitude problems have inverses.

In Chapter V another kind of problem, here called the related linear problem, is introduced and discussed. The basic result here is that the optimal control problem involving a limiting form of a certain integral-type cost functional has the same optimal cost as the limiting cost of a sequence of problems using the stated integral-type cost functional; i. e., that the processes of solution of an optimization problem and passing to the limit on the cost functional can be interchanged in this case.

The approach of Chapter V is applied to certain nonlinear problems in Chapter VI. Steps analogous to those used in solving the related linear problem and passing to the limit can be used in nonlinear problems. Rigorous proofs of convergence are not attempted, but examples are presented to show the convergence in particular cases.

Chapter VII presents computational algorithms applicable to proper linear systems and to a special class of systems for which a certain matrix is invertible. The convergence of these algorithms is discussed both from a theoretical standpoint and by means of

examples. The role of the shape of the reachable set in the convergence of the algorithms is especially emphasized. Finally, additional examples are presented to illustrate the variety of behavior that can be obtained even in simple problems and to demonstrate the capabilities and speed of the computational algorithms.

8.3 Suggestions for Future Research

Three areas touched upon in this work appear to this writer to merit further investigation. The first of these involves the extension of these results to nonlinear systems in a rigorous manner. The most promising area for initial work involves nonlinear systems with linear control, since in such problems the limiting process indicated in Chapter VI can be used to obtain an explicit and simple expression for the "candidate for optimality." The convergence (or lack of convergence) of this limiting process for more general types of nonlinear systems might also be a fruitful area for future work.

A second area of possible research is that of nonlinear mappings between linear spaces. In the problem considered here the original mapping from control space to "final condition" space was linear, but the problem was later reduced to one involving a nonlinear mapping between two finite dimensional spaces. Such mappings might exhibit very interesting properties under further study.

Finally, rigorous proofs of the convergence of the computational algorithms would be desirable. In particular, it would be interesting to know under what conditions (if any) the mapping defined by Eq. 7.7 is indeed a contraction mapping (as it appears to be on the basis of the various numerical examples to which this algorithm was applied).

APPENDIX A

OPTIMAL CONTROL PROBLEMS AS PROBLEMS IN THE CALCULUS OF VARIATIONS

This appendix demonstrates the equivalence of a certain class of optimal control problems to the Lagrange problem of the calculus of variations, gives a derivation of a simplified version of the Lagrange multiplier rule, states the more general form of this rule applicable to problems with separated but otherwise general boundary conditions, and discusses extensions of these results to more general problems.

A. 1 Types of Solutions

Before proceeding with a discussion of variational methods, we shall define what is meant by the term "solution," as applied to optimal control problems:

a) A transformation of the original problem (which involves certain differential equations, boundary values, and constraint equations, plus a statement that a certain functional is to be minimized) into a well-specified boundary value problem solvable (in principle, at least) by conventional techniques is sometimes called a solution to the optimal control problem.

b) A complete analytical or numerical solution of the problem, resulting in a specification of the controlling variables as functions of time, is also called a solution.

c) A determination of the control law, --i. e., an expression which yields the optimal control as a function of the state of the system (rather than as a function of time)--is referred to by some (Pontryagin, Ref. 13) as a solution to the synthesis problem of optimal control.

Variational methods typically yield a solution in the sense of definition a). Solutions in sense b) and/or sense c), if desired, are then obtained by other techniques. In practice, of course, the two-point boundary value problems which result from a) may be extremely difficult to solve.

A. 2 Problem Formulations

A. 2. 1 The Problems of Bolza, Lagrange, and Mayer. An important problem in the calculus of variations is the problem of Bolza with separated end-conditions, which is the problem (as formulated by Bliss, Ref. 8, p. 193) of finding, in the class of arcs

$$\underline{y}(t) = [y_1(t), \ldots, y_m(t)]^T \qquad t \in T$$

lying in (m+1)-dimensional Euclidean space[56] and satisfying differential equations and end conditions of the form

$$g_j(\underline{y}, \dot{\underline{y}}, t) = 0 \qquad j = 1, \ldots, n; \qquad n < m$$

$$h_k(\underline{y}(t_0), t_0) = 0 \qquad k = 1, \ldots, p; \qquad p \leq m+1$$

$$h_k(\underline{y}(t_1), t_1) = 0 \qquad k = p+1, \ldots, p+q; \qquad q \leq m+1$$

one which minimizes a sum of the form

$$J = F_0(\underline{y}(t_0), t_0) - F_1(\underline{y}(t_1), t_1) + \int_T f_0(\underline{y}, \dot{\underline{y}}, t)dt$$

where T is the closed interval $[t_0, t_1]$ of the time axis. The point $(\underline{y}, \dot{\underline{y}}, t)$ is required by Bliss to lie in an open region R in (2m+1)-dimensional Euclidean space, the functions f_0 and g_j are assumed to have continuous partial derivatives of at least third-order in this region, and restrictions are placed on g_j and h_k in order to insure independence. Not all the variables need be present in every case.

This problem has equivalent formulations as a problem of Mayer (in which f_0 is missing) or as a problem of Lagrange (in which F_0 and F_1 are missing), (Bliss, Ref. 8, p. 189, and Meile, Ref. 62, p. 109). For purposes of this discussion, the Lagrange formulation is more convenient. Hence, without losing any generality, we will henceforth restrict attention to this formulation of the problem.

---

[56] The values of the variables $y_1(t), \ldots, y_m(t)$, and t are represented as distances along the coordinate axes of the Euclidean space. This representation of a vector function of time as a moving point in a Euclidean space is used throughout this section.

A. 2. 2 A Class of Optimal Control Problem as Problems of Lagrange.

Optimal control problems are usually stated in a form different from that just given. For example, a certain class of optimal control problems is typically formulated as follows: Assume a system satisfying the independent equations

$$\dot{x}_i = f_i(\underline{x},\underline{u},t) = f_i(x_1,\ldots,x_n,u_1,\ldots,u_r,t); \quad i = 1,\ldots,n; \quad t \in T$$

where $(\underline{x},\dot{\underline{x}},t)$ is constrained to lie in the open set R_x of (2n+1)-dimensional Euclidean space and where f_i (and also the f_0 defined below) are continuous and have continuous partial derivatives of at least third order in R_x. In the class of bounded measurable control functions $\underline{u}(t)$, $t \in T$ lying in the open set R_u of r-dimensional Euclidean space $E^r (r \geq 1)$ which cause $\underline{x}(t)$ to satisfy the independent conditions

$$h_k(\underline{x}(t_0),t_0) = 0 \qquad k = 1,\ldots,p; \qquad p \leq n+1$$

$$h_k(\underline{x}(t_1),t_1) = 0 \qquad k = p+1,\ldots,p+q; \qquad q \leq n+1$$

find one which minimizes

$$J = \int_T f_0(\underline{x},\underline{u},t)dt \ .$$

We will now recast this problem into a form identical to that given above for the problem of Lagrange: Define

$$y_i = x_i \qquad i = 1,\ldots,n$$

$$\dot{y}_{n+i} = u_i \qquad i = 1,\ldots,r$$

$$m = n + r$$

$$\underline{y}(t) = [y_1(t),\ldots,y_n(t), y_{n+1}(t),\ldots,y_m(t)]^T$$

$$g_j(\underline{y},\dot{\underline{y}},t) = \dot{y}_j - f_j(y_1,\ldots,y_n,\dot{y}_{n+1},\ldots,\dot{y}_m,t)$$

$$h_k(\underline{y}(t_0),t_0) = h_k(\underline{x}(t_0),t_0) \qquad k = 1,\ldots,p$$

$$h_k(\underline{y}(t_1),t_1) = h_k(\underline{x}(t_1),t_1) \qquad k = p+1,\ldots,p+q$$

$$f_0(\underline{y},\dot{\underline{y}},t) = f_0(y_1,\ldots,y_n,\dot{y}_{n+1},\ldots,\dot{y}_m,t)$$

$$R = R_x \times R_u \times E^r$$

Then this optimal control problem can be rewritten in a form identical to that of the above-stated calculus of variations problem, as follows:

Find, in the class of arcs $\underline{y}(t)$, $t \in T$, lying in (m+1)-dimensional Euclidean space and satisfying differential equations and end conditions of the form

$$g_j(\underline{y}, \dot{\underline{y}}, t) = 0 \qquad j = 1, \ldots, n; \qquad n < m$$

$$h_k(\underline{y}(t_0), t_0) = 0 \qquad k = 1, \ldots, p; \qquad p \leq n+1$$

$$h_k(\underline{y}(t_1), t_1) = 0 \qquad k = p+1, \ldots, p+q; \qquad q \leq n+1$$

one which minimizes

$$J = \int_T f_0(\underline{y}, \dot{\underline{y}}, t)dt \, .$$

The point $(\underline{y}, \dot{\underline{y}}, t)$ must lie in the open set R of (2m+1)-dimensional Euclidean space. (Note that the inclusion of the whole space E^r in the definition of R implies that the point is restricted in at most only 2n+r+1 of its coordinates, the remaining r being completely free.) The functions f_0 and g_j are continuous and have continuous partial derivatives of at least third-order in R. (Note also that there is no longer any explicit distinction between state and control variables.)

With this equivalence of the two problems in mind, we can proceed to apply the results of the classical calculus of variations to optimal control problems. Furthermore, results from the calculus of variations derived for conditions different from those imposed by Bliss (Ref. 8) in his treatment of the Bolza problem (e.g., the requirement that the set R be open) will apply to optimal control problems with correspondingly different conditions.

A.3 The Lagrange Multiplier Rule

The problems posed in the previous section can be solved with the help of the Lagrange multiplier rule. A simplified version of this rule will now be derived.

A.3.1 Derivation of the Lagrange Multiplier Rule.

Problem: Given a system described by $\dot{\underline{x}} = \underline{f}(\underline{x}, \underline{u}, t)$, where the equations $\dot{x}_i = f_i(\underline{x}, \underline{u}, t)$, $i = 1, \ldots, n$ are consistent and independent. From the set of bounded measurable control functions $\underline{u}(t)$ lying at each moment of time $t \in T$ in the open set R_u of r-dimensional

Euclidean space which cause the system to transfer from the given initial state $\underline{x}(t_o) = \underline{a}$ at initial time t_o to the desired goal state $\underline{x}(t_1) = \underline{b}$ at final time t_1 find the one which minimizes

$$J = \int_T f_o(\underline{x}, \underline{u}, t)\, dt \ .$$

The functions f_o and f_i are continuous in all their arguments and have continuous partial derivatives of at least third order.

Derivation (after Weinstock, Ref. 63, pp. 57-60):

Assume that the above problem has a solution. Denote this solution by $\underline{x}(t)$, $\underline{u}(t)$. Imbed this problem in the more general problem of minimizing

$$J(\epsilon) = \int_T f_o(\underline{X}, \underline{U}, t)\, dt$$

subject to the constraints

$$\begin{aligned} g_i(\underline{X}, \dot{\underline{X}}, \underline{U}, t) &= 0 \\ X_i(t_o) &= a_i \qquad i = 1, \ldots, n \\ X_i(t_1) &= b_i \end{aligned}$$

where

$$\begin{aligned} X_i &= X_i(t) = x_i(t) + \epsilon\, y_i(t) \qquad i = 1, \ldots, n \\ U_j &= U_j(t) = u_j(t) + \epsilon\, v_j(t) \qquad j = 1, \ldots, r \\ g_i(\underline{X}, \dot{\underline{X}}, \underline{U}, t) &= \dot{X}_i - f_i(\underline{X}, \underline{U}, t) \end{aligned}$$

and where the bounded measurable functions $y_1(t), \ldots, y_n(t), v_1(t), \ldots, v_r(t)$ are arbitrary except to the extent implied by the above constraints. Note that, because $\underline{x}(t)$ and $\underline{X}(t)$ satisfy the same boundary conditions,

$$y_i(t_o) = y_i(t_1) = 0\,, \qquad i = 1, \ldots, n\,.$$

$J(\epsilon)$ will be a minimum only if $\frac{dJ(\epsilon)}{d\epsilon} = 0$. But, by definition of $\underline{x}(t)$ and $\underline{u}(t)$, this occurs when $\epsilon = 0$. Thus, a necessary condition that $J(\epsilon)$ be minimum is that

$$\left.\frac{dJ(\epsilon)}{d\epsilon}\right|_{\epsilon=0} = J'(0) = 0 .$$

Performing the indicated differentiation gives

$$J'(\epsilon) = \int_T \left[\sum_{i=1}^{n} \frac{\partial f_o}{\partial \mathbf{X}_i} y_i + \sum_{j=1}^{r} \frac{\partial f_o}{\partial U_j} v_j \right] dt .$$

At $\epsilon = 0$, this becomes

$$J'(0) = \int_T \left[\sum_{i=1}^{n} \frac{\partial f_o}{\partial x_i} y_i + \sum_{j=1}^{r} \frac{\partial f_o}{\partial u_j} v_j \right] dt \qquad \text{(A. 1)}$$

Now, differentiating $g_k(\underline{\mathbf{X}}, \dot{\underline{\mathbf{X}}}, \underline{U}, t) = 0$ with respect to ϵ, we get

$$\sum_{i=1}^{n} \left[\frac{\partial g_k}{\partial \mathbf{X}_i} y_i + \frac{\partial g_k}{\partial \dot{\mathbf{X}}_i} \dot{y}_i \right] + \sum_{j=1}^{r} \left[\frac{\partial g_k}{\partial U_j} v_j \right] = 0, \quad k = 1, \ldots, n .$$

Making use of the facts that $\frac{\partial g_k}{\partial \dot{\mathbf{X}}_i} = \begin{bmatrix} 0 & i \neq k \\ 1 & i = k \end{bmatrix}$, we have that

$$\frac{\partial g_k}{\partial \mathbf{X}_i} = - \frac{\partial f_k}{\partial \mathbf{X}_i} \quad \text{and} \quad \frac{\partial g_k}{\partial U_j} = - \frac{\partial f_k}{\partial U_j} .$$

Setting $\epsilon = 0$, we obtain finally

$$- \sum_{i=1}^{n} \frac{\partial f_k}{\partial x_i} y_i + \dot{y}_k - \sum_{j=1}^{r} \frac{\partial f_k}{\partial u_j} v_j = 0 , \qquad k = 1, \ldots, n . \qquad \text{(A. 2)}$$

Since each of the expressions on the left-hand sides of Eqs. A. 2 is equal to zero along any system trajectory (and hence also along the optimal trajectory), adding any multiples of them to the integrand of Eq. A. 1 does not change the value of the integral. Thus, we may multiply the left-hand sides of Eqs. A. 2 by the as-yet-undetermined Lagrange multipliers

$\psi_1(t), \ldots, \psi_n(t)$, respectively, and add the results to the integrand of Eq. A. 1, yielding

$$\int_T \left[\sum_{i=1}^{n} \left[\left[\frac{\partial f_o}{\partial x_i} - \sum_{k=1}^{n} \psi_k \frac{\partial f_k}{\partial x_i} \right] y_i + \psi_i \dot{y}_i \right] + \sum_{j=1}^{r} \left[\frac{\partial f_o}{\partial u_j} - \sum_{k=1}^{n} \psi_k \frac{\partial f_k}{\partial u_j} \right] v_j \right] dt = 0 \ .$$

Integration by parts of the term $\int_T \sum_{i=1}^{n} \psi_i \dot{y}_i \, dt$ and application of the boundary conditions $y_i(t_0) = y_i(t_1) = 0$ gives

$$\int_T \sum_{i=1}^{n} \psi_i \dot{y}_i \, dt = - \int_T \sum_{i=1}^{n} y_i \dot{\psi}_i \, dt \ .$$

Hence,

$$\int_T \left[\sum_{i=1}^{n} \left[\frac{\partial f_o}{\partial x_i} - \sum_{k=1}^{n} \psi_k \frac{\partial f_k}{\partial x_i} - \dot{\psi}_i \right] y_i + \sum_{j=1}^{r} \left[\frac{\partial f_o}{\partial u_j} - \sum_{k=1}^{n} \psi_k \frac{\partial f_k}{\partial u_j} \right] v_i \right] dt = 0 \ . \qquad \text{(A. 3)}$$

Because of the set of n equations $g_k(\underline{X}, \dot{\underline{X}}, \underline{U}, t) = 0$, $k = 1, \ldots, n$ relating the n + r functions $y_1(t), \ldots, y_n(t), v_1(t), \ldots, v_r(t)$, we cannot regard these n + r functions as being free for arbitrary choice. In fact, there is some subset of n of these functions which are dependent on the choice of the remaining r. We now choose the n arbitrary functions $\psi_1(t)$, $\ldots, \psi_n(t)$ so as to make the coefficients of these n dependent functions in Eq. A. 3 equal to zero. Equation A. 3 then reduces to the integral of a linear combination of the remaining r arbitrary functions. But this integral must be zero for any choice of the arbitrary functions, and therefore the coefficients of these functions must also be zero. Thus, all the coefficients of the variables $y_1, \ldots, y_n, v_1, \ldots, v_r$ in Eq. A. 3 must be zero.

In summary, then, if the problem has solutions, these solutions must satisfy the following equations and conditions:

$$\dot{\underline{x}} = \underline{f}(\underline{x}, \underline{u}, t); \quad \underline{x}(t_0) = \underline{a}; \quad \underline{x}(t_1) = \underline{b}$$

$$\dot{\psi}_i = - \frac{\partial}{\partial x_i} \left[-f_o(\underline{x}, \underline{u}, t) + \sum_{k=1}^{n} \psi_k f_k(\underline{x}, \underline{u}, t) \right]; \qquad i = 1, \ldots, n$$

$$\frac{\partial}{\partial u_j}\left[-f_o(\underline{x},\underline{u},t) + \sum_{k=1}^{n} \psi_k f_k(\underline{x},\underline{u},t)\right] = 0 ; \qquad j = 1, \ldots, r$$

A. 3. 2 Application of the Lagrange Multiplier Rule. The use of the Lagrange Multiplier rule will be illustrated by its application to an optimal control problem less general than that posed in Section A. 2. 2; namely, one in which the set R_x is the whole $(2n+1)$-dimensional Euclidean space and in which the end conditions are completely specified, as follows:

$$x_i(t_o) = a_i \qquad x_i(t_1) = b_i \qquad i = 1, \ldots, n$$

where a_i, b_i, t_o, and t_1 are all given constants.

The Lagrange multiplier rule then gives as a necessary (but not sufficient) condition that a $\underline{u}(t)$ and the corresponding $\underline{x}(t)$ satisfying $\dot{\underline{x}} = \underline{f}(\underline{x},\underline{u})$; $\underline{x}(t_o) = \underline{a}$; $\underline{x}(t_1) = \underline{b}$ be optimal is that there exists a continuous nonzero vector function $\underline{\psi}(t) = (\psi_1(t), \ldots, \psi_n(t))^T$ which satisfies the Euler-Lagrange equations:

$$\dot{\psi}_i = -\frac{\partial H(\underline{x},\underline{u},\underline{\psi},t)}{\partial x_i} \qquad i = 1, \ldots, n \qquad \text{(A. 4)}$$

$$\frac{\partial H(\underline{x},\underline{u},\underline{\psi},t)}{\partial u_j} = 0 \qquad j = 1, \ldots, r \qquad \text{(A. 5)}$$

where, for convenience of notation (and in order to bring out an analogy with classical mechanics), $H(\underline{x},\underline{u},\underline{\psi},t)$ has been used to denote the expression

$$-f_o(\underline{x},\underline{u},t) + \sum_{i=1}^{n} \psi_i(t)\, f_i(\underline{x},\underline{u},t) \ .$$

Here n additional variables $\psi_1(t),\ldots,\psi_n(t)$, n additional differential equations (Eq. A. 4), and r additional finite equations (Eq. A. 5) have been introduced. The result is a problem with 2n differential equations, 2n boundary conditions, r finite equations, and 2n+r dependent variables--a "well-specified" problem to which conventional methods for the solution of systems of differential equations can be applied.

Two assumptions are crucial to these results:

a) In deriving these equations, we assume the existence of an optimal solution to the problem. Thus, the results apply to the solutions, if they exist, but they do not guarantee the existence of a solution or solutions, thus justifying the above statement that the satisfaction of the Euler-Lagrange equations is a necessary but not a sufficient condition that a given set of functions $\underline{x}(t)$, $\underline{u}(t)$, $\underline{\psi}(t)$ be optimal.

b) The control $\underline{u}(t)$ is required to lie in an open set R_u in r-dimensional Euclidean space E^r. J always has an infimum over $\underline{u}$'s in this set, but not necessarily a minimum (Pontryagin, Ref. 13, p. 239; Mortensen, Ref. 22, p. 5). This is, in fact, the major weakness of the Lagrange multiplier rule and of classical calculus of variations--they apply only when $\underline{u}$ lies in an open set, but the resulting transformed problem may have no solution.

In connection with the necessity of these conditions, note that in their derivation the solution which minimizes J is obtained by setting $\frac{dJ(\epsilon)}{d\epsilon} = 0$, where $J(\epsilon)$ is a one-parameter family of functionals and ϵ is the parameter. But obviously a maximum or a zero-slope inflection point of $J(\epsilon)$ would also satisfy the same condition. Thus, the Euler-Lagrange equations alone do not allow one to distinguish between maxima, minima, and zero-slope inflection points. To overcome this difficulty, we add as an additional necessary condition for a minimum the Legendre-Clebsch condition (Mortensen, Ref. 22, p. 7; Meile, Ref. 62, p. 104):

The matrix with elements $\frac{\partial^2 H(\underline{x}, \underline{u}, \underline{\psi}, t)}{\partial u_i \partial u_j}$ must be negative semidefinite.

A point of interest to the discussion of Pontryagin's maximum principle now becomes apparent:

Equation A. 5 indicates that the optimal $\underline{u}$ corresponds to a stationary point of the function $H(\underline{x}, \underline{u}, \underline{\psi}, t)$. The Legendre-Clebsch conditions indicates that this stationary point is a maximum. Both conditions can thus be summarized as follows: $\underline{u}(t)$ with values in R_u at each moment of time $t \in T$ must be chosen so as to maximize $H(\underline{x}, \underline{u}, \underline{\psi}, t)$ at each moment in time.

These results can be generalized to cover problems having the separated but otherwise general boundary conditions of the originally posed Bolza problem by noting that,

in order to be optimal, $\underline{x}(t)$ and $\underline{\psi}(t)$ must satisfy the transversality conditions of the calculus of variations. Using the approach of Bryson, (Refs. 41,42), we say that the transversality conditions are satisfied if $\underline{\psi}$ and $\underline{x}$ are such that the additional conditions

$$\psi_i(t_0) = \frac{\partial Q}{\partial x_i(t_0)}$$

$$\psi_i(t_1) = \frac{\partial Q}{\partial x_i(t_1)} \qquad i = 1,\ldots,n$$

are satisfied, where Q is defined by the expression

$$Q = \sum_{i=1}^{p} \theta_i h_i(\underline{x}(t_0), t_0) + \sum_{i=1}^{q} \theta_{p+i}\, h_{p+i}(\underline{x}(t_1), t_1)$$

and where $\theta_1,\ldots,\theta_{p+q}$ are multiplicative constants to be determined from the boundary conditions. Furthermore, if t_0 and/or t_1 are unspecified, the additional conditions needed for the determination of t_0 and/or t_1 are given by

$$H(t_0) + \frac{\partial Q}{\partial t_0} = 0$$

$$H(t_1) + \frac{\partial Q}{\partial t_1} = 0 \ .$$

A. 4 Extensions of the Classical Calculus of Variations

A. 4. 1 Pontryagin's Maximum Principle. As noted in the previous section, the classical calculus of variations, in its usual formulation, is limited to problems in which the point $(\underline{x},\dot{\underline{x}},\underline{u},t)$ is constrained to lie in an open set in $(2n + r + 1)$-dimensional Euclidean space. Problems of engineering interest often require some subset of these coordinates to lie in closed sets. Therefore, generalization of the classical results to problems with closed sets would be of great practical value.

Note, however, that various degrees of generalization are possible. As a first step, one might consider the case in which the values of $\underline{u}$ are constrained to lie in the closed set R_u in r-space but t, $\underline{x}$, and $\dot{\underline{x}}$ are unconstrained (i. e. , the set R is the cartesian product

of the whole $(2n + r + 1)$-space and the closed set R_u). Pontryagin and his associates (Refs. 10, 13) devised a set of necessary conditions for the solution of this problem. Similar results have also been presented by Berkovitz (Ref. 19) and Larsen (Ref. 21) using the method of Valentine (Ref. 9), as well as by Halkin (Refs. 64, 65), Kalman (Ref. 66), Leitman (Ref. 67), Rozoener (Ref. 16), and others.

The best known of these results, those of Pontryagin and his associates, are discussed below:

In Section A. 3. 1, a set of necessary conditions for the optimality of $\underline{x}$ and $\underline{u}$ was given for the case where R_u is an open set and R_x is the whole Euclidean space. These conditions can be summarized as follows:

A necessary condition that $\underline{u}(t)$ and the corresponding $\underline{x}(t)$, satisfying $\dot{\underline{x}} = \underline{f}(\underline{x}, \underline{u}, t)$ $\underline{x}(t_0) = \underline{a}$, $\underline{x}(t_1) = \underline{b}$, be optimal (i. e., minimize $J = \int_T f_0(\underline{x}, \underline{u}, t)dt$) is that there exists a continuous nonzero vector function $\underline{\psi}(t)$ such that

$$\dot{\psi}_i = -\frac{\partial H}{\partial x_i} \qquad i = 1, \ldots, n$$

and such that $\underline{u}(t)$ with values in R_u at each moment of time $t \in T$ maximizes

$$H(\underline{x}, \underline{u}, \underline{\psi}, t) = -f_0(\underline{x}, \underline{u}, t) + \sum_{i=1}^{n} \psi_i f_i(\underline{x}, \underline{u}, t)$$

at each moment of time. This can also be taken as a statement of the maximum principle[57] by letting R_u be a closed set. Pontryagin et al., proved (Ref. 13) that the maximum principle applies to problems with much less restrictive conditions on $\underline{f}(\underline{x}, \underline{u}, t)$ and $\underline{u}$ than were

---

[57] The maximum principle, as stated by Pontryagin, is more general than this, in that it is necessary only that $\underline{u}$ maximize H almost everywhere on the time interval, and in that a constant multiplier, ψ_0, is also used for $f_0(\underline{x}, \underline{u}, t)$. ψ_0 is required to be nonpositive, and the augmented vector $(\psi_0, \psi_1, \ldots, \psi_m)$ is required to be continuous and nonzero. Since $\underline{\psi}$ is determined only to within a multiplicative constant, however, it follows that the two statements are identical unless ψ_0 turns out to be zero. In the calculus of variations, solutions in which $\psi_0 = 0$ are called "abnormal" solutions (Bliss, Ref. 8, p. 213). Abnormality is often an indication of the fact that there is only one arc in the given neighborhood which satisfies the boundary conditions and constraints, and hence that the optimal solution does not depend on the "cost" function $f_0(\underline{x}, \underline{u}, t)$.

assumed by Bliss (Ref. 8). Specifically, in Pontryagin's treatment $\underline{f}(\underline{x},\underline{u},t)$ must be continuous in $\underline{x}$, $\underline{u}$, and t, but the only partial derivatives that are required to exist are the first partial derivatives of $\underline{f}$ with respect to t and the x_i's (but not the u_j's). Bliss required the existence of partial derivatives of at least third order with respect to all the arguments of $\underline{f}$. This weakening of restrictions is the principal theoretical contribution of Pontryagin (according to Mortensen, Ref. 22, p. 4).

A. 4. 2 Problems with Bounded State Variables. In the problems considered thus far, the state variables $\underline{x}(t)$ have been restricted only to the extent that they must satisfy certain differential equations. If we further restrict $\underline{x}$ by means of inequality constraints of the form

$$Z_i(x_1,\ldots,x_m) \geq 0 \qquad i = 1,\ldots,s$$

the results described thus far no longer apply. However, it has been shown, (Berkovitz, Refs. 19,20) that such problems can be put into the form of a classical Lagrange problem using the method of Valentine (Refs. 9,19,20,21):

Upon definition of additional functions $x_{m+1}(t),\ldots,x_{m+s}(t)$ satisfying

$$(\dot{x}_{m+i})^2 - Z_i(x_1,\ldots,x_m) = 0; \quad x_{m+i}(t_0) = 0; \qquad i = 1,\ldots,s\ ,$$

the problem can now be regarded as an ordinary Lagrange problem in the variables $x_1(t),\ldots,$ $x_{m+s}(t)$, to which all the classical results apply. Any solution to this augmented problem will obviously satisfy the inequality constraints, as well as the usual necessary conditions for an optimal solution. (Of course, additional variables, additional differential equations, and additional Lagrange multipliers have been introduced, so that the augmented problem may be (and usually is) considerably more difficult to solve, in practice, than the corresponding problem without inequality constraints.) Gamkrelidze (Refs. 13,14) proves similar results without using Valentine's method.

A further generalization is the case in which $\underline{x}$ and $\underline{u}$ are jointly constrained to a closed set in (n+r)-dimensional Euclidean space--a set which can not be decomposed into the cartesian product of a constraint set on $\underline{x}$ (independent of $\underline{u}$) in n-space and a constraint set on $\underline{u}$ (independent of $\underline{x}$) in r-space. Berkovitz's treatment (Ref. 19) covers this

situation also. The same approach can be used if $\underline{x}$, $\underline{u}$, and $\dot{\underline{x}}$ are jointly constrained to a closed set in (2n+r)-dimensional Euclidean space.

APPENDIX B

DYNAMIC PROGRAMMING AS A COMPUTATIONAL TECHNIQUE

B. 1 Philosophy and Motivation

"Dynamic programming" is the name given to a computational technique devised by Bellman (Refs. 17, 18) for use in the solution of a class of optimization problems.

The dynamic programming approach, which differs markedly from the usual variational approach, is motivated by the following considerations:

(a) Variational methods invariably yield two-point boundary value problems, which can not be solved analytically except in the simplest cases. Furthermore, even the form of the solution is often not determinable using these techniques. Thus, even though variational methods are analytical, in principle, results can only be obtained by numerical methods in many cases. Even here, variational techniques require the use of iterative methods since two-point boundary value problems generally do not lend themselves to straightforward solution, even on a computing machine. Dynamic programming, on the other hand, is from the outset a numerical method.

(b) Dynamic programming, because it bypasses the difficulties inherent in two-point boundary value problems, often results in reduced computation time.

(c) Recent extensions of the classical calculus of variations involve less restrictive conditions on the system equations than those required in earlier treatments. However, it is easy to find systems (e. g. , systems with multiple modes of operation selected by relays), which do not fit even these relaxed conditions. Dynamic programming, because it makes no use of differential properties of system and cost functions, can handle more general problems than can classical methods.

(d) The insight into the behavior of the system in question (as a function of boundary conditions, system parameters, etc.) that is provided by the analytic solution (when available) generally can not be obtained from one or a few numerical solutions for specific sets of boundary conditions. Thus, when numerical solutions must be resorted

to because of system complexity or nonlinearity, it is really a family of solutions that is needed for intelligent evaluation[58] of the results. Dynamic programming provides such families.

(e) Optimization problems must generally be converted to discrete-time form before they can be solved on digital computers. Thus, even though dynamic programming (as a computational technique) is used primarily in discrete-time problems, it does not necessarily suffer any basic loss of accuracy in comparison to continuous time techniques converted to a form suitable for digital computation.

(f) Variational techniques offer means of finding local maxima (and minima). In a complicated problem, considerable computation might be required in order to obtain and compare all such maxima (or minima) and thus select an absolute maximum (or minimum). Dynamic programming, however, always finds the absolute maximum (or minimum) in the specified region.

(g) Dynamic programming can be extended to apply to stochastic problems, whereas classical techniques are not readily extended in this direction.

The above points, while smacking in some cases of post facto justification of certain unavoidable aspects of the dynamic programming approach (in particular, the discrete-time format and the imbedding, which results in families of solutions), nonetheless summarize a valid point of view which has extensive application. Certain weaknesses inherent in the dynamic programming approach have already been noted (see Section 1. 2. 3).

B. 2 Formulation of the Problem

Assume that the system to be controlled is characterized by a set of difference equations

$$\underline{x}(t_{i+1}) = \underline{x}(t_i) + \underline{F}(\underline{x}(t_i), \underline{u}(t_i)) \quad i = 0, 1, \ldots, N$$

---

58 For example, in practice the boundary conditions are never known with absolute precision. Thus, from an engineering standpoint, it is important to determine the sensitivity of the optimal solution to variations in the boundary conditions. This sensitivity can be determined if families of solutions with the boundary conditions as parameters are provided.

where $\underline{x}(t_{i+1})$, the state of the system at time t_{i+1}, is given as a function of the previous state $\underline{x}(t_i)$ and the control vector $\underline{u}(t_i)$ applied at the previous time t_i. The time points $t_0, t_1, \ldots, t_N$ need not be equally spaced, although in practice they usually are. (If the original system description is in terms of differential equations, then a conversion to difference equation form must be made.)

For simplicity, we denote the vectors $\underline{x}(t_i)$ and $\underline{u}(t_i)$ by $\underline{x}_i$ and $\underline{u}_i$, respectively. The problem to be solved is then written as follows: Given an initial state $\underline{x}_0$, choose an input sequence $\underline{u}_0, \underline{u}_1, \ldots, \underline{u}_N$ (where each vector $\underline{u}_i$ must lie in the given set U) so that the criterion or cost function

$$\sum_{i=0}^{N} c_i F_0(\underline{u}_i, \underline{x}_i)$$

is minimum.[59]

B. 3 Imbedding

Let

$$G_N = \min_{\substack{\underline{u}_i \\ i = 0, \ldots, N}} \sum_{i=0}^{N} c_i F_0(\underline{u}_i, \underline{x}_i),$$

where here and in all that follows each $\underline{u}_i$ is understood to be chosen from the set U. (Such $\underline{u}$'s are said to be "admissible".) Since $\underline{u}_N$ affects only the last term in this series, (because, by assumption, the system is physically realizable, and hence can not respond to an input before the input is applied) it is obvious that G_N can be rewritten[60] as

---

59 No final conditions or bounds on $\underline{x}$ have been specified here, in order to avoid undue complexity in this exposition of the method. Problems with such constraints can be solved by dynamic programming with no increase in difficulty, however. Variations of the set U with time also pose no difficulties.

60 This result is a consequence of Bellman's "principle of optimality" (Refs. 17, 18), which states that an optimal trajectory has the property that, no matter what initial decisions (choices of control vectors) are made, the remaining part of the trajectory must be optimal.

$$G_N = \min_{\substack{\underline{u}_i \\ i = 0,\ldots,N-1}} \left[\sum_{i=0}^{N-1} c_i F_o(\underline{u}_i, \underline{x}_i) + \min_{\underline{u}_N} [c_N F_o(\underline{u}_N, \underline{x}_N)] \right]$$

Continuing this process, we get

$$G_N = \min_{\underline{u}_o} \left[c_o F_o(\underline{u}_o, \underline{x}_o) + \min_{\underline{u}_1} \left[c_1 F_o(\underline{u}_1, \underline{x}_1) + \cdots + \min_{\underline{u}_{N-1}} \left[c_{N-1} F_o(\underline{u}_{N-1}, \underline{x}_{N-1}) + \min_{\underline{u}_N} [c_N F_o(\underline{u}_N, \underline{x}_N)] \right] \cdots \right] \right]$$

a series of "nested" minimizations, each with respect to a single control vector $\underline{u}_i$.

We reason as follows: If we knew the state $\underline{x}_N$, we could compute the cost for the last stage of the process, simply by trying all values of $\underline{u}_N$ and picking the one which gave the lowest cost. We also thereby determine $\underline{u}_N$. But $\underline{x}_N$ is not known at this point, since it is a function of the as-yet-undetermined control vectors $\underline{u}_o, \ldots, \underline{u}_{N-1}$. Therefore, since we can't anticipate which value of $\underline{x}_N$ actually lies on the optimal trajectory, we will tabulate the optimal control and the associated cost for all $\underline{x}_N$ in our range of interest.[61] In practice, a grid of points in the region of interest in state space is established, and the optimal control and cost are computed and tabulated for each of these.

Now we consider the cost for the last two stages of the process. Again, we don't know $\underline{x}_{N-1}$, so we consider all $\underline{x}_{N-1}$'s of interest. For each possible $\underline{x}_{N-1}$, each possible choice of $\underline{u}_{N-1}$ determines a cost for that stage, and it also determines $\underline{x}_N$. Since we have already tabulated the optimal cost for each $\underline{x}_N$ of interest,[62] we can compute the total cost for the last two stages for that choice of $\underline{u}_{N-1}$. The choice

---

[61] It can be seen that bounds on the state variables $\underline{x}$ only serve to make the problem easier, since they limit the range of values of $\underline{x}$ that must be investigated. Any $\underline{u}$ which causes $\underline{x}$ to exceed the stated bounds is, of course, rejected, and the optimum choice is made from the remaining admissible $\underline{u}$'s (if any).

[62] If the $\underline{x}_N$ thus determined falls between the chosen grid points, we interpolate to get the needed data. If it falls outside the chosen grid, we extend the grid.

involving the lowest cost is then the optimal choice for $\underline{u}_{N-1}$ for the given $\underline{x}_{N-1}$. We again tabulate these results for each $\underline{x}_{N-1}$ of interest.

Now that the optimal cost and control are known for each $\underline{x}_{N-1}$ of interest, we can use these results to compute the optimal cost and control for each $\underline{x}_{N-2}$ of interest. And so on, until we have completed the table for each $\underline{x}_0$ of interest.

Entering this table at the given value of $\underline{x}_0$, we can follow through and determine the optimal control at each step of the way, and hence our problem is solved. We can also look at the results for other values of $\underline{x}_0$ with equal ease, and hence we have in effect obtained a family of optimal trajectories, with $\underline{x}_0$ as the parameter. Furthermore, if a certain final state must be reached (i. e., if it is required that $\underline{x}_{N+1} = \underline{b}$, where $\underline{b}$ is some given vector) then the process is unchanged except that in the final stage, only those values of $\underline{x}_N$ which are transformed (by means of the given difference equations) into the required $\underline{x}_{N+1} = \underline{b}$ by an allowable $\underline{u}$ need be considered. If more than one allowable $\underline{u}$ transforms a given $\underline{x}_N$ into the required $\underline{x}_{N+1}$, then the $\underline{u}$ with the lowest last-stage cost is entered in the table.

This process of solving the optimal control problem in question by solving a whole family of problems (a family which includes the given problem) is known as "imbedding", and is characteristic of the dynamic programming approach.

B. 4 Drawbacks and Limitations

Dynamic programming has certain drawbacks and limitations in addition to those already noted, which must be taken into account in determining its domain of practical applicability:

Care must be taken in converting a continuous-time problem to discrete time form. Too coarse a subdivision of the time axis results in excessive errors, and too fine a subdivision causes computing time to increase unreasonably. The same can be said for the choice of grid points in state space (the states for which costs and optimal control are tabulated) and in control space (the various controls which must be tested at each stage and for each state of interest in order to determine the optimal control and cost for that stage and state).

Computing time and/or computer memory requirements very rapidly become excessive as system dimension is increased. Other techniques also suffer from this "curse of dimensionality" (Bellman, Ref. 18, p. 94), although in varying degrees, depending on the type of problem.

A difficulty peculiar to dynamic programming is that referred to by Bellman (Ref. 18) as "the menace of the expanding grid": At some stage in the solution of a problem, the optimal control may transform the state to a point beyond the originally-chosen grid of points. In this event, a new point (or some new points) must be added to the grid and computations must be made at this point (or points) for some or all of the previously-completed columns of the table. These, in turn, may add new points to the grid, and so on. This "doubling-back" to recompute additional points may add considerably to program complexity, computing time, and memory requirements in a particular problem. There is no way to avoid this difficulty, in general.

APPENDIX C

THE FUNCTIONAL ANALYSIS APPROACH TO LINEAR TIME OPTIMAL PROBLEMS

The two principal methods of solution of optimization problems discussed above, dynamic programming and the calculus of variations, apply to problems with involving restrictions on the instantaneous values of the control inputs $\underline{u}(t)$, or perhaps no restrictions on $\underline{u}(t)$ at all. This does not begin to exhaust all the kinds of restrictions on $\underline{u}(t)$ that are of interest, however. For example, one might wish to place limits on the energy required to accomplish a certain task, or on the area under the curve $|u(t)|$, $t_o \leq t \leq t_1$. In many cases, such problems can be put into a form suitable for the application of one or more of the methods mentioned above.[63] However, the resulting nonlinear problems may be difficult to solve.

The functional analysis approach offers an alternative method of solution for certain problems of this type.

---

[63] For example, consider the minimum time problem with control energy E required to be less than or equal to some limit L, energy being defined here as $E = \int_{t_o}^{t_1} (\underline{u}(t), \underline{u}(t))\, dt$. Using the method of Valentine, we can define an additional state variable $x_{n+1}(t)$ by means of the equations $(\dot{x}_{n+1})^2 = \frac{L}{t_1 - t_o} - (\underline{u}(t), \underline{u}(t))$; $x_{n+1}(t_o) = 0$. The right-hand side of the differential equation integrates to $L - \int_{t_o}^{t_1} (\underline{u}(t), \underline{u}(t))\, dt = L - E$. The left-hand side integrates to a number either positive or zero. Thus, any solution to the augmented system of equations must satisfy the condition $E \leq L$, as desired. This is a problem to which classical techniques apply.

C. 1 The Linear Time-Optimal Problem

Krassovski (Ref. 30) has made use of certain of the results of Krein (Ref. 68) in functional analysis to solve a class of time-optimal control problems. This work was extended by Kulikowski (Refs. 31, 32, 33, 34), Kirillova (Ref. 35), and Kranc and Sarachik (Refs. 36, 37).

In its simplest form, the problem considered is that of forcing a system characterized by the equations $\dot{\underline{x}} = A(t)\,\underline{x} + \underline{b}(t)\,u$; (here $\underline{b}(t)$ is an n-vector and $u = u(t)$ = the scalar input or control variable), from its given initial state $\underline{x}(t_o)$ to the origin of the coordinate system in minimum time, subject to the constraint

$$\|u(t)\|_p = \left[\int_{t_o}^{t_1} |u(t)|^p \, dt\right]^{\frac{1}{p}} \leq 1; \qquad p > 1 \ .$$

The quantity $\|u(t)\|_p$ is called the p-norm of u(t). For p = 2, this is equivalent to an "energy" constraint or an "average power" constraint. As p approaches infinity, this condition approaches the amplitude-constraint condition $|u(t)| \leq 1$, which is a problem solvable with the aid of Pontryagin's maximum principle.

By means of Hölder's inequality, an expression for the optimal control is obtained,[64] as well as conditions under which it exists. Kirillova (Ref. 35) showed further that, as p approaches infinity, this result approaches a limit which he proved to be the optimal control for the limiting condition $|u(t)| \leq 1$. This limiting result is, of course, identical to that obtained using the maximum principle.

C. 2 Solution of the Linear Time-Optimal Problem

The problem to be solved here can be stated as follows (see Kranc and Sarachik, Ref. 36): Given a completely controllable[65] system characterized by the set of equations

---

[64] A more detailed description of this step is given in the next section.

[65] Complete controllability is defined in Appendix D. Also, see Kalman (Ref. 25).

$\dot{\underline{x}} = A(t)\,\underline{x} + B(t)\,\underline{u}$, where $\underline{x}$ is the n-dimensional state vector, $\dot{\underline{x}}$ is its time derivative, $\underline{u}$ is the r-dimensional control vector, and A(t) and B(t) are continuous n x n and n x r matrices, respectively. The system is in the given state $\underline{x}(t_o) = \underline{a}$ at initial time t_o. Choose the control vector $\underline{u}(t)$, $t_o \leq t \leq t_1$, from the class of bounded measurable functions satisfying

$$\|\underline{u}(t)\|_p = \left[\int_{t_o}^{t_1} \sum_{j=1}^{r} |u_j(t)|^p \, dt \right]^{\frac{1}{p}} \leq L ,$$

where p is a number greater than one, so that the k-dimensional output vector $\underline{y}(t) = D(t)\,\underline{x}(t)$ satisfies the desired final condition $\underline{y}(t_1) = \underline{b}$ and so that t_1, the final time, is minimum. D(t) is a continuous k x n matrix of rank k.

The solution of the system of differential equations can be written as

$$\underline{x}(t) = \mathbf{X}(t, t_o)\,\underline{a} + \int_{t_o}^{t} \mathbf{X}(t, s)\,B(s)\,\underline{u}(s)\,ds , \qquad (C.1)$$

where $\mathbf{X}(t, s)$ is the fundamental matrix (see Coddington and Levinson, Ref. 57) of the unforced system; i. e.,

$$\frac{d}{dt}\mathbf{X}(t, s) = A(t)\,\mathbf{X}(t, s); \quad \mathbf{X}(t, t) = \mathbf{I}$$

where I is the n x n identity matrix. In these terms, $\underline{y}(t)$ is given by

$$\underline{y}(t) = D(t)\,\mathbf{X}(t, t_o)\,\underline{a} + \int_{t_o}^{t} D(t)\,\mathbf{X}(t, s)\,B(s)\,\underline{u}(s)\,ds .$$

Define $V(t, s) = D(t)\,\mathbf{X}(t, s)\,B(s)$ and $\underline{g}(t) = \underline{b} - D(t)\,\mathbf{X}(t, t_o)\,\underline{a}$. Then the satisfaction of the final condition $\underline{y}(t_1) = \underline{b}$, is equivalent to the satisfaction of

$$\underline{g}(t_1) = \int_{t_o}^{t_1} V(t_1, s)\,\underline{u}(s)\,ds = \int_{t_o}^{t_1} \left[\underline{v}_1^{\,c}(t_1, s) \ldots \underline{v}_r^{\,c}(t_1, s)\right] \underline{u}(s)\,ds \qquad (C.2)$$

where $\underline{v}_i^{\,c}(t_1, s)$ is the ith column of V(t, s).

An intermediate step in the solution of this minimum-time constrained-norm problem is the solution of a class of fixed time minimum norm problems. That is, we assume that the final time t_1 is fixed, and obtain an expression (with t_1 as a parameter) for the control having minimum p-norm, subject to the constraint that the final conditions must be met.

It follows from Hölder's inequality (see Ref. 36) that this minimum-norm control can be written as

$$u_j(t) = \frac{\left|\underline{\bar{c}}\, \underline{v}_j^{\,c}(t_1, t)\right|^{q-1}}{\left[\|\underline{\bar{c}}\, V\|_q\right]^q} \operatorname{sgn}\left[\underline{\bar{c}}\, \underline{v}_j^{\,c}(t_1, t)\right] \qquad j = 1, \ldots, r$$

where $\underline{\bar{c}}$ is the constant k-element row vector which minimizes $\|\underline{\bar{c}}\, V\|_q$ subject to the constraint $\underline{\bar{c}}\, \underline{g}(t_1) = 1$. Here $q = \frac{p}{p-1}$,

$$\|\underline{\bar{c}}\, V\|_q = \left[\int_{t_o}^{t_1} \sum_{j=1}^{r} \left|\underline{\bar{c}}\, \underline{v}_j^{\,c}(t_1, t)\right|^q dt\right]^{\frac{1}{q}}$$

and sgn [Z] is a function which equals +1 if $Z > 0$, -1 if $Z < 0$, and is undetermined but less than one in absolute magnitude when $Z = 0$. The p-norm of this control is

$$\|\underline{u}(t)\|_p = \frac{1}{\|\underline{\bar{c}}\, V\|_q} .$$

The solution to the original minimum-time constrained-norm problem is now obtained[66] by varying t_1 and finding the smallest value of t_1 for which the above p-norm of $\underline{u}(t)$ is less than or equal to L, the given limit. If L is too small, there may be no control that satisfies these requirements.

The limiting cases p = 1 and p = infinity can also be solved by this method, if care is taken in interpretation and definition at various points in the derivation.

---

[66] This is stated without proof by Dranc and Sarachik (Ref. 36).

C. 3 The Minimum-Time Problem with Multi-Norm Constraints

Kranc and Sarachik (Ref. 37) have further extended these results to minimum-time problems in which the individual components of the control vector are constrained to have norms of various types, each norm less than or equal to its respective limit. That is,

$$\| u_i(t) \|_{p_i} = \left[\int_{t_o}^{t_1} |u_i(t)|^{p_i} dt \right]^{\frac{1}{p_i}} \leq L_i \qquad i = 1, \ldots, r \ .$$

The various p_i's and L_i's may be the same or different. These individual constraints are then incorporated into the single constraint equation

$$\| \underline{u}(t) \|_p = \left[\int_{t_o}^{t_1} \sum_{i=1}^{r} L_i^{-p} \left[\| u_i(t) \|_{p_i} \right]^p dt \right]^{\frac{1}{p}} \leq 1$$

and p is allowed to approach infinity. The limiting solution then satisfies all of the individual constraints.

Kreindler (Ref. 61) has given a comprehensive treatment of constrained linear time-optimal problems, using a functional analysis formulation coupled with a geometrical interpretation.

C. 4 Difficulties Arising in the Functional-Analysis Approach

This functional-analysis approach to linear time optimal problems offers solutions to a class of problems some of which are not readily solvable by other methods, but the task of carrying out the computations in an actual problem can be formidable. The difficult step is the minimization of the expression

$$\| \bar{\underline{c}} \ V \|_q = \left[\int_{t_o}^{t_1} \sum_{j=1}^{r} |\bar{\underline{c}} \ \underline{v}_j^{\,c}(t_1, t)|^q dt \right]^{\frac{1}{q}}$$

(or some more complicated expression) over all $\bar{\underline{c}}$ satisfying $\bar{\underline{c}} \ \underline{g}(t_1) = 1$. Furthermore,

this must be done for a range of values of t_1 sufficiently wide either to include the smallest value of t_1 for which $\|\underline{\bar{c}}\, V\|_q = L$, or else to show that such a case does not exist.

APPENDIX D

THE CONTROLLABILITY ASSUMPTION

The concept of controllability, introduced by Kalman (Refs. 25, 69, 70), is defined as follows:

Definition 1

Definition 1: An initial state $\underline{x}(t_o) = \underline{a}$ of a given linear system characterized by the system of equations

$$\dot{\underline{x}} = A(t)\,\underline{x} + B(t)\,\underline{u} \tag{D. 1}$$

is said to be controllable at time t_o if there exists a control function $\underline{u}(t)$, depending on $\underline{a}$ and on the initial time t_o and defined on some closed interval $T = [t_o, t_1]$ such that

$$\underline{x}(t_1) = \underline{0}\,.$$

Definition 2: If Definition 1 is satisfied for every state $\underline{a}$, the system is said to be completely controllable at time t_o.

Definition 3: If Definition 2 is satisfied for all t_o, the system is said to be completely controllable. It has been shown by Kalman (Ref. 25) that

Condition 1: a necessary and sufficient condition that the linear system characterized by (D. 1) be completely controllable at time t_o is that the matrix

$$\int_{t_o}^{t_1} \mathbf{X}(t_o, t)\, B(t)\, B^T(t)\, \mathbf{X}^T(t_o, t)\, dt$$

be positive definite for some $t_1 > t_o$. Here, as in Section 3.1, $\mathbf{X}(t, s)$ is the fundamental

matrix of the unforced system (Ref. 57), defined as the solution of

$$\frac{d}{dt}\,\mathbf{X}(t, s) = A(t)\,\mathbf{X}(t, s) \qquad \mathbf{X}(t, t) = \mathbf{I}\,.$$

These definitions and the equivalent condition are not in a form suitable for application here, for three reasons:

1) They are not stated in terms of the system considered in this work; that is, they do not include the "cost" weighting matrix G(t) or the output matrix D(t).

2) They are framed in terms of forcing the system to the origin (its only finite equilibrium point), whereas the problems considered in this work involve forcing the system from a general initial state to a general final manifold.

3) They involve an unspecified final time $t_1 > t_o$, but the problems considered here usually involve a specified final time.

Therefore, to simplify the discussion here, we shall make additional controllability definitions in terms appropriate to the problem at hand.

Definition 4: A set of boundary conditions

$$\underline{x}(t_o) = \underline{a} \qquad \underline{y}(t_1) = \underline{b}$$

for a given linear system characterized by the system of equations

$$\dot{\underline{x}} = A(t)\,\underline{x} + B(t)\,\underline{u}$$

$$\underline{y} = D(t)\,\underline{x} \qquad \text{(D. 2)}$$

is said to be controllable on the interval T if there exists a bounded measurable control $\underline{u}(t)$, depending on $\underline{a}$, $\underline{b}$, t_o, and t_1 and defined on T, such that the boundary conditions are satisfied.

Definition 5: If Definition 4 is satisfied for all $\underline{a}$ and $\underline{b}$, the system is said to be completely controllable on the interval T.

We obtain a necessary and sufficient condition for the satisfaction of Definition 5 by following essentially the same steps followed by Kalman (Ref. 25):

Condition 2: A necessary and sufficient condition that the linear system characterized by (D. 2) be completely controllable on the interval T is that the matrix

$$\int_T V(t_1, s)\, V^T(t_1, s)\, ds$$

be positive definite, where, as in Section 3. 1,

$$V(t_1, s) = D(t_1)\, \mathbf{X}(t_1, s)\, B(s)\, G^{-1}(s)$$

We prove sufficiency by assuming that Definition 2 is satisfied and then exhibiting a control which allows the boundary conditions to be satisfied: (As in Kalman's proof, this control turns out to be the minimum "energy" control, energy being defined as the integral of $|G(t)\, \underline{u}(t)|^2$.) Such a control can be obtained by setting p = 2 in Eq. 5. 4, using Eq. 5. 5 to eliminate C_p and $\underline{c}_p$, and applying the definition of $\underline{z}(t)$ given by Eq. 3. 5. This gives

$$\underline{u}(t) = G^{-1}(t)\, V^T(t_1, t) \left[\int_{t_o}^{t_1} V(t_1, s)\, V^T(t_1, s)\, ds \right]^{-1} \left[\underline{b} - D(t_1)\, \mathbf{X}(t_1, t_o)\, \underline{a} \right]$$

where, by assumption, the inversion is an allowed operation. This control causes the required final condition $\underline{y}(t_1) = \underline{b}$ to be satisfied, as can be verified by substitution in the expression for $\underline{y}(t)$ given by Eq. 3. 2.

The proof of the necessity of Condition 2 also parallels Kalman's proof:

Assume that Condition 2 is not satisfied but that the system is completely controllable on the interval T. Under the first assumption, there must exist some vector $\overline{c} \neq \underline{0}$ such that

$$\overline{c}^T \left[\int_T V(t_1, s)\, V^T(t_1, s)\, ds \right] \overline{c} = 0$$

Define $\underline{\bar{u}}(t) = G^{-1}(t)\, V^T(t_1, t)\, \underline{\bar{c}}$. Then,

$$\underline{\bar{c}}^T \left[\int_{t_o}^{t_1} V(t_1, s)\, V^T(t_1, s)\, ds \right] \underline{\bar{c}} = \int_{t_o}^{t_1} |G(s)\, \underline{\bar{u}}(s)|^2\, ds = 0$$

which in turn implies that $\underline{\bar{u}}(t) = \underline{0}$ a. e. on T, since G(t) is an invertible matrix for all $t \in T$.

From the second assumption, we know that there must exist some bounded measurable control which causes any specified set of boundary conditions to be satisfied, and therefore that there must exist a bounded measurable control (call it $\underline{u}^*(t)$) which causes the particular set of conditions $\underline{x}(t_o) = \underline{0}$, $\underline{y}(t_1) = \underline{c}$ to be satisfied. Thus, from Eq. 3.2,

$$\underline{\bar{c}} = \int_T D(t_1)\, \mathbf{X}(t_1, s)\, B(s)\, \underline{u}^*(s)\, ds = \int_T V(t_1, s)\, G(s)\, \underline{u}^*(s)\, ds$$

$$\underline{\bar{c}}^T \underline{\bar{c}} = |\underline{\bar{c}}|^2 = \int_T \underline{\bar{c}}^T\, V(t_1, s)\, G(s)\, \underline{u}^*(s)\, ds = \int_T \underline{\bar{u}}^T(s)\, G^T(s)\, G(s)\, \underline{u}^*(s)\, ds\,.$$

Since $\underline{\bar{u}}(t) = \underline{0}$ a. e. on T, from above, it follows that

$$|\underline{\bar{c}}|^2 = 0\,.$$

But this contradicts the initial assumption that $\underline{\bar{c}} \neq \underline{0}$, thus proving the necessity of Condition 2.

APPENDIX E

CERTAIN PROPERTIES OF THE SPACES L_p^r AND L_∞^r

E. 1 Completeness

The results of Chapters III and V depend on the fact that certain function spaces defined there are Banach spaces. The purpose of this appendix is to establish that these spaces are indeed Banach spaces, and to give a representation theorem for bounded linear functionals on the space Z (referred to in this appendix as L_∞^r) defined in Chapter III.

The spaces considered are:

a) the spaces L_p^r of measurable r-component vector functions of time $\underline{z}(t)$ defined on the interval $T = [t_o, t_1]$ for which

$$\int_T |\underline{z}(t)|^p \, dt < \infty \qquad 1 \leq p < \infty$$

with norm given by

$$\|\underline{z}\|_p = \left[\frac{1}{T} \int_T |\underline{z}(t)|^p \, dt \right]^{\frac{1}{p}} \tag{E. 1}$$

where $|\underline{z}(t)|$ is the Euclidean norm of $\underline{z}(t)$.

b) the space L_∞^r of essentially bounded measurable r-component vector functions of time $\underline{z}(t)$ defined on the interval T, with norm

$$\|\underline{z}\|_\infty = \operatorname*{ess\,sup}_{t \in T} |\underline{z}(t)| \tag{E. 2}$$

Using the usual definitions of vector addition and scalar multiplication of the elements $\underline{z}(t)$ of L_p^r or L_∞^r, it is easy to show that these are linear vector spaces. Likewise it

is easy to show from the properties of the Euclidean norm involved in the definition of $\|\underline{z}\|_p$ and $\|z\|_\infty$ that Eqs. E. 1 and E. 2 define norms on the corresponding spaces. To show that L_p^r and L_∞^r are Banach spaces it only remains to be shown that these spaces are complete.

Lemma: The space L_p^r is complete for $1 \leq p < \infty$.

Proof: First we prove that a function $\underline{z}(t)$ is in L_p^r a) if and b) only if each of its components $z_i(t)$ is in $L_p = L_p(t_o, t_1)$.

a) Let $z_i(t)$ be in L_p for $i = 1, \ldots, r$. Then each $z_i(t)$ is a measurable function for which the quantity

$$\|z_i\|_p = \left[\frac{1}{T}\int_T |z_i(t)|^p \, dt\right]^{\frac{1}{p}}$$

is defined and bounded. By repeated application of Minkowski's inequality (Natanson, Ref. 79, p. 198), we can show that

$$\sum_{i=1}^{r}\left[\frac{1}{T}\int_T |z_i(t)|^p \, dt\right]^{\frac{1}{p}} \geq \left[\frac{1}{T}\int_T \left[\sum_{i=1}^{r} |z_i(t)|\right]^p dt\right]^{\frac{1}{p}}$$

From the properties of the Euclidean norm

$$\sum_{i=1}^{r} |z_i(t)| \geq \left[\sum_{i=1}^{r} z_i^2(t)\right]^{\frac{1}{2}} = |\underline{z}(t)| \, ,$$

so that

$$\sum_{i=1}^{r}\left[\frac{1}{T}\int_T |z_i(t)|^p \, dt\right]^{\frac{1}{p}} = \sum_{i=1}^{r} \|z_i\|_p \geq \left[\frac{1}{T}\int_T |\underline{z}(t)|^p \, dt\right]^{\frac{1}{p}} = \|\underline{z}\|_p$$

Since each $\|z_i\|_p$ is bounded, $\|z\|_p$ is also, and hence $\underline{z}(t)$ is in L_p^r.

b) Let $\underline{z}(t)$ be in L_p^r. Then

$$\left[\frac{1}{T}\int_T |\underline{z}(t)|^p \, dt\right]^{\frac{1}{p}} < \infty$$

Since $|\underline{z}(t)| = \left[\sum_{i=1}^{r} z_i^2(t) \right]^{\frac{1}{2}} \geq |z_i(t)|$ for i = 1, ... r, it follows at once that

$$\left[\frac{1}{T} \int_T |z_i(t)|^p \, dt \right]^{\frac{1}{p}} = \|z_i\|_p \leq \|z\|_p < \infty$$

so that each $z_i(t)$ is in L_p.

We now complete the proof of the lemma by showing that every Cauchy sequence in L_p^r converges to an element of L_p^r: Consider any Cauchy sequence $[\underline{z}_n(t)]_{n=1}^{\infty}$ in L_p^r. Then for every $\epsilon > 0$ there exists an $N = N(\epsilon)$ such that for all m and n greater than N

$$\left[\frac{1}{T} \int_T |\underline{z}_m(t) - \underline{z}_n(t)|^p \, dt \right]^{\frac{1}{p}} < \epsilon$$

Since, by the above-noted property of the Euclidean norm, $|z| \geq |z_i|$, i=1, ..., r, it follows that for all m and n greater than N

$$\left[\frac{1}{T} \int_T |z_{mi}(t) - z_{ni}(t)|^p \, dt \right]^{\frac{1}{p}} < \epsilon$$

where $z_{mi}(t)$ and $z_{ni}(t)$ are the ith components of $\underline{z}_m(t)$ and $\underline{z}_n(t)$, respectively. Thus the original Cauchy sequence in L_p^r corresponds to r Cauchy sequences in L_p. Since L_p is known to be a Banach space, and hence a complete space (Dunford and Schwartz, Ref. 85), each of the sequences $[z_{ni}(t)]_{n=1}^{\infty}$ converges to a point in L_p, which we denote by $z_{0i}(t)$, i = 1, ..., r. Let $\underline{z}_0(t) = [z_{01}(t), \ldots, z_{0r}]^T$. Then, by part a) of this proof, $\underline{z}_0(t)$ is in L_p^r. Q. E. D.

Lemma: The space L_∞^r is complete.

Proof: The proof of this lemma is identical to that of the previous lemma, with the appropriate limiting form of Minkowski's inequality being used.

E. 2 A Representation Theorem

Theorem: Each bounded linear functional $z(\underline{v})$ defined on the space L_1^r of r-component row vector functions of time can be represented in the form

$$z(\underline{v}) = \int_T \underline{v}(t)\, \underline{z}(t)\, dt$$

where $\underline{z}(t)$ is an essentially bounded r-component column vector function of time defined on the interval T. Furthermore,

$$\|z\| = \operatorname*{ess\,sup}_{t \in T} |\underline{z}(t)| = \|\underline{z}\|_\infty$$

Proof: Consider elements of L_1^r of the form $\underline{v}(t) = [0, \ldots, 0, v_i(t), 0, \ldots, 0]$. When restricted to such a subspace of L_1^r, the bounded linear functional $z(\underline{v})$ must be equivalent to some bounded linear functional $z_i(v_i)$ operating on elements $v_i(t)$ of L_1, since if $\underline{v}(t)$ is in L_1^r then $v_i(t)$ is in L_1. This same argument applies to each of the components of $\underline{v}(t)$, which leads to the conclusion that $z(\underline{v})$ can be written as the summation of r bounded linear functions $z_i(v_i)$ operating on the components $v_i(t)$ of $\underline{v}(t)$, i. e.,

$$z(\underline{v}) = \sum_{i=1}^{r} z_i(v_i)$$

From the well-known representation theorem for L_1 spaces (Ref. 85, p. 289), each of these r functionals $z_i(v)$ is representable in the form $z_i(v) = \int_T v(t)\, z_i(t)\, dt$ where $z_i(t)$ is an essentially bounded measurable function defined on T. Combining these two results then gives

$$z(\underline{v}) = \sum_{i=1}^{r} \int_T v_i(t)\, z_i(t)\, dt = \int_T \sum_{i=1}^{r} v_i(t)\, z_i(t)\, dt = \int_T \underline{v}(t)\, \underline{z}(t)\, dt$$

where $\underline{z}(t)$ is the r-component column vector function $\underline{z}(t) = [z_1(t), \ldots, z_r(t)]^T$. Since each component of $\underline{z}(t)$ is essentially bounded, $|\underline{z}(t)|$ is essentially bounded, and hence $\underline{z}(t)$ is an element of L_∞^r.

It remains to be shown that $\|z\| = \|\underline{z}\|_\infty$: From the definition of the norm of a functional and from the linearity of $z(\underline{v})$,

$$\|z\| = \sup_{\|\underline{v}\|_1 \leq 1} |z(\underline{v})| = \sup_{\underline{v}(t) \in L_1^r} \frac{\left| \int_T \underline{v}(t)\, \underline{z}(t)\, dt \right|}{\|\underline{v}\|_1} \qquad \text{(E. 3)}$$

From the extention of Hölder's inequality derived in the lemma[67] of Section 3. 4, we have that

$$\|\underline{z}\|_\infty = \operatorname*{ess\,sup}_{t \in T} |\underline{z}(t)| \geq \frac{\left| \int_T \underline{v}(t)\, \underline{z}(t)\, dt \right|}{\|\underline{v}\|_1} \qquad \text{(E. 4)}$$

with equality if and only if

$$\begin{aligned} \underline{v}(t) &= K(t)\, \underline{z}^T \qquad \text{a. e. on } T_2 \\ \underline{v}(t) &= \underline{0} \qquad\quad \text{a. e. on } T - T_2 \end{aligned} \qquad \text{(E. 5)}$$

where K(t) is any function in L_1 and T_2 is the subset of T on which $|\underline{z}(t)| = \operatorname*{ess\,sup}_{t \in T} |\underline{z}(t)|$. (The condition for equality stated in the referenced lemma is framed in terms of a given $\underline{v}(t)$ and an unspecified $\underline{z}(t)$. In the evaluation of $\|z\|$, $\underline{z}(t)$ is given and $\underline{v}(t)$ is varied. Therefore, the condition for equality has been reframed to fit this situation.) Note that $\underline{v}(t)$, as given by Eq. E. 5, is in L_1^r, since $|\underline{z}(t)|$ is essentially bounded. The point of all this is that the largest value that can be attained by the right-hand side of Eq. E. 4 over all $\underline{v}(t)$ in L_1^r can not exceed $\|\underline{z}\|_\infty$, and that this supremum is indeed taken on by an element of L_1^r, as given by Eq. E. 5. Thus $\|z\| = \|\underline{z}\|_\infty$, as claimed. Q. E. D.

---

[67] No circular reasoning is involved here, even though the representation theorem being proven here is applied in a section prior to the section in which this lemma is derived, because the proof of the lemma does not depend on this representation theorem or even on the fact that the spaces involved are Banach spaces.

REFERENCES

1. A. A. Feldbaum, "Optimal Processes in Automatic Control Systems," Automatika i Telemekhanika, Vol. 14, No. 5, 1953.
2. A. A. Feldbaum, "On the Question of Synthesizing Optimal Automatic Control Systems," Trans. 2nd All-Union Conf. on Automatic Control Theory, Vol. 2, Izd-vo, Akad. Nauk SSSR, 1955
3. F. A. Mikhailov, "Integral Indicators of Automatic Control Systems Quality," Chap. 25 of Automatic Control Fundamentals, V. V. Solodovnikov, ed., Mashgiz, 1954.
4. R. Bellman, I. Glicksberg, and O. Gross, "On the 'Bang-Bang' Control Problem," Quart. Appl. Math., Vol. 14, pp. 11-18, 1956.
5. Ostwald, Klassider der exakten Wissenschaften, No. 46, 1894.
6. G. A. Bliss, "The Problem of Mayer with Variable Endpoints," Trans. Am. Math. Soc., Vol. XIX, 1918.
7. O. Bolza, "Über den abnormelen Fall biem Lagrangeschen and Mayerschen Problem mit gemischten Bedingungen and variabeln Endpunkten," Math. Ann., Vol. LXXIV, 1913.
8. G. A. Bliss, Lectures on the Calculus of Variations, Univ. of Chicago, 1946.
9. F. A. Valentine, "The Problem of Lagrange with Differential Inequalities as Added Side Conditions," Contributions to Calculus of Variations 1933-1937, Univ. of Chicago, 1937.
10. V. G. Boltyanskii, R. V. Gamkrelidze, L. S. Pontryagin, "On the Theory of Optimum Processes," Dokl. Akad. Nauk SSSR, Ser. Math., Vol. 110, pp. 7-10, 1956.
11. V. G. Boltyanskii, "The Maximum Principle in the Theory of Optimal Processes," Dokl. Akad. Nauk SSSR, Vol. 119, No. 6, 1958.
12. V. G. Boltyanskii, R. V. Gamkrelidze, L. S. Pontryagin, "The Theory of Optimal Processes I. The Maximum Principle," Izvest. Akad. Nauk SSSR, Ser. Math., Vol. 24, pp. 3-42, 1960.
13. L. S. Pontryagin, et al., The Mathematical Theory of Optimal Processes, Interscience, 1962.

14. R. V. Gamkrelidze, "Optimal Processes with Bounded Phase Coordinates," Izv. Akad. Nauk SSSR, Ser. Math., Vol. 24, pp. 315-356, 1960.

15. R. V. Gamkrelidze, lectures given in University of Michigan seminar on optimal control, September and October, 1964, and at the Symposium on the Mathematical Theory of Optimal Control, Ann Arbor, Oct. 1964.

16. L. I. Rozonoer, "L. S. Pontryagin's Maximum Principle in the Theory of Optimum Systems," (3 parts), Automatika i Telemekhanika, Vol. 20, Nos. 10,11, 12, 1959.

17. R. Bellman, Dynamic Programming, Princeton, 1957.

18. R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton, 1961.

19. L. D. Berkovitz, "Variational Methods in Problems of Control & Programming," J. Math. Anal. & Appl., Vol. 3, pp. 145-169, August 1961.

20. L. D. Berkovitz, "On Control Problems with Bounded State Variables," Rand Corp. Memorandum RM-3207-PR, July 1962.

21. A. Larsen, "Optimal Control and the Calculus of Variations," Ser. 60, Issue 462, Electronics Res. Lab., Univ. of Cal., 1962.

22. R. E. Mortensen,"A Synopsis of Optimal Control Theory," Ser. 60, Issue 487, Electronics Res. Lab., Univ. of Cal., 1962.

23. R. E. Kopp, "Pontryagin Maximum Principle," in Optimization Techniques, G. Leitmann,ed., Academic Press, 1962.

24. S. E. Dreyfus, "Dynamic Programming and the Calculus of Variations," J. Math. Anal. & Appl. Vol. 1, No. 2, 1960.

25. R. E. Kalman, "Contributions to the Theory of Optimal Control," Boletin Soc. Math. Mex., Vol. 5, pp. 102-119, 1961.

26. R. E. Kalman, "The Theory of Optimal Control and the Calculus of Variations," Mathematical Optimization Techniques, Univ. of California Press, pp. 309-331, 1963.

27. "Study of Nonlinear Mechanics," RIAS Final Report, AFOSR Contract AF 49(638)-382, July 1963.

28. E. Roxin, "A Geometric Interpretation of Pontryagin's Maximum Principle," Tech. Report 61-15, RIAS, Dec. 1961.

29. I. Flügge-Lotz and H. Halkin, "Pontryagin's Maximum Principle and Optimal Control," Tech. Report 130, Div. of Eng. Mech., Stanford Univ., Sept. 1961.

30. N. N. Krassovski, "On the Theory of Optimum Regulation," Automation & Remote Control, Vol. 18, pp. 1005-1016, Nov. 1957.

31. R. Kulikowski, "On Optimal Control with Constraints," Bull. Polish Acad. Sci. (Ser. Tech. Sci.), Vol. 7, pp. 285-294, Apr. 1959.

32. R. Kulikowski, "Concerning the Synthesis of the Optimum Nonlinear Control," Bull. Polish Acad. Sci. (Ser. Tech. Sci.), Vol. 7, pp. 391-399, June 1959.

33. R. Kulikowski, "Synthesis of a Class of Optimum Control Systems," Bull. Polish Acad. Sci. (Ser. Tech. Sci.), Vol. 7, pp. 663-671, Nov. 1959.

34. R. Kulikowski, "Optimizing Processes and Synthesis of Optimizing Automatic Control Systems with Nonlinear Invariable Elements," Proc. I. F. A. C. Moscow Conf., pp. 473-477, 1960.

35. R. M. Kirillova, "A Limiting Process in the Solution of an Optimal Control Problem," J. Appl. Math. & Mech., Vol. 24, pp. 398-405, 1960.

36. G. M. Kranc and P. E. Sarachik, "An Application of Functional Analysis to the Optimal Control Problem," ASME Paper 62-JACC-4, 1962.

37. G. M. Kranc and P. E. Sarachik, "On Optimal Control of Systems with Multi-Norm Constraints," Paper No. I-4, JACC, 1963.

38. L. Markus and E. B. Lee, "On the Existence of Optimal Controls," Paper No. 61-JAC-2, JACC, 1961.

39. J. Hadamard, "Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques en castrées," Mem. prés. acad. sci. France [2] 33, No. 4, 1908.

40. F. D. Faulkner, "Direct Methods," in Optimization Techniques, G. Leitmann, ed., Academic Press, 1962.

41. A. E. Bryson and W. F. Denham, "A Steepest Ascent Method for Solving Optimum Programming Problems," J. Appl. Mech., Vol. 29, Ser. E, No. 2, pp. 247-257, 1962.

42. A. E. Bryson, and W. F. Denham, "The Solution of Optimal Programming Problems with Inequality Constraints," IAS Paper No. 63-78, IAS Annual Meeting, New York, Jan. 1963.

43. H. J. Kelley, "Gradient Theory of Optimal Flight Paths," J. Am. Rocket Soc., Vol. 30, pp. 947-953, 1960.

44. H. J. Kelley, "Method of Gradients," in Optimization Techniques, G. Leitmann, ed., Academic Press, 1962.

45. R. G. Graham, "A Steepest-Ascent Solution of Multiple-Arc Vehicle Optimization Problems," Report No. TDR-269(4550-20)-3, Contract No. AF 04(695)-269, Aerospace Corp., El Segundo, Cal., Dec. 1963.

46. L. A. Zadeh, "Optimality and Non-Scalar-Valued Performance Criteria," IEEE Trans. on Automatic Control, Vol. AC-8, No. 1, Jan. 1963.

47. R. Sivan, "The Necessary and Sufficient Conditions for the Optimal Controller to be Linear," Session XI, Paper 2, JACC, 1964.

48. Z. V. Rekasius and T. C. Hsia, "On the Inverse Problem in Optimal Control," Session XI, Paper 4, JACC, 1964.

49. A. A. Feldbaum, "Dual Control Theory," (4 parts), Automatika i Telemekhanika, Vol. 21, Nos. 9 and 11, 1960, Vol. 22, Nos. 1 and 2, 1961.

50. R. E. Kalman and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Trans. ASME, J. Basic Eng., Ser. D, pp. 95-108, Mar. 1961.

51. H. J. Kushner, "Near Optimal Control in the Presence of Small Stochastic Perturbations," Session XIV, Paper 4, JACC, 1964.

52. A. M. Letov, "Analytic Controller Design," Automatika i Telemekhanika, Vol. 21, Nos. 4, 5, and 6, 1960.

53. B. Friedland, "The Design of Optimal Controllers for Linear Processes with Energy Constraints," Melpar Technical Note 62/2, Mar. 1962.

54. C. Sprague, "On the Reticulation Problem in Multivariable Control Systems," Session XVII, Paper 4, JACC, 1964.

55. G. J. Coviello, An Organization Approach to the Optimization of Multivariable Systems," Session XVII, Paper 2, JACC, 1964.

56. J. L. Sanders, "Multi-Level Control," Session XVII, Paper 6, JACC, 1964.

57. E. A. Coddington and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill, 1955.

58. N. I. Akhiezer and I. B. Glazman, Theory of Linear Operators in Hilbert Space, (English transl.), Ungar, 1961.

59. J. P. LaSalle, "The Time Optimal Control Problem," Contrib. to the Theory of Nonlinear Oscillations, Vol. 5, Princeton Univ. Press, 1959.

60. A. R. Stubberud, "A Controllability Criterion for a Class of Linear Systems," Wescon Paper 12.1, 1963.

61. E. Kreindler, "Contributions to the Theory of Time-Optimal Control," J. Franklin Inst., Vol. 275, No. 4, Apr. 1963.

62. A. Meile, "The Calculus of Variations in Applied Aerodynamics and Flight Mechanics," in Optimization Techniques, G. Leitmann, ed., Academic Press, 1962.

63. R. Weinstock, Calculus of Variations, McGraw-Hill, 1952.

64. H. Halkin, "Liapunov's Theorem on the Range of a Vector Measure and Pontryagin's Maximum Principle," Tech. Report 109, Appl. Math. and Stat. Lab., Stanford Univ., May 1962.

65. H. Halkin, "On the Necessary Conditions for Optimal Control of Nonlinear Systems," Tech. Report 116, Appl. Math. and Stat. Lab., Stanford Univ., June 1963.

66. R. E. Kalman, "The Theory of Optimal Control and the Calculus of Variations," RIAS Tech. Report 61-3, 1961.

67. G. Leitmann, "An Elementary Derivation of the Optimal Control Conditions," Lockheed Tech. Report 6-90-61-84, Oct. 1961.

68. M. Krein, "The L-Problem in Abstract Linear Normed Spaces," from Akheiser & Krein, On Some Questions of the Theory of Moments (Russian), 1938.

69. R. E. Kalman, Y. C. Ho, and K. S. Narendra, "Controllability of Linear Dynamical Systems," Contributions to Differential Equations, Vol. 1, Aug. 1962.

70. R. E. Kalman, "Mathematical Description of Linear Dynamical Systems," RIAS Tech. Report 62-18, Nov. 1962.

71. R. Bellman, Introduction to Matrix Analysis, McGraw-Hill, 1960.

72. N. I. Akheizer, The Calculus of Variations, (English transl. by A. H. Frink), Blaisdell, 1962.

73. I. M. Gelfand and S. V. Fomin, Calculus of Variations, (English transl. by R. A. Silverman), Prentice-Hall, 1963.

74. J. T. Tou, Modern Control Theory, McGraw-Hill, 1964.

75. R. E. Kalman, T. S. Englar, and R. S. Bucy, "Fundamental Study of Adaptive Control Systems," Tech. Report No. ASD-TR-61-27, Vol. 1, Flight Control Lab., Aero. Systems Div., Air Force Systems Command; Apr. 1962.

76. S. S. L. Chang, "Sufficient Conditions for Optimal Control of Linear Systems with Nonlinear Cost Functions", Paper No. XI-1, 1964 JACC; Stanford, Cal.

77. L. W. Neustadt, "Optimization, a Noise Problem, and Nonlinear Programming", SIAM J. on Control, Vol. 2, No. 1, 1964, pp. 33-53.

78. N. N. Krasovski, "On the Theory of Optimum Control", J. Appl. Math. & Mech., Vol. 23, No. 4, 1959, pp. 899-919 (in English transl.).

79. I. P. Natanson, Theory of Functions of a Real Variable, v. I, (English transl. by Boron and Hewitt), Ungar, 1961.

80. G. F. Simmons, Introduction to Topology and Modern Analysis, McGraw-Hill, 1963.

81. A. E. Taylor, Introduction to Functional Analysis, Wiley, 1958.

82. W. Rudin, Principles of Mathematical Analysis, McGraw-Hill, 1953.

83. W. A. Porter & J. P. Williams, "Minimum Effort Control of Linear Dynamic Systems," Inst. Sci. & Tech. Univ. of Mich., Aug. 1964.

84. E. Hille & R. S. Phillips, Functional Analysis and Semi-Groups, Am. Math. Soc. Colloq. Pub. Vol. 31, Am. Math. Soc., 1957.

85. N. Dunford and J. T. Schwartz, Linear Operators I, Interscience (2nd printing, 1964).

DISTRIBUTION LIST

No. of Copies	
2	Commanding Officer, U. S. Army Electronics Command, U. S. Army Electronics Laboratories, Fort Monmouth, New Jersey, Attn: Senior Scientist, Electronic Warfare Division
1	Commanding General, U. S. Army Electronic Proving Ground, Fort Huachuca, Arizona, Attn: Director, Electronic Warfare Department
1	Commanding General, U. S. Army Materiel Command, Bldg. T-7, Washington 25, D. C., Attn: AMCRD-DE-E-R
1	Commanding Officer, Signal Corps Electronics Research Unit, 9560th USASRU, P. O. Box 205, Mountain View, California
1	U. S. Atomic Energy Commission, 1901 Constitution Avenue, N.W., Washington 25, D. C., Attn: Chief Librarian
1	Director, Central Intelligence Agency, 2430 E Street, N.W., Washington 25, D. C., Attn: OCD
1	U. S. Army Research Liaison Officer, MIT-Lincoln Laboratory, Lexington 73, Massachusetts
1	Commander, Air Force Systems Command, Andrews Air Force Base, Washington 25, D. C., Attn: SCSE
1	Headquarters, USAF, Washington 25, D. C., Attn: AFRDR
1	Commander, Aeronautical Systems Division, Wright-Patterson Air Force Base, Ohio, Attn: ASRNCC-1
1	Commander, Aeronautical Systems Division, Wright-Patterson Air Force Base, Ohio, Attn: ASAPRD
1	Commander, Aeronautical Systems Division, Wright-Patterson Air Force Base, Ohio, Attn: ASRN-CS
1	Commander, Aeronautical Systems Division, Wright-Patterson Air Force Base, Ohio, Attn: ASNP
1	Commander, Electronic Systems Division, L. G. Hanscom Field, Bedford, Massachusetts
1	Commander, Rome Air Development Center, Griffiss Air Force Base, New York, Attn: RAYLD
1	Commander, Air Proving Ground Center, Eglin Air Force Base, Florida, Attn: ADJ/Technical Report Branch

DISTRIBUTION LIST (Cont.)

No. of Copies	
1	Chief of Naval Operations, EW Systems Branch, OP-35, Department of the Navy, Washington 25, D. C.
1	Chief, Bureau of Ships, Code 691C, Department of the Navy, Washington 25, D. C.
1	Commander, Bu Naval Weapons, Code RRRE-20, Department of the Navy, Washington 25, D. C.
1	Commander, Naval Ordnance Test Station, Inyokern, China Lake, California, Attn: Test Director - Code 30
1	Commander, Naval Air Missile Test Center, Point Mugu, California
1	Director, Naval Research Laboratory, Countermeasures Branch, Code 5430, Washington 25, D. C.
1	Director, Naval Research Laboratory, Washington 25, D. C., Attn: Code 2021
1	Director, Air University Library, Maxwell Air Force Base, Alabama, Attn: CR-4987
1	Commanding Officer-Director, U. S. Navy Electronic Laboratory San Diego 52, California
1	Commanding Officer, U. S. Naval Ordnance Laboratory, Silver Spring 19, Maryland
3	Chief, U. S. Army Security Agency, Arlington Hall Station, Arlington 12, Virginia, 22212 Attn: 2 Cpys - IADEV 1 Copy - EW Div. IATOP
1	President, U. S. Army Defense Board, Headquarters, Fort Bliss, Texas
1	President, U. S. Army Airborne and Electronics Board, Fort Bragg, North Carolina
1	U. S. Army Anti-Aircraft Artillery and Guided Missile School, Fort Bliss, Texas, Attn: ORL
1	Commander, USAF Security Service, San Antonio, Texas, Attn: CLR
1	Chief of Naval Research, Department of the Navy, Washington 25, D. C., Attn: Code 427
1	Commanding Officer, 52d U. S. Army Security Agency, Special Operations Command, Fort Huachuca, Arizona
1	President, U. S. Army Security Agency Board, Arlington Hall Station, Arlington 12, Virginia
1	The Research Analysis Corporation, McLean, Virginia, 22101 Attn: Document Control Officer

DISTRIBUTION LIST (Cont.)

No. of Copies	
10	Headquarters, Defense Documentation Center, Cameron Station, Alexandria, Virginia
1	Commanding Officer, U. S. Army Electronics Research and Development Laboratory, Fort Monmouth, New Jersey, Attn: U. S. Marine Corps Liaison Office, Code: SIGRA/SL-LNR
1	Director, Fort Monmouth Office, Communications-Electronics Combat Developments Agency, Building 410, Fort Monmouth, New Jersey
7	Commanding Officer, U. S. Army Electronics Command, U. S. Army Electronics Laboratories, Fort Monmouth, New Jersey, Attn: 1 Copy - Director of Research 1 Copy - Technical Documents Center - ADT/E 1 Copy - Chief, Special Devices Branch, Electronic Warfare Division 1 Copy - Chief, Advanced Techniques Branch, Electronic Warfare Division 1 Copy - Chief, Jamming and Deception Branch, Electronic Warfare Division 1 Copy - File Unit No. 2, Mail and Records, Electronic Warfare Division 1 Copy - Chief, Vulnerability Br., Electromagnetic Environment Division
1	Commanding Officer, U. S. Army Signal Missile Support Agency, White Sands Missile Range, New Mexico, Attn: SIGWS-MEW
1	Commanding Officer, U. S. Naval Air Development Center, Johnsville, Pennsylvania, Attn: Naval Air Development Center Library
1	Headquarters, Aeronautical Systems Division, Wright-Patterson Air Force Base, Ohio, Attn: ASRNCC-10
1	U. S. A. F. Project Rand, The Rand Corporation, 1700 Main Street, Santa Monica, California
1	Stanford Electronic Laboratories, Stanford University, Stanford, California
1	Director, National Security Agency, Fort George G. Meade, Maryland, Attn: RADE-1
1	Bureau of Naval Weapons Representative, Lockheed Missiles and Space Company, P. O. Box 504, Sunnyvale, California
1	Dr. B. F. Barton, Director, Cooley Electronics Laboratory, The University of Michigan, Ann Arbor, Michigan
55	Cooley Electronics Laboratory, The University of Michigan, Ann Arbor, Michigan

Above distribution is effected by Electronic Warfare Division, Surveillance Department, USAEL, Evans Area, Belmar, New Jersey. For further information contact Mr. I. O. Myers, Senior Scientist, Telephone 59-61252.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Cooley Electronics Laboratory University of Michigan Ann Arbor, Michigan	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP ----

3. REPORT TITLE

Minimum Peak Amplitude Control

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report No. 164

5. AUTHOR(S) (Last name, first name, initial)

Waltz, F. M.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1965	201	85

8a. CONTRACT OR GRANT NO. DA36 039 AMC-03733(E) b. PROJECT NO. 1PO 21101 AO42 01 Task No. -01 c. d. Sub Task No. -02	9a. ORIGINATOR'S REPORT NUMBER(S) 06137-7-T 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) ----

10. AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC.
This report has been released to CFSTI.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
---	U. S. Army Electronics Command Fort Monmouth, N. J. Attn: AMSEL-RD-SE

13. ABSTRACT

This report considers a class of optimal control problems in which the input signal to a given system is to be chosen so as to cause the output of the system to satisfy specified conditions and so that the peak value of the input over the operating interval is a minimum. For cases in which the given system is linear, a theorem guaranteeing the existence of an optimal input and a theorem giving the form of this optimal input are presented, as well as computational algorithms for obtaining numerical values of optimal inputs. Two approaches to minimum peak amplitude problems in which the given system in nonlinear are presented: the first makes use of a certain time-optimal problem, and the second involves a limiting process using an L_p-space norm in place of the originally-specified cost functional.

DD FORM 1 JAN 64 1473

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Optimal Control Functional Analysis Linear-Non Linear Problem Computational Algorithms Pontryagin's Maximum Principle Dynamic Programming						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.