ASD-TDR-63-119

PROCEEDINGS OF THE OPTIMUM SYSTEM SYNTHESIS CONFERENCE

11 - 13 SEPTEMBER 1962

TECHNICAL DOCUMENTARY REPORT NO. ASD-TDR-63-119

February 1963

Flight Control Laboratory
Aeronautical Systems Division
Air Force Systems Command
Wright-Patterson Air Force Base, Ohio

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Qualified requesters may obtain copies of this report from the Armed Services Technical Information Agency, (ASTIA), Arlington Hall Station, Arlington 12, Virginia.

This report has been released to the Office of Technical Services, U. S. Department of Commerce, Washington 25, D. C., in stock quantities for sale to the general public.

Copies of ASD Technical Documentary Reports should not be returned to the Aeronautical Systems Division unless return is required by security considerations, contractual obligations, or notice on a specific document.

26-613M, 900, 5-1-63

FOREWORD

The papers contained herein were presented at the Optimum System Synthesis Conference, Wright-Patterson AFB, Ohio, 11 - 13 September 1962, sponsored by the Aeronautical Systems Division. The program chairman was Mr. L. Schwartz of the System Optimization Section, Aerospace Mechanics Branch, Flight Control Laboratory. Mr. Schwartz was assisted by Lt. R. O. Sickeler, of the same section.

The selection of the papers and speakers was the responsibility of the individual session chairmen, Prof. C. T. Leondes of UCLA, Prof. D. Graham of Princeton University, Dr. J. P. LaSalle of RIAS, and Mr. E. L. Peterson of the General Electric Company.

The program chairman is especially indebted to Mr. Cannon of the Special Activities Division, Directorate of Technical Operations for his invaluable assistance.

ABSTRACT

The proceedings contain a collection of 16 papers presented at the Optimum System Synthesis Conference held at the Aeronautical Systems Division 11 - 13 September 1962. The meeting was directed toward defining the present position of optimum system synthesis and determining guides for future research in both applications and theory. Most papers are concerned with various aspects of recent applications and theoretical developments in optimal control such as steepest descent techniques, suboptimal controllers, optimum filtering, and functional analysis techniques. Some earlier results are also discussed.

PUBLICATION REVIEW

This technical documentary report is a collection of the papers presented at the Optimum System Synthesis Conference and is published to make this information more widely available.

C. B. WESTBROOK
Chief, Aerospace Mechanics Branch
Flight Control Laboratory

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

INTRODUCTION

Since the Optimum System Synthesis Conference was a limited-attendance meeting, which was not meant to be widely publicized, many readers will be interested in some background information. The major aim of the conference was to bring together some prominent mathematicians and engineers working in the various phases of optimum system synthesis to discuss both what has been accomplished and what needs to be done to make optimum synthesis a practical reality. Three sessions were set aside for the presentation of papers; these papers constitute this volume.

In addition to the symposium sessions there was a panel and forum discussion of the general topic, "Is optimum synthesis practical?", and an evening session devoted to informal discussions by small groups. Each group had a particular subject to discuss, which lasted until either the subject or the group was exhausted.

Admittedly, the question, "Is optimum synthesis practical?" is much too ill-defined to admit of a concise reply. However, some attempt should be made to summarize the prevailing opinion -- A task one or two levels less difficult than predicting the wave form of white noise. Despite the clamor for control system optimization techniques, it appears that quite often the "optimum" is a chimera, a situation which results in what might be called The Great Optimization Paradox. On the one hand we find an ever-growing body of theory and technique for solving variational problems of one sort or another. On the other hand we have the problems of the control system designer, many of which cannot or, perhaps, should not be cast in the form of variational problems. Yet there seems to be a vigorous attempt by both sides to get together, even though it is not entirely clear that there is any extensive common ground.

The major difficulty is, apparently, our inability to clearly and unambiguously specify control requirements. Often, the critical factor is not performance, but cost, weight, or reliability. When we seek to "optimize" a control system, we do not have a precise way to characterize the optimum. How much performance are we willing to trade for a given amount of reliability? Indeed, how do we measure performance? These are far from trivial questions; the utility of the powerful synthesis techniques depends upon meaningful answers, answers which are not readily available.

Not all control problems are so ambiguous, particularly the outer-loop or path problems (usually associated with guidance, although guidance and control are the obverse and reverse of the same coin). In cases where there is a clear, quantitative requirement (e. g., minimum time, minimum fuel, minimum expected miss distance) the techniques hold a great deal of promise, and we can find examples of actual optimizations. It is when we misapply the syntheses to ill-defined problems that we suffer our worst failures.

Another obstacle to practicality is the complexity of the optimal control law, often referred to as the problem of orbiting a 7090. Quite often the criterion functional that is being optimized is not too sensitive to changes in the control law, and we can do "almost as well" with a much simpler system. The only other answer seems to be picominiaturization of computing equipment, or similar equipment breakthroughs.

As a final point, the panel-forum session evidenced a healthy trend towards finding out what can and/or cannot be done with optimum synthesis techniques. The practitioner is now attempting to apply the theory.

ASD-TDR-63-119

Feedback Design and Optimal Control Theory

By

John G. Truxal
Peter Dorato

Polytechnic Institute of Brooklyn, New York

Feedback Design and Optimal Control Theory

by
John G. Truxal and Peter Dorato
Polytechnic Institute of Brooklyn

ABSTRACT

The optimization techniques of Pontryagin (maximum principle) and Bellman (dynamic programming) are evaluated as feedback design tools. The first part of the paper is devoted to a discussion of the advantages of optimal closed-loop (feedback) operation. The second part of the paper is devoted to an analysis of the feedback nature of the two optimization techniques cited and an analysis of the current computational difficulties in applying these techniques.

FEEDBACK DESIGN AND OPTIMAL CONTROL THEORY

John G. Truxal and Peter Dorato
Polytechnic Institute of Brooklyn

INTRODUCTION

Recently a great deal of interest has been generated in the application of modern optimization techniques to problems in control. In the past the term "control" has been associated with closed-loop (feedback) systems. Yet the current literature on optimal control theory is largely devoted to open-loop systems. There appears to be some limitation to the application of the modern optimization techniques to the feedback design problem. The purpose of this paper is to evaluate such presently available optimization techniques as the maximum principle and dynamic programming as tools for the design of feedback systems. To simplify the discussion the plant, object being controlled, is assumed to have dynamics given by the first order scalar equation

$$\dot{x} = f(x, u) \tag{1}$$

where x represents the plant output and u the control input. The performance criterion is assumed to be given by

$$S = \int_{t_o}^{T} F(x, u)dt \tag{2}$$

The object of optimal control theory is to determine a control input u which causes S to be an extremal. When the optimal control law, denoted u^o, is obtained as a function of the initial output, i.e.,

$$u^o = \phi\left[x(t_o);\ t\right] \tag{3}$$

the system is said to be operating open-loop. When the optimal control is obtained in terms of the current value of output, i.e.,

$$u^o = \psi\left[x(t);t\right] \tag{4}$$

the system is said to be operating closed-loop.

The first part of the paper is devoted to a discussion of the advantages of closed-loop operation. The second part is a study of the present status of optimization theory as a feedback design technique.

ADVANTAGES OF CLOSED-LOOP CONTROL

Classically feedback systems have been employed to solve the problems of

(1) Plant dynamics modification
(2) Sensitivity reduction

It is fairly easy to demonstrate in linear time-invariant systems that feedback does provide a convenient means of improving system dynamics and reducing plant parameter sensitivity (ref. 1). In nonlinear systems the advantages of feedback are not as easily demonstrated, at least not theoretically. If the system performance is measured in terms of relative stability then such techniques as equivalent linearization, describing functions, and Lyapunov's second method may be used to evaluate the effectiveness of feedback as a plant-dynamics modification tool. To show that a given feedback control law, say $u = \psi(x)$, reduces the sensitivity of the output with respect to a plant parameter w, one might proceed as follows. Express the plant dynamics with parameter w as

$$\dot{x} = f(x, u; w) \tag{5}$$

and define the sensitivity measure as

$$s = \frac{\partial x}{\partial w} \tag{6}$$

From the theory of differential equations with parameters, the function s is known to satisfy the equation (ref. 2)

$$\dot{s} = \frac{\partial f}{\partial x} s + \frac{\partial f}{\partial w} \tag{7}$$

with the initial condition $s(t_o) = 0$. With no feedback, interpreted in this case to mean $u = 0$, obtain a solution of equation (7). Denote this solution s_o. Denote by s_f the solution of (7) with feedback applied, i.e., with $u = \psi(x)$. The sensitivity improvement may then be evaluated, for example, from the relative values of $|s_f|_{max}$ and $|s_o|_{max}$. Note that equation (7) is linear and time-varying and requires a solution of the original nonlinear equation (5) via the term $\partial f/\partial x$. A discussion of computer solutions of (5) and (7) has been given by Miller and Murrary (ref. 3). However to obtain closed form solutions of (5) and (7) is, except for very special cases, almost impossible. Hence the question of sensitivity reduction in nonlinear systems via feedback is still an open question.

In optimal control systems the feedback is applied to modify the plant dynamics in an optimal way, hence plant dynamics modification is automatically taken into account. The stability of optimal closed-loop systems is generally assured by the properties of the performance criterion chosen(ref. 4). The result follows from the fact that S^o, where S^o denotes the value of S when optimal feedback is employed, serves as a Lyapunov function for the optimal closed-loop system. Thus if $F(x, u^o)$, where $u^o = \psi(x(t);t)$, is positive definite, then the Hamilton-Jacobi equations of the optimal path guarantee that $\dot{S}^o$ is negative definite; indeed the Hamilton-Jacobi equation requires that

$$\frac{dS^o}{dt} = -F(x, u^o) \tag{8}$$

In optimal closed-loop systems then, stability does not play the dominant role it does in classical feedback theory. One clear advantage of closed-loop control is that the current state of the output is used to generate the control input u. Thus if the output is suddenly displaced at time t' by some disturbance, closed-loop operation assures optimal operation in the remainder of the optimization interval (t', T), assuming no further disturbances. A similar disturbance in an open-loop system destroys the optimality of operation in the interval (t', T) since the control input is preprogrammed for a fixed initial output $x(t_o)$. A compromise between closed-loop and open-loop operation can be achieved with the use of a periodically updated (with respect to the value of $x(t_o)$) open-loop control law. This approach has some practical advantages which are discussed in the next section. However beyond the point just demonstrated it is difficult to illustrate further advantages of closed-loop optimal control. Although a sensitivity analysis similar to the one previously indicated can be performed, the optimization techniques in current use do not assure that feedback results in a lower sensitivity value than open-loop operation[+], in spite of the fact that experimental studies seem to indicate that feedback in optimal systems does reduce sensitivity. Thus the sensitivity problem in optimal control systems is much like the sensitivity problem in classical nonlinear feedback systems, very little can be said theoretically about the relative merits of feedback as a means of reducing sensitivity. The only distinct advantage of closed-loop optimal control then appears to be in terms of possible output disturbances.

OPTIMIZATION THEORY AS A FEEDBACK DESIGN TECHNIQUE

As is well known the optimal closed-loop control law may be obtained, at least in theory, from Pontryagin's maximum principle in the following steps (ref. 5).

(1) Determine the function u which maximizes, for a minimum of S, the Hamaltonian

$$H = pf(x,u) - F(x,u) \tag{9}$$

Note that the maximizing u, denoted u^o, is a function of both x and p, and hence may be written

$$u^o(t) = \alpha[p(t), x(t)] \tag{10}$$

(2) Solve the set of equations

$$\dot{x} = f(x,u)\Big|_{u = u^o} \tag{11a}$$

$$\dot{p} = -\frac{\partial f}{\partial x}\Big|_{u = u^o} + \frac{\partial F}{\partial x}\Big|_{u = u^o} \tag{11b}$$

[+]Indeed a basic requirement for optimization is complete knowledge of the plant dynamics, in the deterministic case, and complete knowledge of the pertinent statistics, in the stochastic case.

subject to the boundary conditions

$$x(t_o) = x_o \tag{12a}$$

$$p(T) = 0 \tag{12b}$$

It is assumed here that the final time T is fixed while the final output x(T) is free. Note that the solution for p is a function of the initial output $x(t_o)$ and therefore may be written

$$p(t) = \beta\,(x_o; t) \tag{13}$$

(3) Substitute expression (13) back into (10) and let $t = t_o$ to obtain

$$u^o(t_o) = \alpha\,[\beta\,(x_o\,;\,t_o),\ x_o] = \psi\,[x(t_o);\ t_o\,] \tag{14}$$

If now one allows t_o in expression (14) to take on arbitrary values in the original interval of optimization (t_o, T), the closed-loop control law (4) is obtained.

As is also well known, the basic difficulty in applying the above procedure is the problem of obtaining a solution to the nonlinear two-point boundary value problem of step 2. Basically what is required to obtain the closed-loop control law is a mapping of the values of $x(t_o)$ into values of $p(t_o)$. This suggests that arbitrary values of x(T) be chosen and then that equations (11a) and (11b) be run backwards a length of time $T-t_o$ at which time the values of $x(t_o)$ and $p(t_o)$ are recorded for the required mapping. Of course this procedure must be repeated until all the appropriate values of $x(t_o)$ are spanned and all values of time in the interval (t_o, T) are exhausted. The computer time and storage required for such a procedure is however prohibitive, for most realistic problems. A current approach to the problem is to use some steepest descent technique to obtain an open-loop control law and then update from time to time the value of $x(t_o)$ (ref. 6, 7, 8). The success of such a method depends on how quickly the particular steepest descent technique used converges and how often the output data must be updated. Further application of the maximum principle to feedback design depends largely on the development of numerical techniques for the solution of nonlinear two-point boundary value problems.

In some special cases the maximum principle does yield an explicit solution for the optimal closed-loop control law. A notable example is the case of a linear plant with unconstrained input and integral-quadratic performance criterion (ref. 9,10).

Dynamic programming is a feedback design technique by its very nature.[+] Unfortunately, however, dynamic programming also suffers from some basic computational difficulties. To illustrate the feedback nature of dynamic programming and also its computational difficulties the technique is briefly outlined.

[+] Of course in the determinstic case any feedback solution can always be converted, via the plant dynamics, into an open-loop solution. The basic feedback nature of dynamic programming is inescapable, however, in the study of stochastic systems. See, for example, Bellman (ref. 11), sec. 10.4.

First assume that the plant dynamics and performance criterion have been made discrete in time, thus consider

$$x_{k+1} = g(x_k, u_k) \tag{15}$$

and

$$S = \sum_{k=1}^{N} G(x_k, u_k) \tag{16}$$

where, for convenience, the notation $x(k\Delta t) = x_k$ is used. Denote by S_k^o the value of S which results from an optimal input policy u_k over the interval (k, N) and note that S_k^o is a function only of x_k, hence may be written $S_k^o(x_k)$. The optimal control law is then determined in the following steps.

(1) Start at the end of the process and search for a value of u_N which minimizes, assuming one wants to minimize S, the function $G(x_N, u_N)$. Note that u_N is a function of x_N. Let

$$u_N = \psi_N(x_N) \tag{17}$$

and evaluate

$$S_N^o(x_N) = \min_{u_N} G(x_N, u_N) \tag{18}$$

(2) Proceed one step backwards and minimize, with respect to u_{N-1}, the function $\left[G(x_{N-1}, u_{N-1}) + S_N^o(x_N)\right]$, where $x_N = g(x_{N-1}, u_{N-1})$. The results of this step then yields

$$u_{N-1} = \psi_{N-1}(x_{N-1}) \tag{19}$$

and

$$s_{N-1}^o(x_{N-1}) = \min_{u_{N-1}} \left[G(x_{N-1}, u_{N-1}) + S_N^o\left[g(x_{N-1}, u_{N-1})\right] \right] \tag{20}$$

(3) Proceed backwards, iterating the above procedure, until the entire interval (1, N) is exhausted, to obtain the closed-loop control law

$$u_k = \psi_k(x_k), \quad 1 \leq k \leq N \tag{21}$$

Note that the minimization technique just described requires, at any stage k, a search for u_k in terms of all possible values of x_k and storage, for use in the next stage, of the function $S_k^0(x_k)$. In addition all the functions $\psi_k(x_k)$ must also be stored. For practical problems the amount of storage required exhausts the capacity of most modern day computers. The application of dynamic programming as a feedback design technique then seems to depend critically on the development of methods which reduce the amount of storage required and/or the design of computers with larger storage capacity.

Even if the computational difficulties of the maximum principle and dynamic programming are solved, there still remains the problem of implementing the closed-loop control law with currently available hardware. In this connection it should be noted that the control law is generally a nonlinear function of the state of the plant. Here two separate problems arise: one is the problem of measuring the state of the plant, which often requires the measurement of high order derivatives, and the other is the problem of generating the required nonlinear function to the plant state. If a computer (digital) is used as the controller element in the feedback loop, then the computer must accomodate the storage of the control law $\psi_k(x_k)$ for all possible values of x_k and for all k in the range (1, N). One possible solution to this aspect of the problem is the use of "approximations in policy space" which yield simpler control laws, at the price of suboptimal operation (ref. 11,12).

In spite of the difficulties cited above, optimal control theory does provide an organized approach to the design of complex feedback systems. The computational difficulties encountered should, in fairness, be weighted against the rather sophisticated nature of the problem considered. Indeed in many applications, especially nonlinear stochastic systems, the alternative to optimal control theory is no theory at all.

REFERENCES

1. J. G. Truxal, Automatic Feedback Control System Snythesis, McGraw-Hill New York, N. Y., 1955.

2. L. S. Pontryagin, Ordinary Differential Equations, Addison-Wesley, Reading, Mass., 1962, section 22.

3. K. S. Miller and F. J. Murray, "Mathematical Basis for an Error Analysis of Differential Analyzers." J. Math. Phys., vol. 32, nos. 2-3, July-October, 1958.

4. E. G. Al'brekht, "On Optimum Stabilization of Nonlinear Systems," Jour. of Applied Math and Mech., vol. 25, no. 5, 1961.

5. L. I. Rozonoer, "L. S. Pontryagin's Maximum Principle in the Theory of Optimum Systems, I, II, III," Automation and Remote Control., vol. 20, June, July, August, 1960.

6. H. J. Kelly, "Gradient Theory of Optimal Flight Paths," Jour. Amer. Rocket Soc., vol. 30, October, 1960.

7. A. V. Balakrishnan and H. C. Hsieh, "Function Space Methods in Control Systems Optimization," Proceedings of the Optimum System Synthesis Conference held at Wright-Patterson Air Force Base, Dayton, Ohio, September 1962.

8. A. E. Bryson, W. F. Denham, F. J. Carroll, and K. Mikami, "Determination of Lift or Drag Programs to Minimize Re-Entry Heating," Jour. of the Aerospace Sciences, vol. 29, no. 4, April 1962, pp 420-430.

9. R. E. Kalman, The Theory of Optimal Control and the Calculus of Variations, RIAS Technical Report 61-3.

10. G. L. Collina, and P. Dorato, Application of Pontryagin's Maximum Principle: Linear Control Systems, Polytechnic Institute of Brooklyn, Report AFOSR-2655, June, 1962.

11. R. Bellman, Adaptive Control Processes, Princeton U. Press, Princeton, N. J., 1961.

12. M. Aoki, "On Optimal and Suboptimal Policies in the Choice of Control Forces for Final-Values Systems." IRE International Convention, Session No. 1, March, 1960.

ASD-TDR-63-119

Function Space Methods in Control Systems Optimization

By

A. V. Balakrishnan
H. C. Hsieh

University of California
Department of Engineering
Los Angeles, California

FUNCTION SPACE METHODS IN CONTROL SYSTEMS OPTIMIZATION

A.V. Balakrishnan
H.C. Hsieh

Department of Engineering
University of California
Los Angeles

I. INTRODUCTION

Much of the current work on control systems optimization has centered on the establishment of necessary conditions that the optimum solution must satisfy. In many cases these conditions take the form of a functional equation or a vector partial differential equation that the solution must satisfy. Much less volume of work has been devoted to the equally important problem of obtaining efficient computational algorithms based on successive iteration schemes which converge to the optimal solution. [1, 2, 3, 4]. This paper deals with such a procedure introduced in [3] for a class of deterministic control problems using methods of abstract functional analysis which afford a succinct and general solution in many cases and offer a point of departure for problems where any systematic analysis of a general enough nature would appear to be difficult.

The computational method employed here is the method of steepest descent in Hilbert space, an early version of which is due to Kantorovich [5]. In setting the control problem in the function space format, we make use, for instance, the Theory of (Frechet) analytic functions over abstract spaces, specialized to Hilbert space [6].

In Section II, we consider the general problem for linear systems and show how it can be set entirely as a function-space problem leading to very general results. In Section III, we specialize to linear dynamic systems exploiting the compactness of the operation involved. The computation method is explained in Section IV. The specialization to the final value problem for linear systems is contained in Section V. In Section VI, extensions to nonlinear systems are indicated, and finally in Section VII, results of some computer studies based on the proposed method are given.

II. FORMULATION FOR LINEAR SYSTEMS

We begin by considering a linear system, since it is important in its own right and, moreover, provides a point of departure for the extension to nonlinear systems treated in Section VI. Let x(t) represent the output vector of a linear system as a result of the applied control vector u(t). Then x(t) will be related functionally to u(t) often, though not necessarily, by means of a differential equation. Let $x_d(t)$ represent the desired response. The optimization problem is one of minimizing some functional of the difference

$$e(t) = x(t) - x_d(t)$$

The criterion chosen in this paper is the integral of the squared magnitude over some finite interval which we can normalize to [0, T]. Thus we wish to determine u(t) which minimizes

$$\int_0^T \| e(t) \|^2 \rho(t)\, dt\,; \quad \| e(t) \|^2 = e(t)^* e(t)$$

Here $\rho(t)$ is a non-negative weight function and the asterisk denotes the adjoint. To avoid the language of δ-functions, we assume more generally that we have a finite measure m_1 over the Borel sets of [0, T] and we wish to minimize

$$\int_\pi \| e(t) \|^2 dm_1$$

where π denotes the interval [0, T]. Here e(t) will be continuous by virtue of x(t) and $x_d(t)$ being continuous. Usually, there will, in addition, be some constraint on the control vector such as

$$\int_\pi \| u(t) \|^2 dm_2 \leq M^2$$

where m_2 is another finite measure on the Borel sets of π, and again u(t) will be Borel-measurable. By taking the measure m_1 to be a jump at T, we specialize to the "final value problem". But we shall see that this class of problems can be solved constructively with no restriction on m_1 and m_2, other than finiteness. For this it is convenient to introduce the real Hilbert space (actually L_2 space) H_1 of n-dimensional functions square integrable over π with respect to measure m_1 and H_2, the L_2 space of m-dimensional functions square integrable over π with respect to measure m_2. Then x(t), $0 \leq t \leq T$, will be in H_1 and the control vectors can be taken in H_2. Moreover, for a linear system, we have

$$x(t) = \int_0^T W(t, \sigma)\, u(\sigma)\, d\sigma \qquad (1)$$

Of course, for physical realizability, unessential restriction for our problem is

$$x(t) = \int_0^t W(t, \sigma)\, u\,(\sigma)\, d\sigma \tag{2}$$

Such a relation may come from a differential equation but not necessarily so. Now, assuming $W(t, \sigma)$ to be continuous for instance, (1) defines a linear bounded operator from H_2 into H_1. We denote this operator by L, so that we represent (1) by

$$L\,u = x; \quad u \in H_2,\ x \in H_1$$

In this frame then, we want to minimize[†]

$$|| Lu - g ||_1^2$$

The subscripts 1 and 2 denote the norms in H_1 and H_2 respectively. "g" is a given element in H_1 and minimization is over a sphere of radius M in H_2, or more generally it can be over any closed convex set (with non-empty interior usually). The minimization can be made slightly more general by adding another perturbing term and considering

$$C(u) = ||Lu - g||_1^2 + \left[Ku, u \right]_2 \tag{3}$$

where K is a linear bounded non-negative operator on H_2 into H_2. For example, it may be multiplication by a non-negative number corresponding to a "Lagrange multiplier", or more generally it may be multiplication by a positive-definite matrix K(t). Let L^* be the adjoint of L, then L^* maps H_1 and H_2. We can then rewrite C(u) as

$$C(u) = \left[(L^*L + K)\, u, u \right]_2 - 2 \left[u, L^*g \right]_2 + \left[g, g \right]_1$$

Now L^*L is also non-negative, so that so is the sum

$$L^*L + K = R \tag{4}$$

Assume that K is such that

$$\left[Ku, u \right] > 0 \text{ whenever } L^*Lu = 0$$

[†]More generally, we can consider the minimization of

$$\left[P(Lu-g), Lu-g \right]_1^2$$

where P is a linear bounded positive definite operator on H_1 into H_1.

Then R has a bounded Linear inverse. Let

$$R u_o = \left[L^* L + K \right] u_o = L^* g \tag{5}$$

That u_o is then the unique optimal minimizing solution is apparent from [see Reference 4]

$$C(u) = C[u_o] + \left[R(u-u_o),\ u - u_o \right]_2 \tag{6}$$

and the second term is non-negative. This establishes $C(u_o)$ as the unique absolute minimum, whereas the fact that it is a local minimum may be shown in various ways. Moreover, the optimal solution u_o is of the form by using (5)

$$K u_o = L^* \left[g - L u_o \right]$$

or

$$u_o = (L K^{-1})^* \left[g - L u_o \right]$$

If K is just a positive multiplication k, this reduces to

$$u_o = L^* \frac{[g - L u_o]}{k} \tag{7}$$

so that u_o itself is in the range of L^*. Hence using this and setting

$$u_o = L^* y$$

in (5), we obtain an equation in H_1.

$$L L^* y + k y = g$$

This yields a simplification if H_1 is finite dimensional, as would be the case if m_1 has a jump at T.

The main reason for introducing K is the application to the constraint problem: Suppose the convex set Δ is the class of all u such that

$$\| u \|_2 \leqq M$$

Then we have the following theorem: There is a non-increasing sequence of positive number k_n such that

$$\left[L^* L + k_n I \right]^{-1} L^* g = u_n$$

I being the identity operator, where $u_n \in \Delta$ and

$$\lim_n \| L u_n - g \|^2 = \inf_{u \in \Delta} \| L u - g \|^2 \tag{8}$$

The proof of this theorem uses (6) and may be found in [4]. Here we shall sketch it for completeness. Suppose there is an element u in Δ such that

$$L^* L u = L^* g \tag{9}$$

Then it is clear that we can find a decreasing sequence of positive numbers k_n such that

$$u_n = \left[L^* L + k_n I \right]^{-1} L^* g$$

belongs to Δ. Since for each n,

$$\left[L^* L + k_n I \right]$$

has a bounded inverse, then

$$\| u_n \| \leqq \| [L^* L + k_n I]^{-1} L^* L \| \quad \| u \| \leqq \| u \|$$

Thus using (6) we have

$$C(u) = C(u_n) + k_n \left[u_n, u - u_n \right]$$

and from which it follows that, as $k_n \to 0$

$$C(u_n) \to C(u)$$

as required. If there is no u in Δ such that (9) holds, then we can clearly find positive k_o such that

$$\left[L^* L + k_o I \right]^{-1} L^* g = u_o$$

and

$$\| u_o \| = M$$

It may be shown as in [4] that this element yields the unique minimum, since for any $k < k_o$

$$\left\| \left[L^* L + k_o I \right]^{-1} L^* g \right\| > M$$

and hence is not in Δ. In particular, we note that the optimal element is always in the range of L^*, or possibly its closure.

III. THE MINIMIZATION PROBLEM FOR LINEAR DYNAMIC SYSTEMS

In this section, we shall specialize our discussion to the minimization problem for a linear dynamic system. We shall take m_1 and m_2 to be Lebesgue measure. Suppose now that the dynamic system is described by a differential equation:

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

where A(t) is an n x n matrix and B(t) is an n x m matrix. We shall have

$$x(t) = x_o(t) + \int_o^T W(t, \tau)\, u(\tau)\, d\tau \tag{10}$$

with

$$W(t, \tau) = 0 \quad \text{for } \tau > t$$

Here $W(t, \tau)$ is an n x m weighting function matrix of the system and $x_o(t)$ is the output due to initial condition. We can now define a linear operation L from H_2 into H_1 as

$$v(t) = \int_o^T W(t, \tau)\, u(\tau)\, d\tau \; ; \; L u = v$$

$$0 \leqq t \leqq T$$

The main simplification that we shall exploit in this section is the compactness of the operator L by virtue of

$$\int_o^T \int_o^T \left\| W(t, \tau) \right\|^2 dt\, d\tau < \infty$$

Here the adjoint L^* mapping H_1 into H_2:

$$h(\tau) = \int_o^T W^*(t, \tau)\, g(t)\, dt; \; L^* g = h$$

$$0 \leqq \tau \leqq T$$

Now

$$C(u) = \|L u - g\|_1^2$$

$$= \left[Lu, Lu \right]_1 - 2 \left[g, Lu \right]_1 + \|g\|_1^2$$

$$= \left[L^* Lu, u \right]_2 - 2 \left[L^* g, u \right]_2 + \|g\|_1^2$$

where

$$g(t) = x_d(t) - x_o(t).$$

Denoting $L^* L$ by R, we note that R is non-negative, self-adjoint and an integral operator with kernel

$$G(\tau_1, \tau_2) = \int_o^T W^*(t, \tau_1)\, W(t, \tau_2)\, dt$$

It is also compact and thus is characterized by a countable number of non-negative eigenvalues. Moreover, for any u in H_2, we have

$$R u = \sum_{i=1}^{\infty} \lambda_i \left[u, \phi_i \right] \phi_i$$

where $\{\lambda_i\}$ are the set of non-zero eigenvalues and $\{\phi_i\}$ are the set of corresponding orthonormalized eigenfunctions, and $\{\lambda_i\}$ are arranged to be monotone nonincreasing.

We can now phrase the minimization problem entirely in terms of the eigenfunctions. For this we note that for any u in H_2, we have the unique orthogonal decomposition

$$u = \sum_{i=1}^{\infty} a_i \phi_i + u_o \tag{11}$$

where

$$a_i = \left[u, \phi_i \right]; \left[L u_o, L u_o \right] = \left[R u_o, u_o \right] = 0$$

Similarly, since $\{L\phi_i\}$ are orthogonal, we have the orthogonal decomposition

$$g = \sum_{i=1}^{\infty} \frac{[g, L\phi_i]}{\lambda_i} L\phi_i + g_o$$

$$L^* g_o = 0 \; ; \quad \sum_{i=1}^{\infty} \frac{[g, L\phi_i]^2}{\lambda_i} < \infty$$

Hence we have

$$C(u) = \sum_{i=1}^{\infty} \lambda_i a_i^2 - 2 a_i [g, L\phi_i] + ||g||_1^2$$

and completing the square in each term

$$C(u) = \sum_{i=1}^{\infty} \lambda_i \left[\left(a_i - \frac{[g, L\phi_i]}{\lambda_i} \right)^2 - \frac{[g, L\phi_i]^2}{\lambda_i^2} \right] + ||g||_1^2$$

$$\geq - \sum_{i=1}^{\infty} \frac{[g, L\phi_i]^2}{\lambda_i} + ||g||_1^2 = ||g_o||^2$$

Hence it is clear that if we set

$$u_n = \sum_{i=1}^{n} \frac{[g, L\phi_i]}{\lambda_i} \phi_i$$

We have that

$$\lim_n C(u_n) = ||g_o||^2 = \operatorname{Inf}_u C(u)$$

and

$$\lim_n R u_n = L^* g$$

A necessary and sufficient condition for an element h in H_2 such that

$$\operatorname{Inf}_n C(u) = C(h)$$

is then of course that

$$\sum_{i=1}^{\infty} \frac{[g, L\phi_i]^2}{\lambda_i^2} < \infty$$

and in that case

$$h = \sum_{i=1}^{\infty} \frac{[g, L\phi_i]}{\lambda_i} \phi_i \quad ; \qquad Rh = L^* g$$

Suppose now, in addition, that h is such that

$$\operatorname{Inf}_u C(u) = 0$$

Then we notice that $g_o = 0$ and

$$L h = g$$

and what is more important, for any u such that

$$L u = g$$

we have

$$\left[u - h, \phi_i \right] = 0 \text{ for every } i$$

and

$$||u|| \geqq ||h||$$

by virtue of the orthogonal decomposition (11). If we want

$$\operatorname{Inf}_u C(u) = 0$$

for every g in H_i, then of course we must have that $\{L\phi_i\}$ is complete in H_1. On the other hand, the more common problem is one which constraints u to be in some bounded convex set. Suppose we enclose this set in a sphere of radium M and consider, instead, u such that

$$||u||^2 \leqq M^2$$

First of all, if we take a $k > 0$ and minimize

$$C(u) = ||Lu - g||_1^2 + k\,||u||_2^2 \tag{12}$$

we have in terms of the decomposition (11)

$$C(u) = \sum_{i=1}^{\infty} (\lambda_i + k) \left[\left(a_i - \frac{[g, L\phi_i]}{\lambda_i + k} \right)^2 - \frac{[g, L\phi_i]^2}{(\lambda_i + k)^2} \right]$$

$$+ ||g||_1^2 + k\,||u_o||_2^2$$

Hence setting

$$a_i = \frac{[g, L\phi_i]}{\lambda_i + k} \quad ; \; u_o = 0$$

We note that

$$[R + k]\, u = (R + k) \sum_{i=1}^{\infty} \frac{[g, L\phi_i]}{\lambda_i + k} \phi_i$$

$$= \sum_{i=1}^{\infty} [g, L\phi_i]\, \phi_i = L^* g$$

or

$$u = [R + k]^{-1} L^* g$$

To answer the constraint problem, we may look upon (12) as the usual perturbation using a Lagrange multiplier, and hence all we have to do is to use the smallest k so that

$$\sum_{i=1}^{\infty} \frac{[g, L\phi_i]^2}{(\lambda_i + k)^2} = M$$

unless of course

$$\sum_{i=1}^{\infty} \frac{[g, L\phi_i]^2}{\lambda_i^2} < M$$

That this does yield the absolute minimum has already been shown generally in Section II.

The derivation given above for obtaining the optimal control depends on the knowledge of the eigenvalues and eigenfunctions of the operator R. Since they are, in general, infinite in number, this approach is not suitable for solving the synthesis problems. In the next section, we shall present a method of steepest descent in Hilbert space, which will overcome these difficulties.

IV. THE METHOD OF STEEPEST DESCENT FOR MINIMIZING A QUADRATIC FUNCTIONAL

In this section we shall indicate a computational method based on the steepest descent method in Hilbert space for solving the minimization problem associated with a quadratic functional. An early exposition of such a method is due to Kantorovich [5]. The essential idea of this method is contained in the following: In seeking the minimum of a quadratic functional Q(u), an arbitrary initial guess u_o is assumed. We obtain the gradient at this

point, i.e., finding an element z such that $\frac{d}{d\epsilon}\left(Q(u_o + \epsilon z)\right)$ will be maximized at $\epsilon = 0$. Let z_o be such an element. Since $Q(u_o + \epsilon z_o)$ is a second degree polynomial of ϵ, it will attain a minimum for some ϵ_o. Then the element $u_1 = u_o + \epsilon_o z_o$ will be adopted as the next approximation and the whole procedure can be repeated as many times as accuracy required.

A geometrical interpretation of this method is fairly clear. In the space where Q(u) is defined, the surfaces Q(u) = C are, in general, a family of similar ellipsoids with center at the minimum point. The initial approximation u_o shall lie on a certain ellipsoid of the family. From this initial point, we shall move along the direction of the gradient at this point, i.e., along the normal to this ellipsoid. We shall reach a point u_1 where the value of Q(u) is the least on this normal, i.e., a point where the normal line is tangent to some ellipsoid of the family. From this point u_1, we shall go along the new normal to this ellipsoid $Q(u) = Q(u_1)$, and the whole process is then repeated.

The quadratic functional to be minimized has the general form

$$Q(u) = \left[Tu, u\right]_2 - 2\left[h, u\right]_2 \tag{13}$$

where T is a positive definite or a non-negative definite operator in H_2. Let us now substitute u by $u_o + \epsilon z$. Then

$$Q(u_o + \epsilon z) = Q(u_o) + 2\epsilon\left[T u_o - h, z\right]_2 + \epsilon^2\left[Tz, z\right]_2 \tag{14}$$

In seeking the direction of the gradient, it is necessary to find z such that

$$\frac{d}{d\epsilon}\left(Q(u_o + \epsilon z)\right)_{\epsilon=o} = 2\left[T u_o - h, z\right]_2$$

is maximized. According to Schwarz's inequality, it will be achieved if

$$z_o = T u_o - h \tag{15}$$

Here R u - L^*g shall be referred to as the gradient of the functional Q. With the choice of z_o as given by (15), (14) will attain its minimum if

$$\epsilon_o = - \frac{\| z_o \|_2^2}{\left[T z_o, z_o\right]_2} \tag{16}$$

By repeating the above procedure, it is evident that at the nth step, we have

$$u_n = u_{n-1} + \epsilon_{n-1} (T u_{n-1} - h) \tag{17a}$$

with

$$\epsilon_{n-1} = - \frac{\| z_{n-1} \|_2^2}{[T z_{n-1}, z_{n-1}]_2} \tag{17b}$$

$$z_{n-1} = T u_{n-1} - h \tag{17c}$$

We note that we improve performance at each step, that is to say:

$$Q(u_{n+1}) = Q(u_n) - \frac{\| z_n \|_2^4}{[T z_n, z_n]_2} \tag{18}$$

It can be shown [3] that $Q(u_n)$ decreases to the optimum:

$$\lim_n Q(u_n) = \inf_u Q(u)$$

Many variations on this basic idea are possible. However, the method has to be substantially modified when the control u is constrained, such as u is restricted to some closed bounded convex set with the origin as an interior point. In this case, we can try adjusting ϵ_n and z_n to satisfy the constraint in many ways but the convergence of the sequence to the optimum would appear to be difficult to obtain in general. By slightly enlarging the convex set to be the smallest sphere containing it, we can obtain a general convergence proof for a modified steepest descent method, as shown in [4]. Let us assume that the radius of the sphere is M. Then we have seen that the optimal solution is given by

$$u = (R + k)^{-1} L^* g$$

for some k such at $\| u \|_2 = M$.

Let us assume that at the nth iteration, we have

$$\| u_n \|_2^2 \leqq M^2$$

We shall now consider, as a function of positive k,

$$u_{n+1}(k) = u_n - \epsilon_n z_n$$

where

$$\epsilon_n = \frac{\| z_n \|_2^2}{[(R+k)z_n, z_n]_2}$$

$$z_n = (R+k)\, u_n - L^* g$$

Now

$$\| u_{n+1}(k) \|_2^2$$

is a continuous function of k and approaches to zero as k goes to infinity. Suppose

$$\| u_{n+1}(0) \|_2^2 \leqq M^2$$

Then we define

$$u_{n+1} = u_{n+1}(0)$$

Otherwise, let k_n be such that

$$\| u_{n+1}(k_n) \|_2^2 = M^2$$

and define

$$u_{n+1} = u_{n+1}(k_n)$$

It can then be shown that

$$\lim_n Q(u_n) = \inf_u Q(u)$$

for $\| u \|_2 \leqq M$. It must be noted here that the sequence is not necessarily monotone at the early stage of the iteration.

In practice, we are not really interested to have $\| u \|_2$ exactly equal to M. A more realistic constraint for u is that for some small $\delta > 0$, either

$$\| u \|_2^2 \leqq M^2$$

or

$$M^2 - \delta \leqq \| u \|_2^2 \leqq M^2$$

Thus we shall define

$$u_{n+1} = u_{n+1}(0)$$

if

$$\| u_{n+1}(0) \|^2 \leqq M^2$$

and

$$u_{n+1} = u_{n+1}(k_n)$$

if

$$M^2 - \delta \leqq \| u_{n+1}(k_n) \|_2^2 \leqq M^2$$

This is actually what we would do when we carry out the computation on a digital computer.

V. THE FINAL-VALUE PROBLEM FOR LINEAR SYSTEMS

In solving the general control problem, the steepest descent method can be used to obtain the optimal solution. However, when we are only interested in controlling the final output vector, then there is a reduction in dimensionality affording a corresponding simplification. In fact, the range of L becomes finite dimensional so that matrix methods suffice.

More specifically, we want to minimize

$$|| x(T) - g ||^2 \; ; \; x(T) = \int_0^T W(T,t)\, u(t)\, dt$$

where now g is a finite dimensional vector. In our general treatment, we have only then to take m_1 to be a jump at T. Hence H_1 is finite dimensional. For simplicity let us take m_2 to be Lebesgue measure. Then again L is compact, and moreover

$$L^* g = W(T,t)^* g$$

so that, using (7) we know that the optimal solution is of the form

$$u = L^* y$$

Also we need to consider

$$\left[L^*L + kI \right] L^*y = L^* \left[(LL^* + kI)\, y \right] = L^*g$$

and now

$$LL^*$$

being an operator on a finite dimensional space must be a square matrix, and indeed it is the n x n matrix

$$A(T) = LL^* = \int_0^T W(T, \tau)\, W(T, \tau)^*\, d\tau$$

and the optimal vector y is given by

$$\left[A(T) + k \right] y = g \tag{19}$$

for some k. The constraint

$$\| u \|_2 \leqq M$$

now becomes

$$y^*\, A(T)\, y \leqq M^2$$

The steepest descent can then proceed using (17a) and (17b), so that for instance

$$y_{n+1} = y_n + \epsilon_n \left[\left(A(T) + k_n \right) y_n - g \right]$$

On the other hand, we can also use the eigenfunction approach, the eigenfunctions of R being finite in number, since range of R is finite dimensional. In particular we note from Section III that for the minimum to be zero for every g, we must have that R must have exactly n eigenfunctions corresponding to non-zero eigenvalues. It should be noticed that the eigenvalues for R is exactly the eigenvalues for A(T). We can, of course, readily translate (19) in terms of the eigenvectors of A(T). Let β_i be the orthonormalized eigenvectors corresponding to the non-zero eigenvalues λ_i of A(T). Then the optimal y is given by

$$y = \sum_{i=1}^{p} \frac{[g, \beta_i]}{\lambda_i + k}\, \beta_i$$

It may also be noted that a necessary and sufficient condition for A(T) have a zero eigenvalue is that

$$W(T,t)^* y \equiv 0 \text{ for } 0 \leqq t \leqq T \tag{20}$$

for some non-zero y, since

$$y^* A(T)\, y = \int_o^T \| W(T,t)^* y \|^2 \, dt = 0$$

For dynamical systems characterized by differential equations, (20) can of course be carried further.

VI. NONLINEAR SYSTEMS

In this section, we shall indicate some extensions of the previous theory to nonlinear systems. We shall not strive for maximum generality here, although much of the theory is capable of further generalization. A key point in dealing with nonlinear systems is the specification of the nonlinear system. We may consider a dynamic system characterized by a nonlinear differential equation

$$\dot{x} = F(x,u,t) \tag{21}$$

where x(t) is the system state vector and u(t) is the control vector. We want to minimize, as before

$$C(u) = \int_\pi \| Ax(t) - g(t) \|^2 \, d\, m_1 + k \int_\pi \| u(t) \|^2 \, d\, m_2 \tag{22}$$

where $\| \quad \|$ denotes the Euclidean norm. Assuming there is a unique solution to the Equation (21) for fixed initial conditions, we have then a mapping from H_2 into H_1 as before, except that now in

$$A\, x = AN(u)$$

N is not a linear operator but rather a nonlinear operator or function. We shall assume that F(. , .) is such that N(u) is analytic--that is, locally bounded and G-differentiable. Then we wish to minimize

$$C(u) = \| AN(u) - g \|_1^2 + k \| u \|_2^2 \tag{23}$$

From this point of view then, all we need is N(u), and it does not matter whether it comes from a dynamic system or not. It is clear that (23) is also analytic. Moreover, the Frechet derivative is (in the notation of Reference 6)

$$\delta\Big[\, C(u);\, h \,\Big] = 2\Big[\, A\delta[\, N(u);h \,] \,,\, AN(u) \,\Big]_1 - 2\Big[\, A\delta\,[\, N(u);h \,] \,,\, g \,\Big]_1 + 2k\Big[\, u, h \,\Big]_2 \tag{24}$$

But, since $\delta[C(u);\, h]$ is a continuous linear homogeneous polynomial, we must have

$$\delta\Big[\, N(u);\, h \,\Big] = L\,h$$

$$\delta\Big[\, C(u);\, h \,\Big] = \Big[\, v,\, h \Big]_2$$

where, now, from (24), we note that

$$v = 2\Big[\, AL \,\Big]^* A\, N(u) - 2\,\Big[\, AL \,\Big]^* g + 2\,k\,u \tag{25}$$

and moreover, the derivative zero leading to an extremum for C such that

$$0 = v = \Big[\, AL \,\Big]^* AN(u) - \Big[\, AL \Big]^* g + k\,u$$

or

$$k\,u = \Big[\, AL \,\Big]^* \Big[\, g - AN(u) \,\Big] \tag{26}$$

This result generalizes (7). Substituting

$$u = (AL)^*\, y$$

in (26), we again obtain for y the equation in H_1

$$AN\Big[\, (AL)^*\, y \,\Big] + k\,y = g \tag{26a}$$

We note that (26) and (26a) are no longer linear equations, although if H_1 is finite dimensional, (26a) is of course also a finite dimensional vector equation. On the other hand, steepest descent method in this context becomes:

$$u_{n+1} = u_n + \epsilon_n z_n \tag{27}$$

where

$$z_n = \left[AL \right]^* AN(u_n) - \left[AL \right]^* g + k\, u_n$$

and ϵ_n is chosen to minimize

$$C(u_{n+1})$$

Although we can thus inprove the approximation at every stage, convergence to the optimum is, in general, much harder to prove.

The main calculation is the derivative $\delta[N(u); h]$. We shall now examine this in some special cases. First of all, being analytic, we know that we can expand N(u) as a power series which can then be approximated by a polynomial of degree m say. For example, suppose that u(t) is one dimensional, and we take m = 2:

$$x(t) = \int_0^T W_1(t, s)\, u(s)\, ds + \int_0^T \int_0^T W_2(t, s_1, s_2)\, u(s_1) u(s_2)\, ds_1\, ds_2 \qquad (28)$$

then

$$\delta \left[N(u); h \right] = L\, h$$
$$= \int_0^T W_1(t, s)\, h(s)\, ds + \int_0^T \left[\int_0^T W_2(t, s, \sigma)\, u(\sigma)\, d\sigma \right] h(s)\, ds$$

So that, taking A as the identity matrix, and m_1, and m_2 as Lebesgue measure, we have

$$v = \int_0^T \left[W_1(t, s) + \int_0^T W_2(t, s, \sigma)\, u(\sigma) d\sigma \right] \left[\int_0^T W_1(t, \sigma) u(\sigma)\, d\sigma \right.$$
$$\left. + \int_0^T \int_0^T W_2(t, s_1, s_2) u(s_1)\, u(s_2)\, d\, s_1 ds_2 \right] dt \qquad (29)$$
$$- \int_0^T \left[W_1(t, s) + W_2(t, s, \sigma)\, u(\sigma)\, d\sigma \right] g(t)\, dt$$
$$+ k\, u(s)$$

Setting this equal to zero, we have a nonlinear integral equation to solve for obtaining a local extremum. On the other hand, since the explicit functional representation x(t) in terms of u(t) is known, the steepest descent procedure can be applied directly.

When u(t) is of dimensionality higher than one, the representation of N(u) by an approximating polynomial is somewhat more complex. A polynomial of degree m being written as the sum of homogeneous polynomial of lower degree will be:

$$N(u) = \sum_{k=o}^{m} \int_o^T \cdots \int_o^T L_K \left(t, u(s_1), u(s_2) \cdots u(s_k) \right) ds_1 \cdots ds_k \qquad (30)$$

where for each t

$$L_k(t, x_1, x_2, \ldots, x_k)$$

is a vector, each component of which is a continuous K-linear form in the vector $x_1, x_2, \ldots x_k$. The Frechet derivative can again be calculated in a similar manner.

If the system is described in terms of the nonlinear differential Equation (21), it is possible to obtain $\delta[N(u); h]$ as a solution to a linear time varying equation. For, in

$$x = F(x, u + \lambda h) \qquad (31)$$

for fixed u and h, we have a function of λ for each t, and at $\lambda = 0$

$$\frac{dx}{d\lambda} = \nabla_x F[x, u] \frac{dx}{d\lambda} + \nabla_u F[x, u] h \qquad (32)$$

where $\nabla_x F(x, u)$ and $\nabla_u F(x, u)$ are the gradient matrices. Now (32) is a linear equation for $\frac{dx}{d\lambda}$, and taking initial condition for (21) to be zero, so that

$$\frac{d\,x(t)}{d\lambda} = 0 \text{ for } t = 0, \lambda = 0$$

we have

$$\frac{dx}{d\lambda} = \int_o^t \Phi(t)\, \Phi^{-1}(s)\, \nabla_u F \left(x(s), u(s) \right) h(s)\, ds \qquad (33)$$

where $\Phi(t)$ is the fundamental matrix solution of

$$\dot{Y} = \nabla_x F[x, u]\, Y$$

From (33), it follows that

$$L\, h = \delta[N(u) ; h]$$

corresponds to the operation

$$\int_o^t \Phi(t)\Phi\ (s)^{-1}\ \nabla_u F\left[\ x(s),\ u(s)\ \right]\ h\ (s)\ ds$$

To use this, x(s) must be known, so that if one solution is known, we can perturb it to obtain a better approximation to the minimization problem using Equation (27). Some simplification in these considerations occurs, if, as is usually the case,

$$F(x, u) = f_1(x) + f_2\ (u)$$

so that

$$\nabla_x F(x, u) = \nabla_x f_1(x)$$
$$\nabla_u F(x, u) = \nabla_u f_2(u)$$

VII. COMPUTER STUDIES

In this section we shall present the results of some computer studies of the steepest descent methods espoused in the paper. These applications are not meant to be exhaustive or even representative but rather to illustrate some of the specific considerations that arise in such application.

Example 1. Let us now consider a second order linear system which is described by the differential equation

$$\frac{d^2x}{dt^2} + \frac{dx}{dt} = u(t)$$

Assume that the desired output over the control interval [0, 1] is a step function, i.e.,

$$x_d(t) = 1 \qquad 0 \leqq t \leqq 1$$

Let the initial conditions be

$$x(0) = 1 \qquad x'(0) = -1$$

Then

$$x_o(t) = e^{-t}$$

and

$$g(t) = 1 - e^{-t}$$

The problem then is to have the position output of the system following the step function subject to the energy constraint that

$$\| u \|_2^2 \leqq 3$$

This is a one-dimensional problem.

The weighting function of the system is

$$W(t) = 1 - e^{-t}$$

Then

$$G(\tau_1, \tau_2) = - \text{Max}\,(\tau_1, \tau_2) + e^{-(1-\tau_1)} + e^{-(1-\tau_2)} - \frac{1}{2}\left(e^{-|\tau_1 - \tau_2|} + e^{-(2-\tau_1-\tau_2)} \right)$$

and

$$L^* g = 0.3678 - \tau - 0.5\, e^{-\tau} + 0.3001\, e^{\tau}$$

Here Max (τ_1, τ_2) denotes the maximum of τ_1 and τ_2.

In carrying out the computation, we take $\delta = 0.2$. The initial guess for the solution is arbitrarily chosen to be $u_o(t) = 1$. Some of the computational results are listed in Table I. At each iteration step, we have to adjust k_n such that

$$2.8 \leqq \| u_n \|_2^2 \leqq 3$$

The increment of adjustment for k_n in this problem is chosen to be 0.005. It is interesting to see that, after the first iteration, the number k_1 is already very close to its optimal value. At the third iteration, k_n has reached its optimal value.

The closeness of the approximate solution for u(t) to its optimal one can be measured by the ratio

$$ER = \frac{\| (R + k_n)\, u_n - L^* g \|_2}{\| L^* g \|_2}$$

Thus at the 4th iteration with ER = 0.0044, we can stop the process. The system performance is plotted in Figure 1 and the sequence of control inputs is

TABLE I

COMPUTATIONAL RESULTS FOR EXAMPLE 1

No. of Iteration Steps n	k_n	$\|u_n\|_2^2$	$\frac{\|(R+k_n)u_n - L^* g\|_2}{\|L^* g\|_2}$	$Q(u_n)$	$Q(u_n)+k_n\|u_n\|_2^2$	$\frac{\|L u_n - g\|_1^2}{\|g\|_1^2}$
1	0.0065	2.949360	0.515050	-0.155592	-0.136421	0.074357
2	0.0060	2.859088	0.069710	-0.159816	-0.142657	0.049229
3	0.0055	2.996001	0.024473	-0.160658	-0.144180	0.044222
4	0.0055	2.969426	0.004438	-0.160531	-0.144200	0.044976
5	0.0055	2.999596	0.005363	-0.160701	-0.144203	0.043963
6	0.0055	2.992307	0.000992	-0.160662	-0.144205	0.044196
7	0.0055	2.999195	0.001222	-0.160700	-0.144205	0.043969
8	0.0055	2.997514	0.000226	-0.160691	-0.144205	0.044024
9	0.0055	2.999088	0.000279	-0.160700	-0.144205	0.043973
10	0.0055	2.998706	0.000052	-0.160698	-0.144205	0.043985

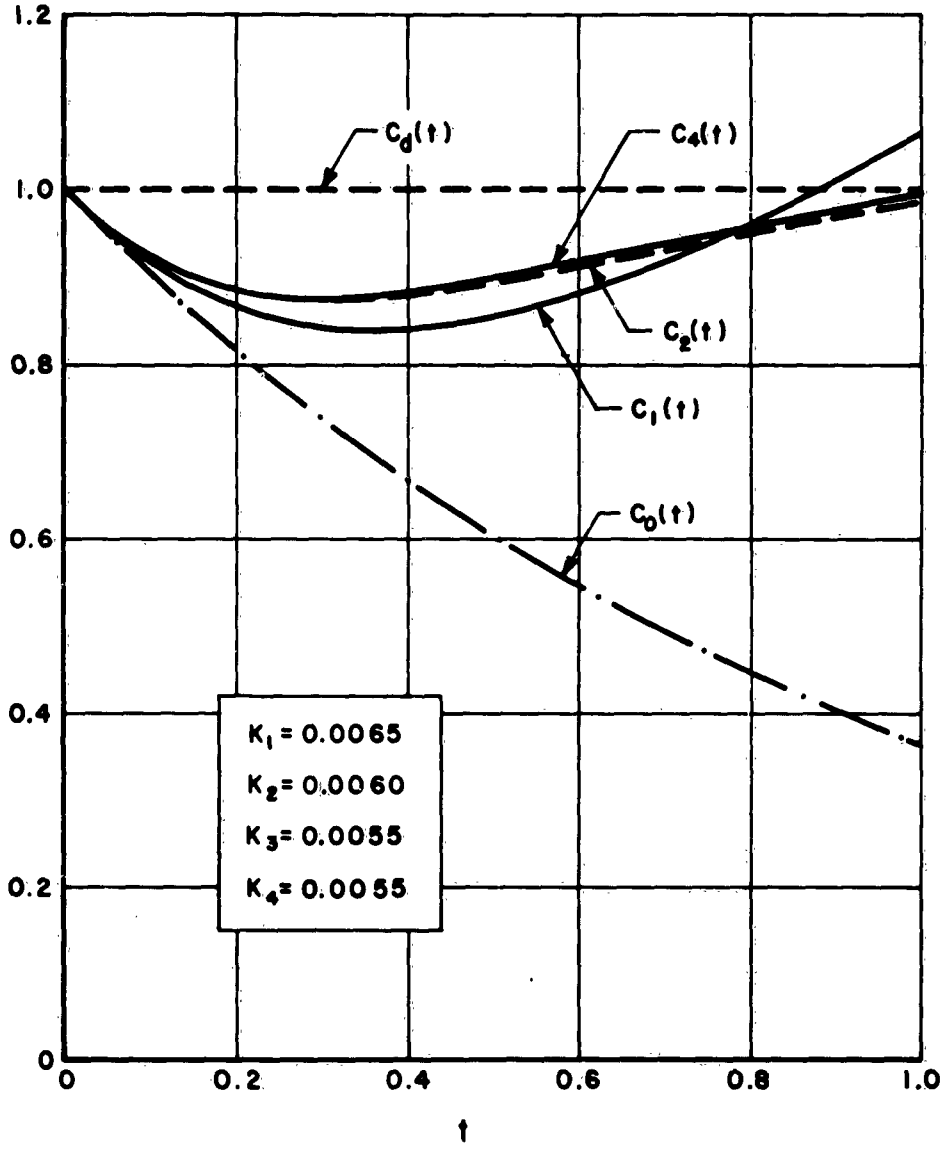


FIGURE 1. PERFORMANCE OF THE SYSTEM IN EXAMPLE 1

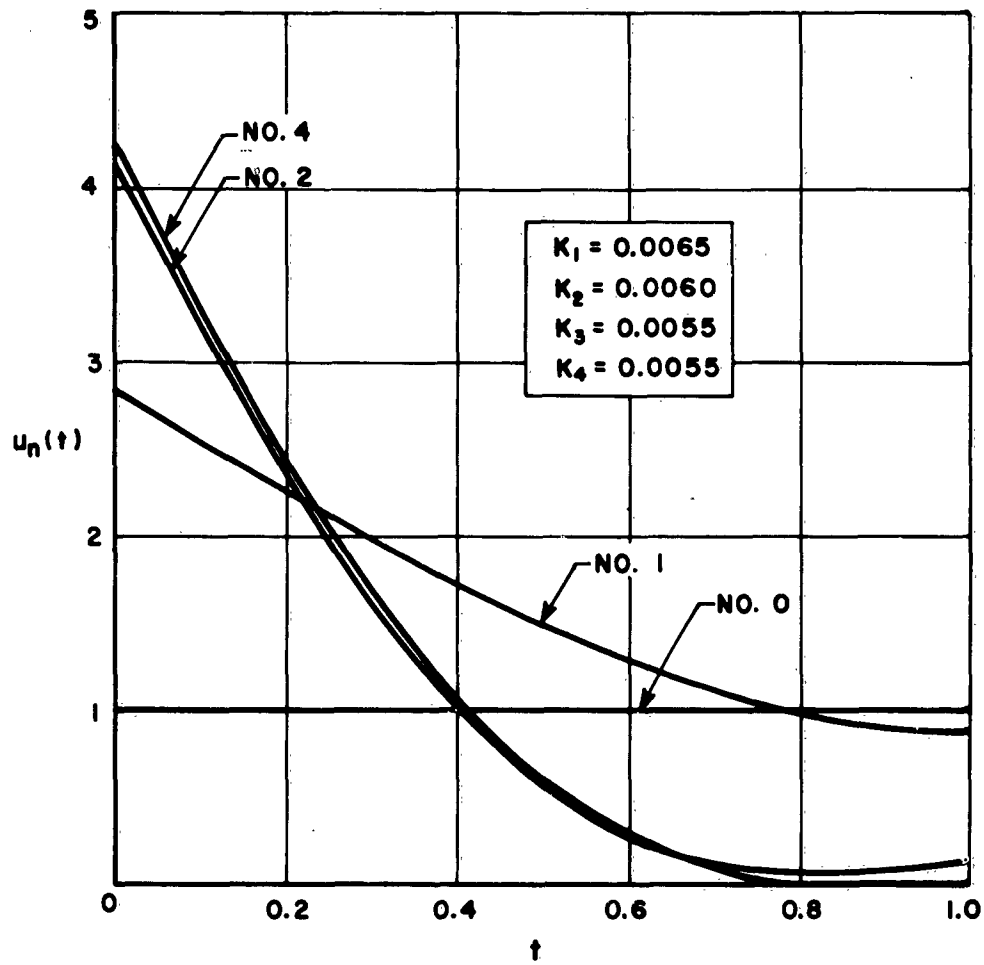


FIGURE 2. THE APPROXIMATION SEQUENCE $u_n(t)$ FOR THE SYSTEM IN EXAMPLE 1

plotted in Figure 2. It should be noticed that the oscillation of $Q(u_n)$ after k_n has reached its steady value of 0.0055 is due to the variation of $\| u(k_n) \|_2^2$. However, $Q(u_n) + k_n \| u_n \|_2^2$ is indeed a non-increasing sequence for the same values of k_n.

Example 2. Let us consider the same system as given in Example 1. However, we shall now try to control both the position and velocity outputs of the system to follow the ideal step function by using a single control input. In this case, we have a 2 x 1 system. Thus

$$x_d = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$W(t) = \begin{bmatrix} 1 - e^{-t} \\ e^{-t} \end{bmatrix}$$

$$x_o(t) = \begin{bmatrix} e^{-t} \\ -e^{-t} \end{bmatrix}$$

$$g(t) = x_d(t) - x_o(t)$$

$$= \begin{bmatrix} 1 - e^{-t} \\ e^{-t} \end{bmatrix}$$

Then

$$G(\tau_1, \tau_2) = \text{Max}\,(\tau_1, \tau_2) + e^{-(1-\tau_1)} + e^{-(1-\tau_2)} - e^{-(2-\tau_1-\tau_2)}$$

$$L^* g = 0.3678 - \tau - 0.2324\, e^{\tau}$$

The constraint for the input is taken to be

$$\| u \|_2^2 \leqq 5$$

In carrying out the computation, we again choose $\delta = 0.2$. Hence when $\| u \|_2^2$ exceed 5, we shall adjust k_n such that

TABLE II

COMPUTATIONAL RESULTS FOR EXAMPLE 2

No. of Iteration Steps n	k_n	$\|u_n\|_2^2$	$\frac{\|(R+k_n)u_n - L^*g\|_2}{\|L^*g\|_2}$	$Q(u_n)$	$Q(u_n)+k_n\|u_n\|_2^2$	$\frac{\|Lu_n - g\|_1^2}{\|g\|_1^2}$
1	0	2.515420	0.396180	-0.437727	-0.437727	0.396180
2	0	1.970645	0.241534	-0.487548	-0.487548	0.241534
3	0	3.240699	0.196627	-0.513064	-0.513064	0.145496
4	0	3.295242	0.126567	-0.527287	-0.527287	0.121808
5	0	4.184334	0.112038	-0.535891	-0.535891	0.107479
6	0	4.307971	0.079227	-0.541763	-0.541763	0.097699
7	0	4.975983	0.078128	-0.546090	-0.546090	0.090491
8	0.001	4.978586	0.059593	-0.548912	-0.544033	0.085792
9	0.008	4.866785	0.030549	-0.549684	-0.510750	0.084506
10	0.005	4.961765	0.029163	-0.551068	-0.526259	0.082201
11	0.009	4.953561	0.018552	-0.551326	-0.516744	0.081772
12	0.008	4.908812	0.014031	-0.551118	-0.510848	0.082118
13	0.008	4.984433	0.012147	-0.551844	-0.511969	0.080909
14	0.008	4.955171	0.011334	-0.551709	-0.512068	0.081134
15	0.009	4.927553	0.008125	-0.551580	-0.507232	0.081348
16	0.008	4.942081	0.009304	-0.551775	-0.512238	0.081024
17	0.009	4.947917	0.005447	-0.551861	-0.507330	0.080880
18	0.009	4.919139	0.006306	-0.551625	-0.507353	0.081273
19	0.009	4.933637	0.003887	-0.551774	-0.507371	0.081026
20	0.009	4.914428	0.004794	-0.551614	-0.507384	0.081291
21	0.009	4.930254	0.003020	-0.551767	-0.507395	0.081036
22	0.009	4.916224	0.003880	-0.551650	-0.507404	0.081232
23	0.009	4.930914	0.002415	-0.551789	-0.507411	0.081000
24	0.009	4.920170	0.003067	-0.551698	-0.507416	0.081152
25	0.009	4.933092	0.001959	-0.551819	-0.507422	0.080951

$$4.8 \leqq \|u\|_2^2 \leqq 5$$

The increment of adjustment for this problem is taken to be 0.001. The initial guess is still chosen to be $u_o(t) = 1$. The computational results are listed in Table II. When k_n are slightly different, $Q(u_n)$ sometimes show oscillations as in Steps No. 11 and No. 12, or in Steps No. 14 and No. 15. This is due to the fact that we are not adjusting $\|u\|_2^2$ to be exactly equal to 5 and it is not practical and necessary to do so. It is seen that larger norm always corresponds to smaller Q. Thus the finer the adjustment for k_n and the smaller the value for δ, the better will be the monotonocity for Q.

Figure 3 shows the system performance and Figure 4 shows the sequence of controls. It is seen that the iteration process can be stopped at the 15th step with ER = 0.008. In the practical application of this method when on-line computation is required, a proper threshold level for ER will be chosen. The choice of this threshold should be based on the desired accuracy of the solution and the computing time which we can afford.

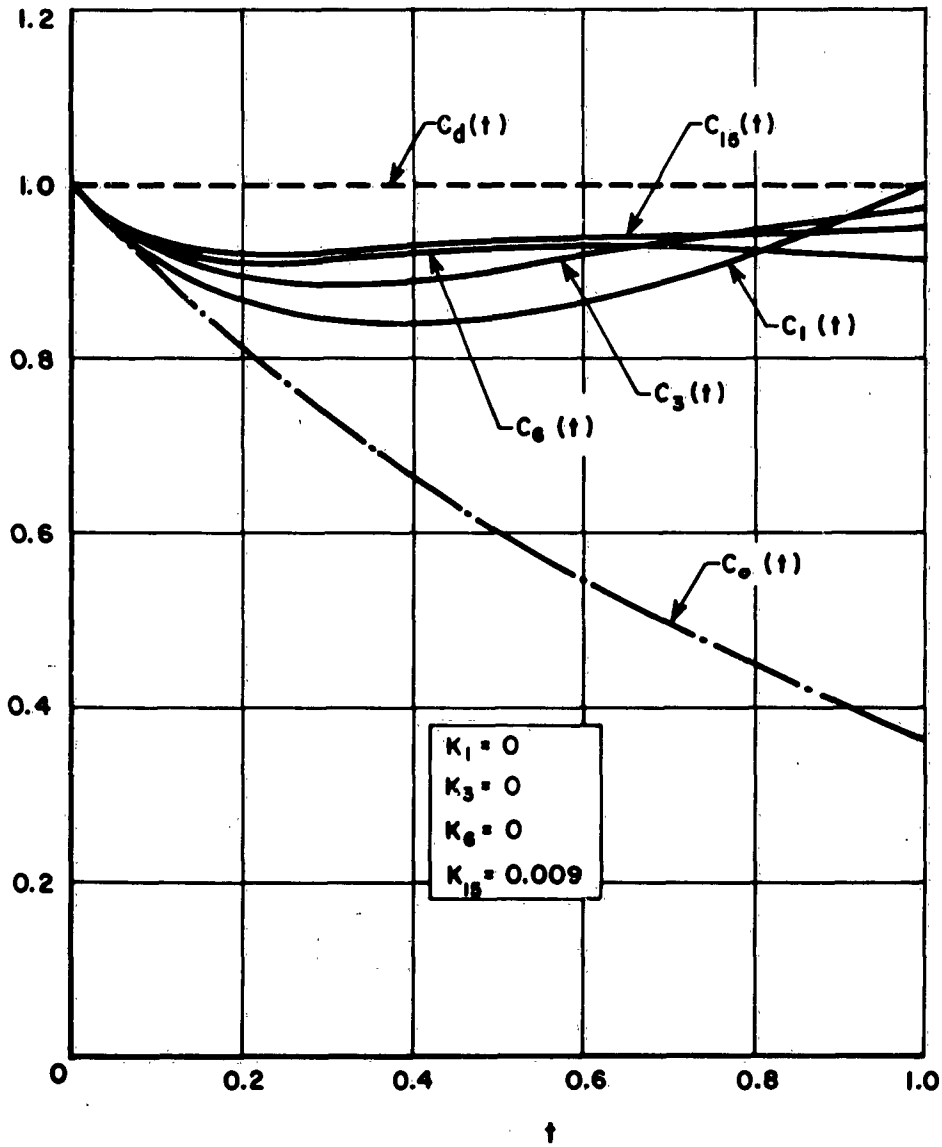


FIGURE 3.— PERFORMANCE OF THE SYSTEM IN EXAMPLE 2

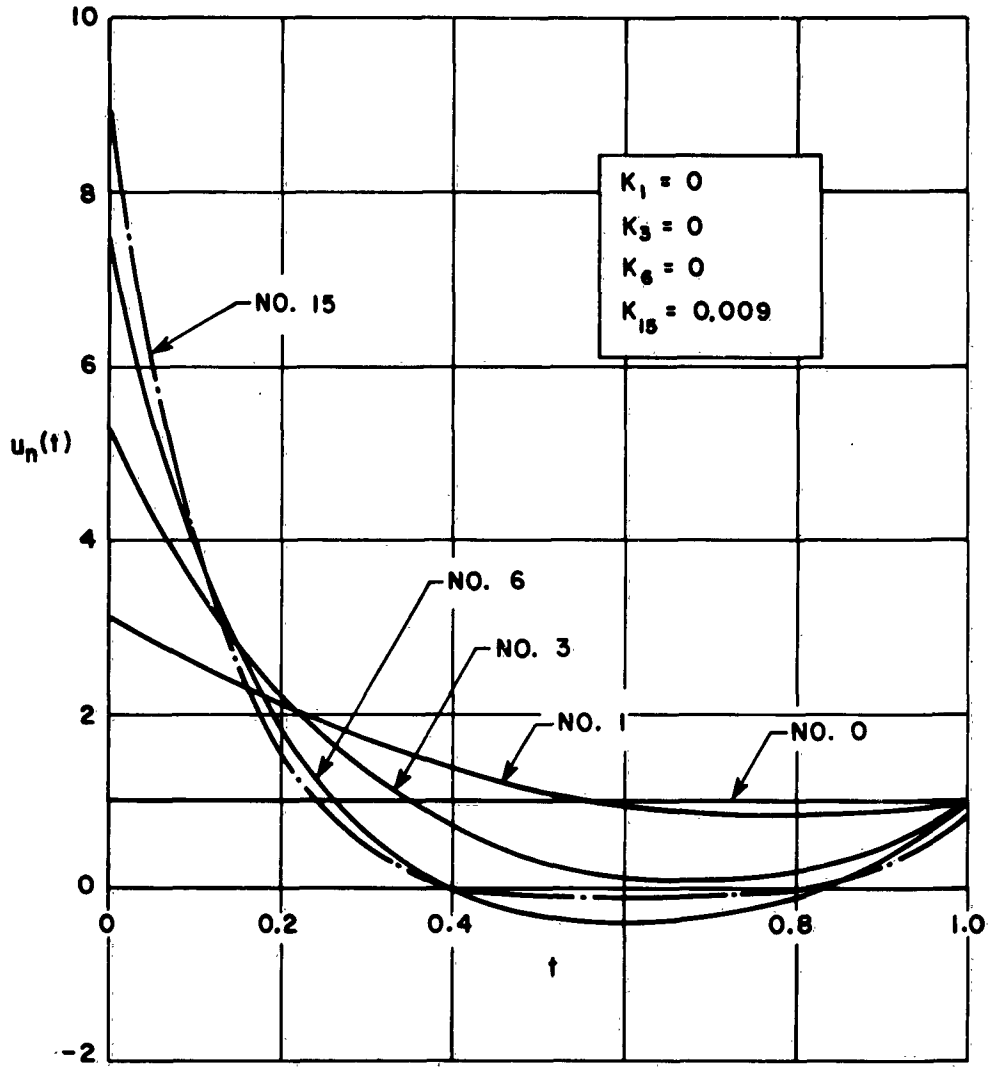


FIGURE 4. THE APPROXIMATION SEQUENCE $u_n(t)$ FOR THE SYSTEM IN EXAMPLE 2

REFERENCES

1. Neustadt, L. W., "Synthesizing Time Optimal Control Systems," J. of Math. Anal. and Applic., Vol. 1, No. 4 (1960), pp. 484-493.

2. Ho, Y. C., "A Successive Approximation Technique for Optimal Control Systems Subject to Input Saturation," Trans. ASME, Series D, J. of Basic Engineering, Vol. 84, No. 1 (1962), pp. 33-40.

3. Balakrishnan, A. V., "On a General Theory of Nonlinear Estimation Problems in Control Systems," invited lecture, Symposium on Mathematical Problems in Control Systems, Washington, D. C., November, 1961.

4. Balakrishnan, A. V., "An Operator-Theoretic Formulation of a Class of Control Problems and a Steepest Descent Method of Solution," to be published in J. SIAM.

5. Kantorovich, L. V., "Functional Analysis and Applied Mathematics," Usp. Mat. Nauk, Vol. 3, pp. 89-185.

6. Hille, E. and R. S. Philips, "Functional Analysis and Semigroups," American Mathematical Society Colloquium Publications, Vol. 31, 1957.

ASD-TDR-63-119

The Problem of Optimal Mode Switching

By

R. A. Nesbit

University of California, Los Angeles
Aerospace Corporation,
El Segundo, California

THE PROBLEM OF OPTIMAL MODE SWITCHING

R. A. Nesbit

University of California, Los Angeles
Aerospace Corporation, El Segundo, California

ABSTRACT

Control systems may be designed to operate in different modes. In some cases these modes may be switched to obtain various system trajectories. The optimization of the trajectories can be considered as a determination of the switching times and other mode parameters. The optimization problem is to determine these times and parameters. First order computations can yield a descent method but convergence is not assured. Theorems which deal with the existence of solutions to optimal control problems which have been recently proved indicate that it is better to relax the dynamic restrictions during the optimization than to work only with the admissible class of switched controls. The resulting optimal trajectory is then approximated by one of the trajectories from the admissible class.

INTRODUCTION

Some vehicles have control systems which operate in various modes, such as an angle of attack hold mode, a bank angle hold, a flight path angle hold, and a pressure altitude hold. For the control of a re-entry vehicle one might include other hold modes such as acceleration, attitude, thrust, or temperature. One approach to trajectory optimization is to select the appropriate combination of modes at each time. For some optimization criteria, however, the optimum switching will "chatter" or change rapidly from one position to another. This type of behavior has already been noted in connection with relay controllers. A necessary condition for the existence of chattering in optimal controllers can be based on the application of the maximum principle. This condition turns out to give information about the optimal trajectory in trivial examples, although it is questionable whether or not this result is of practical significance. Warga, Roxin, Filippov, and Chang (References 1 through 4), deal with the problem of existence of solution for the general control problem. This general theory seems to indicate that optimization of the mode switching problem should be accomplished by generalizing the problem. The method consists of first relaxing the restrictions on the switching, finding the optimum, and then approximating this optimum by the restricted switching trajectories.

A sub-optimal trajectory can be constructed using a fixed switching sequence and adjusting the switching times by a steepest descent procedure.

DISCUSSION

A class of mode switching systems will be considered. These systems will be represented by the vector equations

$$\frac{d\underline{x}}{dt} = \underline{f}(\underline{x}, \underline{u})$$

$$t_0 \leq t \leq t_2 \qquad (1)$$

$$\underline{x}(t_0) = \underline{x}_0 \quad \underline{x}_0 \in B_0$$

$$\underline{x}(t_1) \in B_1$$

For each admissible $\underline{u}(t)$ the solution of the differential equation is assumed to be well defined. The initial condition on the n-dimensional state vector is in some region B_0 and the terminal condition is required to be in some region B_1. The r-dimensional control vector $\underline{u}$ is chosen from a set of functions

$$\underline{u}_k(\underline{x} - \underline{r}_k) \quad k = 1, 2 \ldots m$$

$$\underline{r}_k \text{ constant vectors} \qquad (2)$$

The value of $\underline{u}$ is chosen from the set of possibilities. This method of formulation is made in an attempt to incorporate the design restrictions into the optimization problem. Instead of formulating a completely "open loop" set of dynamics, a set of closed loop systems is considered. The formulation is motivated by the current practice of designing control systems. For example, the attitude control system of some satellites have different sensor and feedback configurations for acquisition, earth pointing, and eclipse operation. The application of these systems requires some sort of switching logic, but in many cases the logic is very simple. Since the system designer is generally willing to sacrifice some performance for a simple mechanization, it is of interest to explore the possibilities of mode switching logic.

A particular application which motivates this study is the trajectory control of a vehicle, and while this problem has been explored in some detail to find optimum trajectories and means of following them, it seems reasonable to consider "sub-optimal" trajectories if they are easier to implement.

One requirement of the trajectory problem is to transfer the system from an initial state in some region B_0 at t_0 to a terminal state in another region B_1 at some reasonable time $t_1 \leq t_2$. The fundamental question about whether or not the vehicle is capable of making the trip is of interest to those concerned with the journey. To satisfy doubts on this subject, it is worthwhile to find at least one control sequence which makes the trip possible.

If the journey is so difficult that only one control scheme will work, then the question is how to find and mechanize it. But if there are many possibilities, then some choice of route must be made even if it is made by chance. In some "games" random selection is the "optimal strategy"; in

other cases there is some "cost" which is to be minimized. Again, existence and uniqueness are of interest. When there is a control which completes the journey at some cost, then the existence of a minimizing trajectory is of less importance since the cost can be made reasonably close to the minimum.

The lack of uniqueness of the optimizing trajectories merely requires an additional decision. Some writers indicate that the problem "is not properly formulated" as long as any ambiguity remains, but even without admitting this one can conceive of a hierarchy of "play-offs" leading to the final choice of controls. For this discussion the first component of the state vector $x^1(t_1)$ at the terminal time t_1 will be taken as the cost function to be minimized.

THE SWITCHING CONCEPT

The control problem is to select one of the possible switch positions. A block diagram of the system described above is shown in Figure 1.

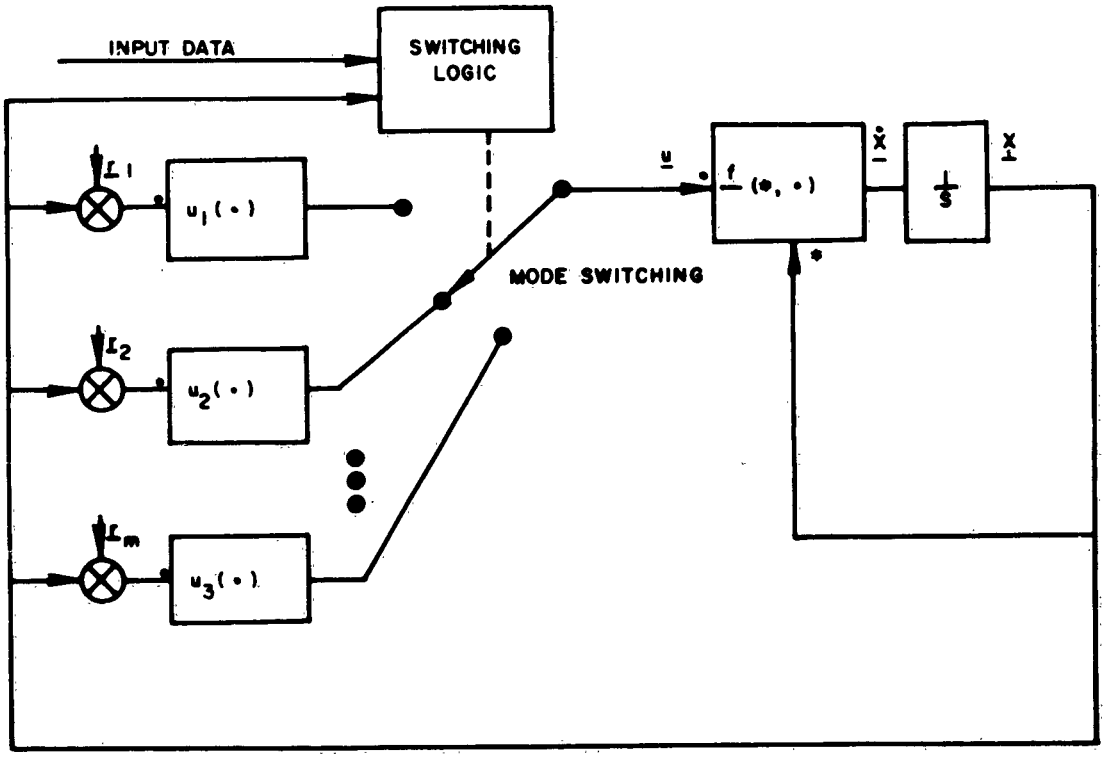


Figure 1. Block Diagram of Mode Switching System.

The switching logic is required to give a trajectory which satisfies terminal conditions. For linear systems this requirement may be implemented by a final value controller (Reference 5). The significant computation carried out in the final value control problem is the conditional response prediction. This computation utilizes the known dynamics of the system to extend the trajectory to some future time under the assumption of a particular control sequence.

This idea is applicable to nonlinear as well as linear systems. Let $\underline{\phi}_k(t, \underline{x}_0, t_0)$ designate the solution of the equations at time t with the initial state $\underline{x}_0$ at time $\underline{t}_0$ and using the k^{th} mode of control. This solution is a conditional response predictor if the present state is used as initial condition and if the terminal time is used. That is,

$$\underline{x}(t_1) = \underline{\phi}_k\left[t_1, \underline{x}(t), t\right]$$

if

$$\underline{u} = \underline{u}_k(\underline{x}) \quad \text{for} \quad t \le \tau \le t_1$$

The system may start in mode j and switch to mode k at time t_1 in this case

$$\underline{x}(t_2) = \underline{\phi}_k\left[\underline{t}_2, \underline{\phi}_j(t_1, \underline{x}_0, t_0), t_1\right]$$

The effect of small changes in the switching time t_1 are of interest and can be estimated on the basis of first order computations.

Computation of the First Order Influence of Switching Time and Initial State

The computation of the first order effects of changes in t_1, $\underline{x}_0$, and t_0 on $\phi(t_1 \underline{x}_0, t_0)$ can be made for the general mode and the mode subscript will be dropped. The first order effects describe deviations from some "nominal" trajectory. The deviations from the nominal will be designated by the vector z, and the nominal by the vector $\underline{y}(t)$ with $\underline{y}(t_0) = \underline{x}_0$. Then $\underline{z} = \underline{x} - \underline{y}$ and the dynamics of the deviation are

$$\dot{\underline{z}} = \underline{f}(\underline{y} + \underline{z}) - \underline{f}(\underline{y}) \tag{3}$$

Assuming that the first partial derivatives of $\underline{f}$ are continuous, define the matrix A with elements

$$a^i_j(t) = \left.\frac{\partial f^i}{\partial x^j}\right|_{\underline{y}} \tag{4}$$

As usual, the fundamental solution of the adjoint equation will be used. Define the matrix $W(\tau)$ such that

$$\frac{dW}{d\tau} = - WA \qquad W(t_1) = I \tag{5}$$

By combining Equations (3) and (5) one may obtain

$$\frac{d}{dt}(W\underline{z}) = W\left[\underline{f}(\underline{y} + \underline{z}) - \underline{f}(\underline{y}) - A\underline{z}\right]$$

and thus

$$\frac{d}{dt}(W\underline{z}) = 0\left\|\underline{z}\right\|^2$$

where $0\left\|\underline{z}\right\|^2$ indicates the terms of second and higher order. Integrating, one obtains

$$W(t_1)\,\underline{z}(t_1) = W(t_0)\,\underline{z}_0 + \int_{t_0}^{t_1} 0\left\|z\right\|^2 d\tau$$

or

$$\underline{z}(t_1, \underline{z}_0, t_0) = W^{-1}(t_1)\,W(t_0)\,z_0 + 0_2\left\|z\right\|^2 \tag{6}$$

Thus, to a first order of approximation in $\left\|z\right\|$, one obtains

$$\underline{z}(t_1) = W^{-1}(t_1)\,W(t_0)\,\underline{z}(t_0) \tag{7}$$

This will be designated by

$$\underline{z}(t_1) = K(t_1, t_0)\,\underline{z}(t_0) \tag{8}$$

A first order approximation to the effect of switching time on the final state can be obtained by considering the equivalent perturbation $\Delta\underline{z}$ due to a change in switching time and propagating the perturbation to the final time. For switching from the j^{th} mode to the k^{th} mode this perturbation is

$$\Delta\underline{z}(t_i) = \left[\underline{f}(\underline{x}, \underline{u}_j) - f(\underline{x}, \underline{u}_k)\right]\Delta t_i \tag{9}$$

and the effect on the final state is

$$\begin{aligned}\Delta\underline{z}(t_n) &= K(t_n, t_i)\,\Delta\underline{z}(t_i) \\ &= K(t_n, t_i)\left[\underline{f}(\underline{x}, \underline{u}_j) - \underline{f}(\underline{x}, \underline{u}_k)\right]\Delta t_i\end{aligned} \tag{10}$$

Defining

$$\underline{v}_i = K(t_n, t_i)\left[\underline{f}(\underline{x}, \underline{u}_j) - \underline{f}(\underline{x}, \underline{u}_k)\right] \tag{11}$$

one may write the final perturbation vector as

$$\Delta\underline{z}(t_n) = K(t_n, t_0)\,\Delta\underline{z}(t_0) + \sum_{i=1}^{n=1} \underline{v}_i\,\Delta t_i \tag{12}$$

A linearized mode switching controller could be designed by setting $\underline{z}(t_n) = 0$ and solving for Δt_i using a psuedo inverse if necessary. The effect of small changes in the switching times Δt_i can be used to "improve" the nominal by the usual steepest descent method. The switching influence vectors $\underline{v}_i$ may span a subspace of the state space $\underline{z}$ even if the number of switching times is large.

At each step of a descent procedure the values of Δt_i must be chosen. This can be done by making Δt_i the "solution" to the equation

$$\underline{v}_1\,\Delta t_1 + \underline{v}_2\,\Delta t_2 + \ldots + \underline{v}_s\,\Delta t_s = \Delta\underline{z}(t_n)$$

using a psuedo inverse if necessary.

This procedure will not converge in many cases, and the real problem is to determine the conditions under which it will converge. A fairly trivial example will now be discussed to illustrate the descent procedure.

EXAMPLE

The computations of the descent procedure will be illustrated in the usual first order system with mean square error. The system is shown in Figure 2 and is described by

$$\dot{x}^1 = (x^2 - 1)^2$$

$$\dot{x}^2 = \begin{cases} (1-a)(x^2+2) \\ -a\,2x^2 \end{cases} \qquad a = 0 \text{ or } 1$$

$$x^1(0) = 0 \qquad \begin{matrix} \dot{x}^2(0) = -0.5 \\ x^2(2) = 1.5 \end{matrix}$$

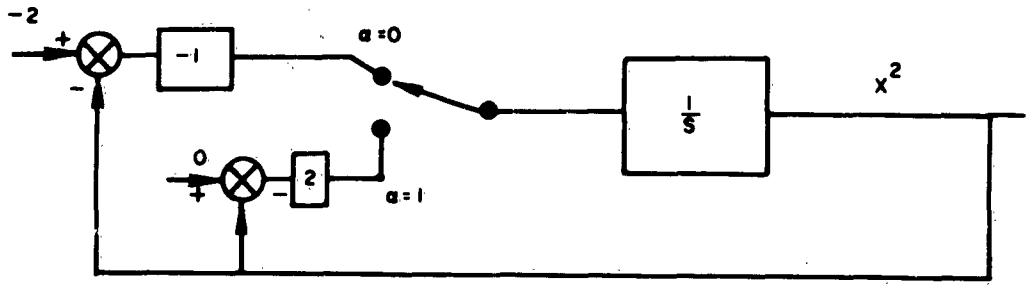


Figure 2. System for Example.

As will be shown later, this system will be required to chatter in order to approach the non-attainable minimum of $x^1(2)$. During this chattering the process will "slide" along the trajectory $x^2 = 1$. For the present, a sub-optimal trajectory with only two switches will be considered. The switch parameter a will be taken as

$$0 \text{ for } 0 < t \leq t_1$$

$$1 \text{ for } t_1 < t \leq t_2$$

$$0 \text{ for } t_2 < t \leq 2$$

and the nominal trajectory will switch at $t_1 = 0.5$ $t_2 = 1.5$.

The adjoint equations are

$$\begin{bmatrix} \dot{w}_{11} & \dot{w}_{12} \\ \dot{w}_{21} & \dot{w}_{22} \end{bmatrix} = - \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} 0 & 2(x^2-1) \\ 0 & -2a + (1-a) \end{bmatrix}$$

and since x^2 does not depend upon x^1 these equations may be solved along the nominal

$$W_3 = \begin{bmatrix} 1 & [x^2(t_2) + 2][\exp 2(2 - t_2) - 1] \\ & \quad + 6[1 - \exp(2 - t_2)] \\ 0 & \exp(2 - t_2) \end{bmatrix}$$

$$W_2 = \begin{bmatrix} 1 & -\frac{x^2}{2}(t_1)\left[\exp -4(t_2 - t_1) - 1\right] \\ & \qquad -\left[1 - \exp -2(t_2 - t_1)\right] \\ 0 & \qquad\qquad \exp -2(t_2 - t_1) \end{bmatrix}$$

The velocity difference $(\underline{f}_0 - \underline{f}_1)$ is given by

$$\underline{f}_0 - \underline{f}_1 = \begin{pmatrix} 0 \\ 3x^2 + 2 \end{pmatrix}$$

Along the nominal mentioned above

$$\underline{v}_1 = \begin{pmatrix} -1.62 \\ 0.762 \end{pmatrix} \quad \underline{v}_2 = \begin{pmatrix} -0.766 \\ -3.61 \end{pmatrix}$$

Thus the switching time increments are related to the changes in the terminal state through the equation

$$\begin{bmatrix} -1.62 & 0.766 \\ 0.762 & -3.61 \end{bmatrix} \begin{bmatrix} \Delta t_1 \\ \Delta t_2 \end{bmatrix} = \begin{bmatrix} \Delta x^1(2) \\ \Delta x^2(2) \end{bmatrix}$$

to a first order approximation in Δt_k and $\Delta x^i(t)$. The nominal trajectory does not quite satisfy the terminal conditions, and the increment in terminal state is taken as

$$\begin{bmatrix} \Delta x^1 \\ \Delta x^2 \end{bmatrix} = \begin{bmatrix} -0.1 \\ 0.07 \end{bmatrix}$$

This choice gives switching increments of $\Delta t_1 = 0.0556$ $\Delta t_2 = -0.0133$.

The modification of the switching times does improve the trajectory. Repeated application of the procedure gives the sequence of trajectories tabulated in Figure 3.

	t_0	$x^2(t_0)$	t_1	$x^2(t_1)$	t_2	$x^2(t_2)$	t_3	$x^2(t_3)$
NOM.	0	−0.5	0.5	0.437	1.500	0.0640	2	1.403
TRIAL 1	0	−0.5	0.537	0.566	1.481	0.0857	2	1.504
TRIAL 2	0	−0.5	0.585	0.692	1.494	0.1123	2	1.503

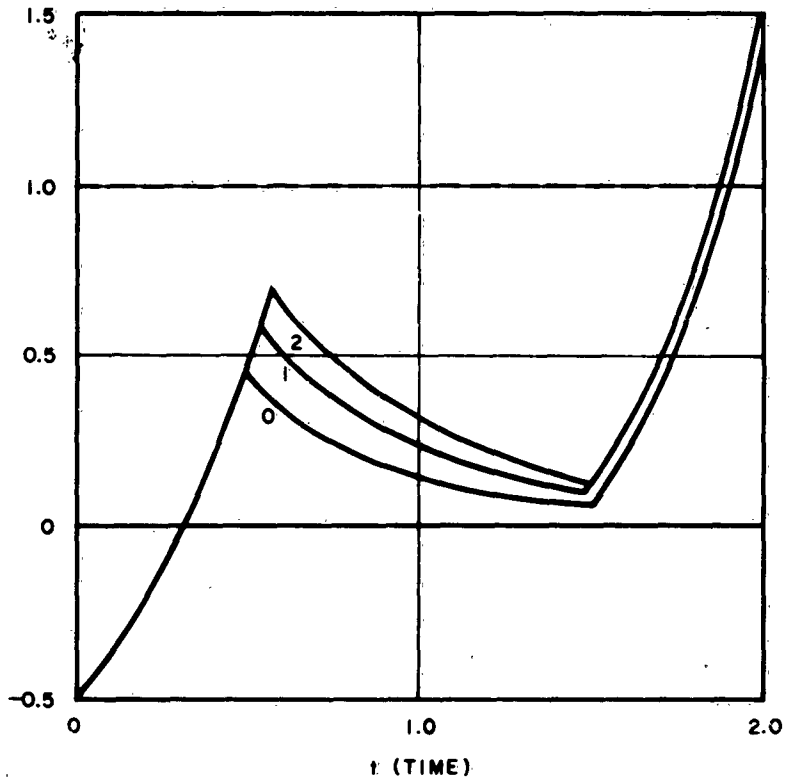


Figure 3. Sketch of Trajectories.

The steepest descent procedure outlined above can be used in more complicated examples, and the principle saving of computation and storage is due to the fact that the influence matrices only need to be stored at the switching instants. If two switching times t and z approach each other nearer than some specified ϵ, that is if $|t_1 - t_2| < \epsilon$ the intermediate mode may be omitted and the number of switches reduced. This procedure leads to a sub-optimal policy in the mathematical sense, but there are often practical reasons for avoiding a large number of switches. The available knowledge about the desired trajectory and mode properties can be used to select a reasonable nominal sequence.

A Necessary Condition for Optimal Switching

The system described by Equations (1) and (2) has the following adjoint to the first variation.

$$\dot{p}_i = - \sum_{j=1}^{n} \sum_{k=1}^{m} a_k p_j \left[\frac{\partial f^j}{\partial x^i} + \sum_{\alpha=1}^{m} \frac{\partial f^j}{\partial u^\alpha} \frac{\partial u^\alpha}{\partial x^i} \right]$$

The Hamiltonian is

$$H = \sum_{k=1}^{m} a_k \left[\underline{p} \cdot \underline{f} (\underline{x}, \underline{u}_k) \right]$$

For a minimizing trajectory the maximum principle (Reference 6) requires that the a_k be chosen to maximize H for suitably chosen boundary conditions on the multipliers p_i. As an alternative to steepest descent computation of a fixed number of switching times, a steepest descent procedure for determining the initial conditions on the multipliers may be attempted. The relative merits of these two procedures are of interest. One problem which may arise in the optimal switching case is the problem of chattering. The necessary condition outlined above gives an unambiguous specification of the switch position when one of the quantities $\underline{p} \cdot \underline{f} (\underline{x}, \underline{u}_k)$ is actually greater than any of the others. However, when there are two or more of these quantities equal to the maximum, then the choice is not clear. This equality occurs at each switching instant, and if the equality only holds for an interval of measure zero, there is no ambiguity of the resulting trajectory.

If the multiple maximum persists, then the switching choice is not defined by the maximum principle, although it may possibly be defined by the persistence of the multiple maximum. The optimum switching choice may not exist in piecewise continuous form, and the optimal control choice "chatters" between the two or more possibilities. In this case the suboptimal choice of a fixed number of switches may be more desirable computationally.

A more rigorous discussion of chattering may be based on the work of Warga, Roxin, Filippov, and Chang (References 1 through 4). The switching problem is first associated with a relaxed problem by considering values of a_k: $0 \leq a_k \leq 1$ $\Sigma a_k = 1$ instead of only $a_k = 0$ or 1, $\Sigma a_k = 1$. This relaxation allows the velocity vector to take on all the values in the simplex with corners at $\underline{f} (\underline{x}, \underline{u}_k)$ instead of only the value of the corners. For the relaxed problem a solution exists as shown in the above papers. Warga also shows that this optimal solution can be approximated arbitrarily closely by using only the original switching control. The application of this result to the problem of optimum mode switching seems to indicate that the relaxed problem should be solved for the optimal trajectory and then approximated by the switched trajectory. However, if one is going to attempt to follow a nominal, then one of the other perturbation techniques (Reference 7) may be

more feasible for closed loop operation. It appears that switching modes does not simplify the optimum trajectory computation unless a sub-optimal trajectory with a fixed number of switches is accepted.

EXAMPLE

The dynamics of the previous example are used, but an attempt is made to find the "optimal" switching. The multipliers p_i satisfy the equations

$$\dot{p}_1 = 0 \qquad p_1 = -1$$

$$\dot{p}_2 = -2(x^2 - 1)p_1 - p_2\left[(1 - a_2) - 2a_2\right]$$

and the Hamiltonian is

$$H = p_1(x_2 - 1)^2 + p_2\left[(1 - a_2)(x^2 + 2) - a_2 2x^2\right]$$

This quantity is a constant and switching can only occur when $x^2 = -2/3$ or $p_2 = 0$. If $p_2 = 0$, $\dot{p}_2 = 0$ only if $x^2 = 1$ and thus a necessary condition for chattering is $x^2 = 1$.

CONCLUSION

A steepest descent modification of the switching times in a mode switching control system can be based on first variation computations, and the computer storage requirement less than that required for functional descent. By introducing control variables the mode switching system can be put in the format usually used in the study of control optimization. The application of the maximum principle yields a necessary condition for optimum switching and a necessary condition for chattering. If the switching problem is relaxed to assure the existence of an optimal solution the computational problem of finding this optimum is not simplified by the switching approach although mechanization of the control system may be simplified. The use of steepest descent procedure for determining the sub-optimal control using a fixed number of switches is only based upon differential arguments and some knowledge of the global properties of the trajectory is required to assure the usefulness of the procedure.

REFERENCES

1. Warga, J., "Relaxed Variational Problems," J. of Math. Anal. and Appl., Vol. 4, No. 1, pp 111-128, 1962.

2. Roxin, Emilio, "The Existence of Optimal Controls," Michigan Mathematical Journal, Vol. 9, 11 109-119, 1962.

3. Filippov, A. F., "On Certain Questions in the Theory of Optimal Control," Vestnik Moskouskovo Universiteta, Ser. Mat., Mekh., Astron., Fiz., Khim. No. 2, pp 25-32, 1969.

4. Chang, S. S. L., An Extension of Ascoli's Theorem and Its Applications to the Theory of Optimal Control, New York University, College of Engineering, Technical Report 400-51.

5. Mathews, M. V. and C. W. Steeg, "Final Value Control System Synthesis," IRE Transactions on Automatic Control, pp 6-16, February 1957.

6. Boltyanskii, V. G., R. V. Gamkrelidze, and L. S. Pontryagin, "The Theory of Optimal Processes I. The Maximum Principle," Izv. Akad. Nauk. SSSR. Ser. Mat. 24, pp 3-42, 1960 (Russian).

7. Bryson, Arthur E. and Walter F. Denham, "A Guidance Scheme for Supercircular Re-entry of a Lifting Vehicle," ARS Journal, Vol. 32, No. 6, pp 894-898, June 1962.

BIBLIOGRAPHY

Balakrishnan, A. V., An Operator Theoretic Formulation of a Class of Control Problems and a Steepest Descent Method of Solution, Aerospace Corporation Report No. A-62-1730-202, Revision I, 26 July 1962.

Bellman, R., "Dynamic Programming," Chapter IX, Princeton University Press, Princeton, New Jersey, 1958.

Bellman, R., I. Glicksberg and O. Gross, "On the 'Bang-Bang' Control Problem," Quarterly Applied Mathematics 14, pp 11-18, 1956.

Bushaw, D. W., Contributions to the Theory of Nonlinear Oscillations IV - Optimal Discontinuous Forcing Terms, Princeton University Press, Princeton, New Jersey, 1958.

Eggleston, H. G., "Convexity," Cambridge Tracts in Mathematics and Mathematical Physics, No. 47, Cambridge University Press, pp 34-38, 1958.

Gamkrelidze, R. V., "Optimum-Rate Processes with Bounded Phase Coordinates," Doklady Akad. Nauk. SSSR. 125, No. 3, pp 475-478, 1959.

Gamkrelidze, R. V., "The Theory of Optimum Processes in Linear Systems," Doklady Akad. Nauk. 116, No. 1, pp 9-11, 1957.

Gamkrelidze, R. V., "Theory of Time Optimal Processes for Linear Systems," Izv. Akad. Nauk. SSSR. Ser. Mat. 22, pp 469-476, 1958 (Russian).

Ho, Y. C., "A Successive Approximation Technique for Optimal Control Systems Subject of Input Saturation," J. of Basic Engin., Vol. 84, Ser. D, No. 1, pp 33-40, 1962.

LaSalle, J. P., "The Time Optimal Control Problem," Contributions to the Theory of Nonlinear Oscillations, Vol. V, Annals of Mathematics Studies No. 45, Princeton, pp 1-24, 1960.

Lee, E. B. and L. Markus, "Optimal Control for Nonlinear Processes," Arch. Rational Mech. Anal. 8, pp 36-58, 1961.

Neustadt, L. W., "Synthesizing Time Optimal Control Systems," J. of Math. Anal. and Applic., Vol. 1, No. 4, 11 484-493, 1960.

Warga, J., "Necessary Conditions for Minimum in Relaxed Variational Problems," J. Math. Anal. and Appl. 4, 11 129-145, 1962

ASD-TDR-63-119

Optimal Thrust Programming for Minimal Fuel Midcourse Guidance

By

James S. Meditch

Aerospace Corporation
El Segundo, California

OPTIMAL THRUST PROGRAMMING FOR MINIMAL FUEL MIDCOURSE GUIDANCE

James S. Meditch

Aerospace Corporation
El Segundo, California

ABSTRACT

The problem of specifying the minimal fuel thrust program for guiding a space vehicle in its midcourse phase is formulated as an optimization problem. The task of the thrust program is to transfer the vehicle from a given initial error state (position and velocity) to a terminal error state of zero with minimum fuel subject to a magnitude constraint on the thrust vector.

The form of the optimal thrust program is derived and shown to be of an "on-off" nature with time-varying direction during periods of thrusting.

The problem of synthesizing the optimal thrust program is reduced to the problem of maximizing a particular function whose gradient is readily computed. An iterative computational procedure for synthesizing the optimal thrust program is then developed utilizing the method of steepest ascent.

INTRODUCTION

Let us consider the problem of guiding a space vehicle during the midcourse phase of its mission. Basically, the problem is that of correcting the space vehicle's ballistic or "free fall" trajectory. In general, this involves thrusting at low levels in contrast to the large amounts of thrust required for launch or deboost operations. We indicate a typical mission involving a midcourse maneuver in Figure 1.

There are two primary functions which a midcourse guidance system must perform. The first is one of sensing and processing information to determine the vehicle's state, such as position and velocity, in an appropriate coordinate system. This is the navigation function. The second is that of utilizing the navigation information to generate thrust control signals so that mission objectives are achieved. This is the thrust programming function.

Numerous problems dealing with the optimization of flight paths and guidance systems have been treated in the literature (References 1 through 6). Here we shall consider the specific problem of minimal fuel thrust programming for midcourse flight where the corrective thrust vector is amplitude constrained. The central result of the paper will be the development of an iterative computational technique for generating the thrust program. With respect to the navigation functions, we shall assume that midcourse guidance is to be initiated at a specified instant of time at which the vehicle's state is known to a desired degree of accuracy.

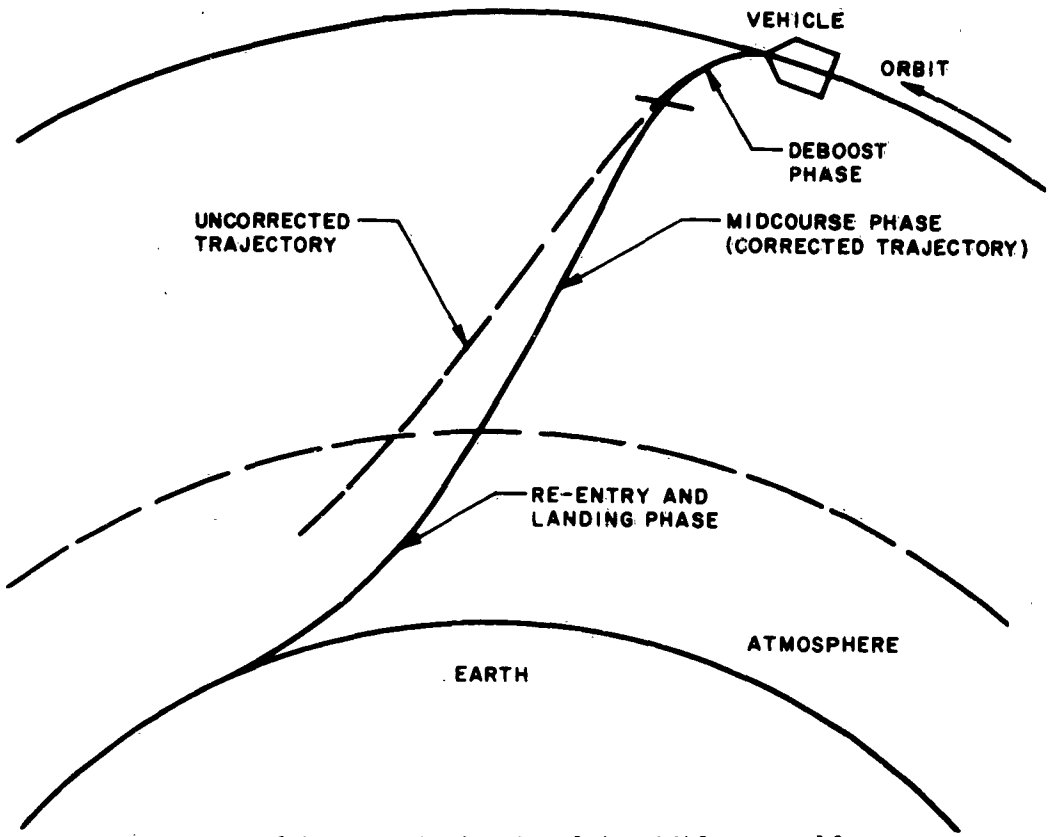


Figure 1. Typical Space Mission Involving Midcourse Maneuver.

THE MINIMAL FUEL PROBLEM

If launch or deboost operations, that is, phases of a space mission involving periods of relatively high level thrusting, could be executed perfectly, a given space vehicle would free fall to the desired destination along a path which we shall term the nominal trajectory. However, imperfect launch or deboost operations cause errors to exist along the ballistic or free fall portion of the mission. We use the term errors here to denote position and velocity deviations from the nominal trajectory. We assume these errors are small enough to permit a linearization of the vehicle's equations of motion.

To within the accuracy limitations of the guidance elements utilized, we shall require that the corrections reduce the errors to zero at the conclusion of the midcourse phase. We shall also require that the midcourse phase be executed in a fixed time so that the vehicle will arrive at the same terminal state at the same time as a vehicle following the nominal trajectory. By

imposing these two requirements, we have considerably reduced the task of the terminal guidance system in executing re-entry, rendezvous, or docking type maneuvers.

In order to maximize the useful payload, we shall choose quantity of fuel as the system parameter to be minimized. Moreover, for obvious practical reasons, we shall assume the corrective thrust is amplitude constrained.

The equations of motion of a space vehicle subject only to gravitational and propulsive forces may be written as

$$\dot{y} = P(y, t) + Q(y)\, T(t) \tag{1}$$

for either cartesian or spherical coordinates. In Equation (1), y is a six-dimensional column vector whose elements are the position and velocity coordinates of the vehicle, P is a six-dimensional column vector representing the gravitational forces, and Q (y) is a 6 x 3 matrix which relates the three-dimensional thrust vector T to $\dot{y}$.

Assuming small errors, we shall now linearize Equation (1). We let $y = Y + \delta y$ where Y is the nominal trajectory which satisfies the free fall equation

$$\dot{Y} = P(Y, t). \tag{2}$$

The initial conditions on Equation (2) are those which would result from an ideal (perfect) launch or deboost operation.

Since y satisfies Equation (1), δy satisfies the equation

$$\delta\dot{y} = \left(\frac{\partial P}{\partial y}\right) \delta y + Q(y)\, T(t) + \ldots, \tag{3}$$

where $(\partial P/\partial y)$ is a 6 x 6 matrix (evaluated along the nominal trajectory Y), and the dots on the right denote higher order terms in δy.

Assuming Q (y) is approximated closely enough by Q (Y) and neglecting the higher order terms, Equation (3) becomes

$$\delta\dot{y} = \left(\frac{\partial P}{\partial y}\right) \delta y + Q(Y)\, T(t). \tag{4}$$

Since the time history of the nominal trajectory is known, we observe that $(\partial P/\partial y)$ and Q (Y) are explicit functions of time. Hence, defining

$$x(t) = \delta y(t) \qquad A(t) = \left(\frac{\partial P}{\partial y}\right)$$

$$u(t) = T(t) \qquad B(t) = Q(Y),$$

Equation (4) becomes

$$\dot{x}(t) = A(t)\,x(t) + B(t)\,u(t). \tag{5}$$

In Equation (5), $x(t)$ is a six-dimensional column vector, the vehicle state vector; $u(t)$ is a three-dimensional column vector, the thrust vector; and $A(t)$ and $B(t)$ are 6×6 and 6×3 matrices, respectively.

We shall let the time interval over which midcourse guidance is performed be the interval $0 \le t \le \tau$, where τ is a known constant. We shall denote the initial errors or deviations from the nominal trajectory by $x(0)$. The task of the midcourse guidance system is then to null these errors so that $x(\tau) = 0$.

For low-level thrust propulsion systems operating in free space, it is reasonable to assume the specific impulse I_{sp}, given by

$$I_{sp} = \frac{|T(t)|}{g_o\,\dot{m}}, \tag{6}$$

is constant (Reference 7). In Equation (6), $|T(t)|$ is the magnitude of the thrust vector, g_o is the acceleration of gravity at sea level, and $\dot{m}$ is the time rate of change of the fuel mass. Cross-multiplying in Equation (6), and integrating the result between the limits 0 and τ, we obtain

$$\int_o^\tau |T(t)|\,dt = g_o I_{sp} \int_o^\tau \dot{m}\,dt = g_o I_{sp} \int_o^\tau dm. \tag{7}$$

But $g_o \int_o^\tau dm$ is simply the total weight of the fuel consumed during the midcourse phase. From Equation (7) then, we see that the time integral of the thrust vector magnitude is proportional to the fuel consumed. Therefore, in order to minimize the fuel consumed during the midcourse phase, we shall choose

$$S = \int_o^\tau |T(t)|\,dt,$$

or equivalently,

$$S = \int_o^\tau \|u(t)\|\,dt, \tag{8}$$

as the design parameter or cost to be minimized. In Equation (8), the symbol $\|\ \|$ denotes the Euclidean norm as given by

$$\|u(t)\| = \sqrt{u_1^2(t) + u_2^2(t) + u_3^2(t)}.$$

This is simply the magnitude of the thrust vector.

In minimizing Equation (8), we shall restrict ourselves to thrust vectors for which: (a) $\|u(t)\| \leq 1$, and (b) each component of $u(t)$ is piecewise continuous. Thrust vectors which satisfy these two restrictions for $0 \leq t \leq \tau$ are termed admissible.

In practice, condition (a) is equivalent to requiring $|T(t)| \leq T_o$ where T_o is a known constant equal to the maximum thrust available from the propulsion system. For analytic simplicity, we have normalized this constraint. Condition (b) is one of requiring that the thrust vector be physically realizable.

We now summarize the problem formulation: We are given the linearized system of Equation (5) of a space vehicle in midcourse flight. We wish to determine an admissible thrust program $u(t)$ which will transfer the vehicle from a known initial state $x(0)$ to the terminal state $x(\tau) = 0$ in a fixed time τ such that the cost (Equation (8)) is minimized. Such an admissible thrust program will be called an optimal thrust program.

DERIVATION OF THE OPTIMAL THRUST PROGRAM

For a given initial state $x(0)$ and an admissible thrust program $u(t)$, the state of the system (Equation (5)) at any time t, $0 \leq t \leq \tau$, is given by

$$x(t) = X(t)\left[x(0) + \int_o^t X^{-1}(s)\, B(s)\, u(s)\, ds\right], \qquad (9)$$

where $X(t)$ is the 6×6 matrix solution of $\dot{X}(t) = A(t)\, X(t)$, and $X(0) = I$ = identity matrix.

From Equation (9), we see that we can achieve the desired terminal state $x(\tau) = 0$ if and only if

$$-x(0) = \int_o^\tau X^{-1}(s)\, B(s)\, u(s)\, ds. \qquad (10)$$

Let us define an additional state variable y by

$$y(t) = \int_o^t \|u(s)\|\, ds.$$

We note from Equation (8) that $S = y(\tau)$.

Let us consider the set Ω defined by the relation

$$\Omega = \left\{ \left[\int_0^\tau X^{-1}(t)\, B(t)\, u(t)\, dt, \int_0^\tau \|u(t)\|\, dt \right] : u(t) \text{ admissible} \right\}.$$

We see that Ω is a set in seven-dimensional Euclidean space, and that a point (x, y) belongs to Ω if and only if there exists an admissible thrust program $u(t)$ which transfers the initial state $-x$ to the origin in time τ with a cost y.†

It is shown in Reference 8 that Ω is convex, closed, and bounded, and that it is symmetric about the y (cost) axis.

For a given $x(0)$, the minimum cost to reach the origin in time τ is the least $y(\tau)$ for which $(-x(0), y(\tau)) \in \Omega$. We shall denote this minimum cost by $y^o(\tau)$. Clearly, $(-x(0), y^o(\tau))$ is a boundary point of Ω. We let $\underset{\sim}{x}_o = (-x(0), y^o(\tau))$. (Bold faced letters will be used throughout to indicate seven-dimensional vectors.)

Since Ω is convex, there exists at least one seven-dimensional column vector $\underset{\sim}{\eta}$ such that

$$\underset{\sim}{\eta}' \cdot \underset{\sim}{x}_o \geq \underset{\sim}{\eta}' \cdot \underset{\sim}{\omega}, \tag{11}$$

for all $\underset{\sim}{\omega} \in \Omega$ where the prime denotes the transpose. That is, we may construct a hyperplane of support to Ω at $\underset{\sim}{x}_o$ with $\underset{\sim}{\eta}$ as any vector normal to this hyperplane at $\underset{\sim}{x}_o$.

From the properties of Ω it can be shown that the last (seventh) component of $\underset{\sim}{\eta}$ is nonpositive. With little loss of generality, we shall assume it is negative. Moreover, since the length of $\underset{\sim}{\eta}$ is immaterial, we shall set its last component equal to -1 for all the work which follows.

From Equation (11), we see that the function $\underset{\sim}{\eta}' \cdot \underset{\sim}{\omega}$ attains its maximum in Ω when $\underset{\sim}{\omega} = \underset{\sim}{x}_o$. For any $\underset{\sim}{\omega} \in \Omega$ given by

$$\underset{\sim}{\omega} = \left[\int_0^\tau X^{-1}(t)\, B(t)\, u(t)\, dt, \int_0^\tau \|u(t)\|\, dt \right],$$

this function becomes

†It is conceivable that for sufficiently large initial errors $x(0)$, it would be impossible to achieve $x(\tau) = 0$ even if full thrust $\|u(t)\| = 1$ were utilized throughout the interval $0 \leq t \leq \tau$. In our work here, we shall consider only those initial errors which can be transferred to the origin in time τ using admissible thrust programs.

$$\underset{\sim}{\eta}' \cdot \underset{\sim}{\omega} = \int_0^\tau \left[\eta' \cdot X^{-1}(t)\, B(t)\, u(t) - \|u(t)\|\right] dt, \tag{12}$$

where, from above, $\underset{\sim}{\eta} = (\eta, -1)$ and $\eta = (\eta_1, \ldots, \eta_6)$.

We wish to select an admissible thrust program $u(t)$ which maximizes $\underset{\sim}{\eta}' \cdot \underset{\sim}{\omega}$. We maximize this function by maximizing the integrand of Equation (12) for all t, $0 \le t \le \tau$. Thus, we wish to maximize the function

$$f(t) = \eta' \cdot X^{-1}(t)\, B(t)\, u(t) - \|u(t)\| \tag{13}$$

with respect to the variable u subject to the constraint $\|u\| \le 1$.

We observe that $\eta' \cdot X^{-1}(t)\, B(t)$ is a three-dimensional row vector. If $\eta' \cdot X^{-1}(t)\, B(t) = 0$, we maximize $f(t)$ by setting $u(t) = 0$. If $\eta' \cdot X^{-1}(t)\, B(t) \neq 0$, $f(t)$ will be maximized only if $u(t)$ has the same direction as $\eta' \cdot X^{-1}(t)\, B(t)$. We shall assume this is the case in the work which follows. Hence, for $\eta' \cdot X^{-1}(t)\, B(t) \neq 0$, Equation (13) becomes

$$f(t) = \|u(t)\| \left[\|\eta' \cdot X^{-1}(t)\, B(t)\| - 1\right]. \tag{14}$$

If $\|\eta' \cdot X^{-1}(t)\, B(t)\| > 1$, $f(t)$ is maximized by making $\|u(t)\| = 1$, the maximum allowable. Since $u(t)$ has the same direction as $\eta' \cdot X^{-1}(t)\, B(t)$, we obtain

$$u(t) = \frac{\left[\eta' \cdot X^{-1}(t)\, B(t)\right]'}{\|\eta' \cdot X^{-1}(t)\, B(t)\|} .$$

If $\|\eta' \cdot X^{-1}(t)\, B(t)\| < 1$, $f(t)$ is maximized by setting $\|u(t)\| = 0$ which means $u(t) = 0$.

Finally, if $\|\eta' \cdot X^{-1}(t)\, B(t)\| = 1$, we have $f(t) \equiv 0$ for any admissible $u(t)$. For purposes of our work here, we shall assume that the set of points in the interval $[0, \tau]$ at which $\|\eta' \cdot X^{-1}(t)\, B(t)\| = 1$ is of measure zero for every vector η. For the sake of completeness, however, we set $u(t) = 0$ whenever $\|\eta \cdot X^{-1}(t)\, B(t)\| = 1$. For the above assumption then, we have that the thrust program $u(t)$ which maximizes $\underset{\sim}{\eta}' \cdot \underset{\sim}{\omega}$ is unique and defined almost everywhere.

Because of the uniqueness, we can replace inequality (11) by the strict inequality

$$\underset{\sim}{\eta}' \cdot \underset{\sim}{x}_0 > \underset{\sim}{\eta}' \cdot \underset{\sim}{\omega}, \quad \underset{\sim}{\omega} \in \Omega \text{ and } \underset{\sim}{\omega} \neq \underset{\sim}{x}_0 \tag{15}$$

In summary, the thrust program which maximizes $\underset{\sim}{\eta}' \cdot \underset{\sim}{\omega}$ (and thereby minimizes the cost) is given by the relation

$$u(t) = \begin{cases} \dfrac{\left[\eta' \cdot X^{-1}(t)\, B(t)\right]'}{\left\|\eta' \cdot X^{-1}(t)\, B(t)\right\|} & \text{whenever } \left\|\eta' \cdot X^{-1}(t)\, B(t)\right\| > 1 \\ 0 & \text{whenever } \left\|\eta' \cdot X^{-1}(t)\, B(t)\right\| \leq 1, \end{cases} \tag{16}$$

for $0 \leq t \leq \tau$.

SYNTHESIS OF THE OPTIMAL THRUST PROGRAM

We see from Equation (16) that the problem of generating the optimal thrust program is one of computing an η for a given initial state $x(0)$.

Let us consider a seven-dimensional column vector $\underset{\sim}{\lambda} = (\lambda, -1)$ where $\lambda = (\lambda_1, \ldots, \lambda_6)$. For a given λ, let us define the function

$$v(t, \lambda) = \begin{cases} \dfrac{\left[\lambda' \cdot X^{-1}(t)\, B(t)\right]'}{\left\|\lambda' \cdot X^{-1}(t)\, B(t)\right\|} & \text{whenever } \left\|\lambda' \cdot X^{-1}(t)\, B(t)\right\| > 1 \\ 0 & \text{whenever } \left\|\lambda' \cdot X^{-1}(t)\, B(t)\right\| \leq 1, \end{cases} \tag{17}$$

for $0 \leq t \leq \tau$. Let us introduce the vector function

$$\underset{\sim}{z}(\lambda) = \left[\int_0^\tau X^{-1}(t)\, B(t)\, v(t, \lambda)\, dt, \int_0^\tau \left\|v(t, \lambda)\right\| dt\right] \tag{18}$$

Analogous to Equation (15), we obtain

$$\underset{\sim}{\lambda}' \cdot \underset{\sim}{z}(\lambda) > \underset{\sim}{\lambda}' \cdot \underset{\sim}{\zeta};\ \underset{\sim}{\zeta} \in \Omega \text{ and } \underset{\sim}{\zeta} \neq \underset{\sim}{z}(\lambda). \tag{19}$$

Hence, $\underset{\sim}{z}(\lambda)$ is a boundary point of Ω and a hyperplane of support to Ω at $\underset{\sim}{z}(\lambda)$ has the normal vector $\underset{\sim}{\lambda} = (\lambda, -1)$.

Now if a vector η is known for a given initial state $x(0)$, the corresponding point on the boundary of Ω has the coordinates

$$\underset{\sim}{z}(\eta) = (-x(0), y^{o}(\tau)). \tag{20}$$

Let us consider any vector λ for which $\underset{\sim}{z}(\lambda) \neq \underset{\sim}{z}(\eta)$. We shall denote the hyperplane of support to Ω at $\underset{\sim}{z}(\lambda)$ by $B(\lambda)$. If $\underset{\sim}{\xi} = (\xi_1, \ldots, \xi_6, \xi_7)$ is any point in the seven-dimensional Euclidean space which also lies in $B(\lambda)$, we obtain

$$\underset{\sim}{\lambda}' \cdot \underset{\sim}{z}(\lambda) = \underset{\sim}{\lambda}' \cdot \underset{\sim}{\xi} \tag{21}$$

as the equation for this plane.

The line through the point $(-x(0), 0)$ parallel to the y axis intersects $B(\lambda)$ at some point $\underset{\sim}{\xi} = [-x(0), \rho]$. Setting $\underset{\sim}{\xi} = \underset{\sim}{\Xi}$ in Equation (21), we obtain

$$\underset{\sim}{\lambda}' \cdot \underset{\sim}{z}(\lambda) = \underset{\sim}{\lambda}' \cdot (-x(0), \rho) . \tag{22}$$

Recalling $\underset{\sim}{z}(\lambda) \neq \underset{\sim}{z}(\eta)$, it follows from Equation (19) that

$$\underset{\sim}{\lambda}' \cdot \underset{\sim}{z}(\lambda) > \underset{\sim}{\lambda}' \cdot \underset{\sim}{z}(\eta). \tag{23}$$

From Equations (20), (22), and (23), we obtain

$$\underset{\sim}{\lambda}' \cdot (-x(0), \rho) > \underset{\sim}{\lambda}' \cdot (-x(0), y^{o}(\tau)). \tag{24}$$

From Equation (24), it then readily follows that

$$\rho < y^{o}(\tau). \tag{25}$$

Solving Equation (22) for ρ yields

$$\rho(\lambda) = -\lambda' \cdot x(0) - g(\lambda), \tag{26}$$

where we have made the definition $g(\lambda) = \underset{\sim}{\lambda}' \cdot \underset{\sim}{z}(\lambda)$ and have introduced the notation $\rho(\lambda)$ to emphasize the fact that ρ is a function of λ. We note, from Equation (18), that

$$g(\lambda) = \underset{\sim}{\lambda}' \cdot \underset{\sim}{z}(\lambda) = \int_{o}^{\tau} \left[\lambda' \cdot X^{-1}(t)\, B(t)\, v(t, \lambda) - \|v(t, \lambda)\|\right] dt, \tag{27}$$

which is clearly a function of λ.

Now if $\underset{\sim}{z}(\lambda) = \underset{\sim}{z}(\eta)$, we see from Equations (20) and (22) that

$$\rho(\lambda) = y^{o}(\tau). \tag{28}$$

From Equation (25), we recall $\rho(\lambda) < y^{o}(\tau)$ if $\underset{\sim}{z}(\lambda) \neq \underset{\sim}{z}(\eta)$. Thus, Equation (28) holds if and only if $\underset{\sim}{z}(\lambda) = \underset{\sim}{z}(\eta)$.

From Equations (25) and (28) then, we see that the problem of synthesizing the optimal thrust program is one of determining a six-dimensional vector λ which maximizes the function $\rho(\lambda)$ given in Equation (26). We shall perform this maximization using the method of steepest ascent (Reference 9).

The method involves making λ a function of some parameter p and solving the differential equation

$$\frac{d\lambda}{dp} = k \nabla\rho(\lambda) \tag{29}$$

for $\lambda(p)$. In Equation (29), k is some p onstant and ∇ denotes the gradient.

Let us now consider the limit

$$\lambda = \lim_{p\to\infty} \lambda(p) .$$

We shall show that if this limit exists, it is precisely a desired λ for synthesizing the optimal control for a given $x(0)$.

It can be shown (Reference 10) that $\nabla\rho(\lambda)$ (note Equations (26) and (27)) is continuous in λ and is given by

$$\nabla\rho(\lambda) = -x(0) - \int_0^\tau \mathcal{X}^{-1}(t)\, B(t)\, v(t, \lambda)\, dt , \tag{30}$$

Assuming

$$\lim_{p\to\infty} \lambda(p) = \lambda^o$$

where λ^o is a constant six-dimensional column vector, we obtain

$$\lim_{p\to\infty} \nabla\rho\left[\lambda(p)\right] = \nabla\rho(\lambda^o),$$

since $\nabla\rho(\lambda)$ is continuous. From Equation (29) it now follows that

$$\frac{d\lambda}{dp} \xrightarrow[\lambda\to\lambda^o]{} k\,\nabla\rho(\lambda^o) = 0 .$$

Thus, as $\lambda\to\lambda^o$, we have $d\lambda/dp\to 0$. (Otherwise, λ would grow without bound as $p\to\infty$ contradicting our assumption that λ approaches a limit.) Then, for $\lambda = \lambda^o$, we see from Equation (30) that

$$-x(0) = \int_0^\tau X^{-1}(t)\, B(t)\, v(t, \lambda^o)\, dt . \tag{31}$$

This means $v(t, \lambda^o)$ is precisely the thrust program required to transfer the space vehicle from the point $x(0)$ to the desired terminal state $x(\tau) = 0$. It is very easy to show (Reference 8) that $\underset{\sim}{z}(\lambda^o) = \underset{\sim}{z}(\eta)$, and, therefore, that $v(t, \lambda^o)$ is an optimal thrust program.

If we substitute Equation (30) into Equation (29), we obtain

$$\frac{d\lambda}{dp} = -k \left[x(0) + \int_0^\tau X^{-1}(t)\, B(t)\, v(t, \lambda)\, dt \right],$$

or equivalently,

$$\frac{d\lambda}{dp} = -k\, X^{-1}(\tau) \left\{ X(\tau) \left[x(0) + \int_0^\tau X^{-1}(t)\, B(t)\, v(t, \lambda)\, dt \right] \right\}. \tag{32}$$

We observe that the term within the braces is simply the solution $x(t)$ of Equation (5) with $u(t) = v(t, \lambda)$ evaluated at the terminal time $t = \tau$. Since this solution depends upon the choice of λ, we shall adopt the notation

$$x(\tau, \lambda) = X(\tau) \left[x(0) + \int_0^\tau X^{-1}(t)\, B(t)\, v(t, \lambda)\, dt \right]. \tag{33}$$

Substituting Equation (33) into Equation (32), we have

$$\frac{d\lambda}{dp} = -k\, X^{-1}(\tau)\, x(\tau, \lambda),$$

or, in discrete form,

$$\lambda^{(i+1)} = \lambda^{(i)} - KX^{-1}(\tau)\, x(\tau, \lambda^{(i)}), \tag{34}$$

where K is some positive constant and i is the index of iteration.

We now outline the steps of an iterative computational procedure for determining a correct λ:

(1) Make an arbitrary initial "guess" $\lambda^{(o)}$ for λ.

(2) For this value of $\lambda^{(i)}$, compute $v(t, \lambda^{(i)})$ from Equation (17).

(3) Integrate the equations of motion (Equation (5)) using the thrust program of step (2) to determine $x(\tau, \lambda^{(i)})$. Equation (33) may be used directly for this purpose.

(4) Substitute this value of $x(\tau, \lambda^{(i)})$ into Equation (34), solve for $\lambda^{(i+1)}$, and return to step (2).

(5) Terminate the iteration process when $\|\lambda^{(i+1)} - \lambda^{(i)}\| < \epsilon$ where ϵ is a small positive number.

The optimal thrust program is then given by Equation (17) using the last iteration on λ.

DISCUSSION OF RESULTS

The utility of space missions is directly dependent on their payload capability. In minimizing the amount of fuel used in the midcourse phase, our intent has been to increase this capability, that is, to permit the inclusion of additional useful equipment. We must bear in mind that a savings of only a few pounds may permit a mission to include scientific experiments which would not be possible otherwise. This is a significant factor when viewed in terms of the over-all cost of a space mission.

In developing the equations of motion of a space vehicle in midcourse flight, we assumed the mass M of the vehicle was appreciably larger than the mass of fuel m consumed during the midcourse phase. More realistically, of course, we should have considered the variations in total mass along the flight. However, if the fuel mass for the midcourse phase should turn out to represent a significant percentage of the total vehicle mass, this would raise the question of the effectiveness of the launch or deboost guidance system. In general, we would expect errors to be small at the conclusion of launch or deboost operations. The task of midcourse guidance then becomes that of making small corrections so that these errors do not propagate along the trajectory and cause large misses at the destination. Hence, for purposes of this preliminary study, it is felt the assumption $M >> m$ is reasonable.

We observe from Equation (17) that whether or not the vehicle is thrusting is controlled by whether or not $\|\lambda' \cdot X^{-1}(t)\, B(t)\|$ exceeds a threshold equal to unity. Moreover, we note that when the vehicle is thrusting, the system utilizes the full thrust capability by making $\|v(t, \lambda)\| = 1$ and changes only the thrust direction in accordance with the components of $\lambda' \cdot X^{-1}(t) B(t)$. As a result of this "on-off" nature of the thrust program, non-throttleable propulsion can be used. However, since the direction of thrust is time-varying, the attitude control system must be utilized during periods of thrusting.

Since the initial error $x(0)$ is a measured quantity, i.e., the output of the navigation scheme employed, the thrust program performance is directly dependent upon the quality of navigation. By allowing the navigation system more smoothing time, i.e., deferring the initiation of midcourse guidance, one obtains a more accurate estimate of $x(0)$. However, in waiting, one allows the launch or deboost burnout errors to propagate unchecked. A trade-off between these two factors is essential in the selection of the time at which midcourse guidance should be initiated.

In conclusion, we point out that whether or not the midcourse guidance scheme developed here is practical from a systems-hardware point of view is a question requiring further study. In this section, we have pointed out a number of factors, pro and con. Certainly, the discussion is not exhaustive. In any event, it is felt that results developed here could be exploited as guide lines in other similar studies. For example, the amount of fuel required to execute a mission using the above scheme could be used as an ultimate performance limitation in the design of other midcourse guidance systems for the same mission.

REFERENCES

1. Kelley, H. J., "Gradient Theory of Optimal Flight Paths," ARS Semi-Annual Meeting, Los Angeles, California, 9-12 May 1960.

2. Lukes, D., "Application of Pontryagin's Maximum Principle in Determining the Optimal Control of a Variable Mass Vehicle," ARS Paper No. 1927-61, August 1961.

3. Isaev, V. K., "L. S. Pontryagin's Maximum Principle and Optimal Programming of Rocket Thrust," Automation and Remote Control, Vol 22, No. 8, pp 881-893, 1961.

4. Kelley, H. J., "Guidance Theory and Extremal Fields," National Aerospace Electronics Convention, Dayton, Ohio, 14-16 May 1962.

5. Bryson, A. E., F. J. Carroll, W. F. Denham, and K. Mikami, "Determination of the Lift or Drag Program that Minimizes Re-entry Heating with Acceleration or Range Constraints Using a Steepest Descent Computation Procedure," IAS Meeting, New York, N. Y., January 1961.

6. Leitmann, G., "On a Class of Variational Problems in Rocket Flight," Journal of Aero Space Science, Vol 26, No. 9, pp 586-591, 1959.

7. Seifert, H., Editor, Space Technology, John Wiley and Sons, Inc., New York, 1959.

8. Meditch, J. S., and L. W. Neustadt, "An Application of Optimal Control to Midcourse Guidance," submitted to the Second Congress of IFAC, Basle, Switzerland, September 1963.

9. Rosenbloom, P. C., "The Method of Steepest Descent," Proc. of the Sixth Symposium in Applied Mathematics, McGraw-Hill Book Co. New York, N. Y., pp 127-176, 1956.

10. Neustadt, L. W., "On Synthesizing Optimal Controls," submitted to the Second Congress of IFAC, Basle, Switzerland, September 1963.

ASD-TDR-63-119

The Synthesis of Optimal Controllers

By

Bernard H. Paiewonsky

Aeronautical Research Associates
of
Princeton, Inc.
Princeton, New Jersey

THE SYNTHESIS OF OPTIMAL CONTROLLERS

Bernard H. Paiewonsky

Aeronautical Research Associates of Princeton, Inc.

SUMMARY

A review of the literature shows that the problem of optimal control synthesis is not completely solved although many aspects of the problem have received considerable study. An iterative procedure for the synthesis of time optimal controllers is described and some experimental results are presented. It is well known that the optimal control law for the time optimal control of linear systems with bounded control amplitude is of bang-bang type while the time optimal rocket steering control law is continuous. In both cases, the solution of the associated two-point boundary problem requires the determination of (n - 1) parameters for an n^{th} order system. The computational procedure described here is based on the idea of optimal evolution and uses the method of steepest descent to obtain the proper values of the n - 1 parameters. The geometrical significance of these parameters as well as their relationship to the reachable sets is discussed. An experimental investigation of the convergence of the iterative procedure is described for both regulator and controller problems. The use of this iterative technique for variable input signals has been found to work in those cases investigated. The development of the theory shows that convergence to the correct solution cannot always be guaranteed when working with variable inputs. For example, it is possible for moving targets to "out run" or "out maneuver" the system.

The results presented for the bang-bang problem were obtained on second order systems. The rocket steering problems provided third and fourth order examples. A closed loop optimal controller for optimal rocket steering has been programmed and is presently being studied on a digital computer.

INTRODUCTION

A. A dynamic system is to be controlled in such a way as to maximize or minimize some performance functional. The state of the system at any instant is described by n-independent variables $\chi_i(t)$, $1 \leq i \leq n$. These variables change with time according to a set of differential equations $\dot{\chi}_i = f_i(\chi_j, U_K, t)$ where the U_K are independent controls, $1 \leq K \leq R \leq n$. These are the governing equations for the system. The variables χ_i may or may not correspond directly to the physical variables naturally associated with the process. In addition to the governing equations there

are constraints on the state variables and on the controls. These constraints are often in the form of inequalities $A(\chi_i, U_K) \leq 0$ or $\int_o^T B(\chi_i, U_K) \leq M$. The functional to be extremized is often taken to be an integral of the form $S = \int_o^T L(\chi_i, U_K, t)dt$ where the function L is called the error criterion and the functional S is called alternatively the performance index, the payoff or the cost.

In realistic cases unwanted disturbances enter the system along with noise in the sensors that observe the state variables or the measurable physical variables. The controller may also suffer from imperfections of various types and in addition the controller may have only imperfect knowledge of the process or system being controlled.

The initial and desired terminal states of the system are prescribed. The problem consists in finding a control law u(t) in the class of allowable controls that brings the state point from initial to terminal locations and extremizes the payoff or cost.

A synthesis procedure for an optimal system results in a closed feedback loop representing the system being controlled and a computer to select the optimal control law. The block diagram in Figure 1 illustrates the main ideas.

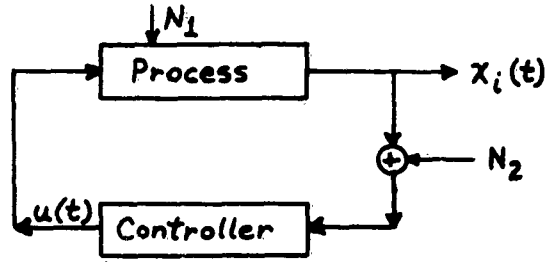


FIGURE 1. OPTIMAL FEEDBACK CONTROLLER

It is necessary to incorporate the noise and random inputs from the beginning and to determine the extent to which these disturbances must be taken into account in selecting a performance criterion.

The synthesis procedure involves several distinct steps. Each of these steps is an interesting area for research on its own merits. However, no single step, by itself is sufficient to constitute a synthesis procedure. These steps include the selection of a performance criterion, determination of the system state point from noisy data, estimation of statistical properties of noise and random inputs, and computation of the optimal control law.

B. Historical background. A search of the literature discloses many examples of a specific nature as well as several thorough treatments of general methods of formulating problems of

optimal control. However, very little has been published regarding actual synthesis of real time optimal controllers. Research on real time optimal control appears to be confined in the main to separate studies of one or more of the individual steps described previously. However, there are some results on synthesis of predictive but not necessarily optimal controllers.

The reduction of the synthesis problem to studies of individual simpler problems follows the traditional pattern. General studies of dynamical systems lead to special studies of properties of controlled systems. The historical development of control theory usually emphasizes stability. Although this is always a consideration, it is not the main concern here. The optimization problem is frequently treated separately and ordinarily is formulated in terms of calculus of variations, either in classical form or by using direct methods. The treatment of the optimization problems, in a great many cases, is aimed at finding solutions for a few initial states and limited number of terminal states. The two-point boundary problem is ordinarily solved for only a restricted class of initial and end conditions. The solution of two-point boundary problems constitutes another large area of research and again there are a number of general studies and a number of very specific studies.

Finally, there is a class of papers devoted mainly to applications of previously derived theoretical and computational results to specific systems. These papers encompass aircraft performance studies, space travels, chemical plant operation, instrument servo designs, optimal noise filters, and a variety of other special applications. It goes without saying that there are papers that fit into more than one of the general categories described as well as articles of interest that do not fit well into any one of these general areas.

The behavior of dynamic systems subject to random inputs or disturbances has some connection with research on optimal control synthesis. There is a relation between certain random problems of control theory and problems arising from physics, for example Brownian motion. The similarity between problems is very close and the same kind of equation (Fokker-Planck) arises from both classes of problems. Blackman (1) Hopkin and Wang (2) and Barrett (3) have studied this. The book by Wax (4) contains a collection of stochastic problems of physics that relate to control theory. Texts by Tsien (5) Lanning and Battin (6) and Petersen (7) have discussed random processes and control theory in detail. In addition, recent papers by Aoki (8) Bellman (9) Adorno (10) and Zadeh (11) have also been concerned with control and random processes.

The design of optimal controllers depends upon the selection of performance criteria. Various kinds of standards, performance criteria, and error indices have been proposed and investigated (12,13).

The theory of optimization has developed along two parallel branches stemming from mathematical and engineering interests. Formulations of the mathematical problem have been made by Bliss (14) Hestenes (15) Kalman (16) Marcus and Lee (17) Breakwell (18) Bryson (19) Bellman (20) Kalaba (21) La Salle (22) Pontriagin (23) Kipiniak (24) Chang (25) DeSoer (26) and many others. A review of the literature relating to time optimal control has been prepared by Paiewonsky (27). At the present time, the formulation of optimization problems and the resulting necessary conditions are fairly well understood. A possible exception is the case of inequality constraints on the state variables.

Once the problem has been formulated in precise mathematical terms, the next step is to obtain solutions. The solutions can be the optimal path that the system should follow or the optimal control law to be applied to the system.

The solution of the optimization equations has been attempted by many investigators using several different techniques. For the most part, these investigations have been directed towards solutions of a special class of optimization problems.

The direct integration of the Euler equations has been tried many times. Integrations are tried forward, backwards, or in combination. There are several ways to satisfy the boundary conditions. These include (a) trial and error searching for the Lagrange multipliers, (b) flooding or calculating a great number of trajectories by systematically varying multipliers at initial or terminal points, (c) guessing and refining the initial results by Newton's method, influence functions methods and other means.

It has been known for a long time that the direct integration of the equations of the system coupled to the Euler equations results in a set of equations with extreme sensitivity to initial or terminal conditions.

Another approach to the solution of the optimization equations is the partial differential equation or payoff surface approach. The justification for these methods is based on the fact that for most systems the surfaces of constant optimal payoff are boundaries of convex sets and the gradient of the surface (if it exists) corresponds to the Lagrange multipliers of the classical approach or the adjoint variables of the maximum principle. These optimal payoff surfaces are solutions of a partial differential equation analogous to the Hamilton-Jacobi equation obtainable from the maximum principle. The properties of these surfaces, also called reachable sets, have been studied by Halkin (28) Paiewonsky (27) Roxin (29) and Anderson (30). The characteristics of this partial differential equation satisfy the Euler equations. A discrete version of this partial differential equation is the basis for the method of Dynamic Programming developed by Bellman. This approach has been applied to many problems with varying degrees of success. A technique based on results of Neustadt (31) appears promising for optimal problems with linear processes. This will be dis-

cussed in more detail in a following section.

The methods listed above have in common the idea of searching for a function that satisfies initial and terminal conditions among a class of functions satisfying certain necessary conditions. An alternative approach has been taken by several investigators. In particular, Kelley (32) Bryson (33) Swanlund (34) Kipiniak (24) Ho (35) and Dreyfus (36) have investigated the gradient methods. These methods seek a control function that extremizes the performance functional from the class of functions that satisfy the boundary conditions. The method depends upon successive approximations to continually improve the value of the performance index. In the course of the "descent" it is necessary to adjust the procedure to satisfy the boundary conditions which may be drifting off. Several special procedures developed to do this have been reported on and are included in the bibliography. Analog computers for solving these problems have been investigated by Kipiniak (24) and Meers (37).

The optimal control of quantized or discrete systems has been studied by DeSoer (38) Friedland (39) Kulikowski (40) Neustadt (41) and Zadeh (11) among others.

The idea of a field of extremals is fairly old, however, it has been only recently that this notion has been applied to optimal control problems to obtain neighboring solutions to a known optimal trajectory.

Perturbation techniques designed to obtain additional solutions to boundary problems by examining neighboring solutions have been studied by Bryson (42) Kelley (43) Kipiniak (24) Swanlund (34) and Dow (44) among others.

There is such an abundant literature on applications of optimization procedures to problems in flight mechanics that it is impractical to do more here than to cite some of the more prominent works and to describe in generalities what has been accomplished.

It is important to observe that the objective in many of these studies is to find only a few solutions to a limited class of boundary conditions and the question of synthesis of a controller does not arise.

Classical variational theory has been used by Leitmann (45) Behrbohm (46) Miele (47) Melbourne (48) and Edelbaum (49) among others to study optimal paths for aircraft, rockets, and space craft. Lawden (50) employed direct methods for optimal orbital transfers, while Bellman (51) and Dreyfus (52) and Smith (53) use dynamic programming.

Optimal reentry and optimal climbs have been among the applications of the gradient methods employed by Bryson (54) and Kelley (55). These investigators have also given some consideration to the synthesis problem. Friedland (56) has considered the

synthesis problem and has described in general terms the structure of a class of optimal controllers.

At this point it is useful to look back for a moment and take stock of what has been said. Optimal control and optimal path problems can be formulated and solved for a number of special applications. In general, though, these are idealized cases. That is, given a typical problem as posed in the preceding sections, it is usually possible to find a solution to the two-point boundary problem, if one exists, by some one of a variety of means. Nevertheless, this is not enough to constitute a synthesis procedure, although it is a step in the right direction.

A closed loop optimal controller must be able to determine the optimal control law at each instant of time. Except for some special cases, i.e. time invariant linear processes with quadratic error and quadratic fuel cost, the two-point boundary problem requires iterative solution in one form or another.

The determination of the optimal control may be accomplished by looking up stored results of previous calculations, by actual solution of the two-point boundary problem with fast-time-scale computers, or by a combination of these methods. For example, controllers have been described by Bryson (57) and Kelley (43) that would use stored nominal optimal paths together with a fast computer to determine neighboring optimal paths. Alternatively, nominal paths can be stored and the optimal control for an "initial" point not on a stored path can be approximated by an interpolation scheme.

There are many alternatives available and only a few seem to have received any serious study. For the most part, these studies are idealized by excluding noise, random disturbances, and component imperfections. There are exceptions however. Studies of satellite attitude controllers have included these effects and considered the behavior of the systems, (usually relay type), in the small signal region with component imperfections present such as dead time and hysteresis (58). The effect of these real characteristics on limit cycle size and fuel consumption are especially important in those applications involving a service lifetime that is long compared to system time constants. An optimal system designed, for example to reduce large errors in least time, should be expected to spend most of its life in the small signal region where the effects of imperfections and noise are important.

Returning again to the synthesis question, it has been found worthwhile to examine some results of studies of non-optimal controllers that use techniques similar in principle to those required for optimal synthesis. These are the predictive controllers and the terminal controllers. Some results of automatic predictive controllers are available. (Chesnut and Sollicito (59) Steeg and Morris (60).) The applications studied include reentry guidance, satellite rendezvous and airplane automatic landings. In addition, studies of predictive controllers with human pilots in the

loop have also been made (NASA) (61,62). These have also been idealized problems as various simplifications or approximations have been made to facilitate the studies.

These predictive controllers have been made to work at least in ideal simulation and there seems to be no fundamental reason why similar controllers based on an optimal control logic could not be made to work under the same ideal conditions. The important question is how to design an optimal controller to operate under real conditions and very little is known about this at present.

The next section will describe some results of optimal synthesis studies for linear systems.

SYNTHESIS OF OPTIMAL CONTROLLERS FOR LINEAR SYSTEMS

A. Preliminary remarks. The discussion in the preceding section defined an optimal controller as a closed-loop feedback system that repetitively solves the two-point boundary problem. It was pointed out that there are several ways to accomplish this depending on the amount of pre-computation and storage allowed. The studies reported on here use the minimum amount of precomputation. The minimum time performance criterion was selected for these synthesis studies for several reasons. The main reason being the belief that a synthesis procedure could be achieved for this case. The reachable sets are generally easy to obtain, yet the problems are not trivial. The next step was to try to obtain a computational solution to the time optimal control problem. The closed-loop system that was envisioned would frequently sample the measurable system output variables and estimate the system state point. On the basis of this estimate, the controller rapidly obtains the optimal control manipulation by finding the proper state for the adjoint system. Section B contains a description of the studies directed towards developing a useable computational procedure. The incorporation of this procedure into a feedback controller is described in section C. Some applications to simple problems of flight mechanics were used to provide a framework for the synthesis studies.

B. Solution of the time optimal control problem. It is already well known that the solution for the classical time optimal control problem is of bang-bang type and the control law can be given in terms of the response of the system adjoint to the given system. It is possible to obtain these results by several different procedures and a review of the field is contained in reference (27).

In spite of the fact that the optimal control law has been known for a long time, very few optimal controllers have been built for systems of order higher than two. The principal drawback has been the difficulty in solving the two-point boundary problem, or what is the same thing, finding the optimal initial state of the adjoint. This section is intended primarily to show how previously published techniques can be combined to successfully solve this

problem. In addition, some experimental results are presented that provide verification of the convergence of the proposed procedure for a class of linear time invariant systems. The problem can be stated as follows: Given a linear system of the form

$$\dot{X} = A(t)X + B(t)u \tag{1}$$

where

X is an n-vector
A is n x n matrix
B is n x r matrix
u is r vector,

It is desired to find the control law u^o that takes the system from initial state $X(0)$ to terminal state X_f (or catches input signal $\xi(t)$ in least time, T^o.

It is well known that the optimal control law is given by the expression

$$u^o = \text{sgn}\ (\eta \cdot X^{-1}B) \tag{2}$$

where $X^{-1}(t)$ satisfies the adjoint differential equation

$$(\dot{X}^{-1}) = -(X^{-1})A$$

The solution of the equation (1) can be written as

$$x(t) = X(t)\left[\, x(o) + \int_o^T X^{-1}(\tau)\ B(\tau)u(\tau)d\tau \,\right] \tag{3}$$

After substituting the terminal condition $x_i(t) = \xi_i(t)$ and equation (2) the following expression

$$-x(o) = X^{-1}(t)\xi(t) + \int_o^T X^{-1}(\tau)\ B(\tau)\text{sgn}\left(\eta X^{-1}(\tau)B(\tau)\right)d\tau \tag{4}$$

is obtained.

This equation in effect is a mapping of vectors η into vectors

$$Z = -\int_o^t \ldots\, d\tau \qquad \xi(t) = 0\ .$$

The problem is to find an η that maps Z into $-x(o)$. This was recognized in a somewhat different form by Krasovskii (63). However, Neustadt provided a key to a steepest descent procedure by means of an elegant result (31).

Before proceeding directly to the main idea, it is desirable to insert a brief discussion of some geometrical aspects of the problem. The idea of optimal isochrones, or the boundaries of the reachable sets, plays an important role in the development of optimal control theory. To each point in the x-phase space, there

is assigned a value of the optimal time required to reach the target. Through each point there passes a portion of a surface of constant optimal time to reach the target (Figure 2). The normal to this surface defines the direction of the optimal η vector. These surfaces need not be smooth, in fact, there may be corners, sharp edges and points. Therefore, the gradient may not exist at certain points or along certain curves or hypersurfaces. Figure 3 shows a typical set of isochrones for a second-order system.

The hyperplane passing through the point $-x(o)$ and defined by the η^o vector is a support plane for the convex set formed by the optimal isochrones. Therefore, for all unit vectors η, the inner product $(\eta, -x(o)) < (\eta^o, -x(o))$ or η^o maximizes $(\eta, -x(o))$. Neustadt (31) suggests the following iteration procedure. Select a trial value of η. This determines a trial hyperplane at $-(x(o))$ in the Z-space. Examine the locus of points mapped by η into Z using t as a parameter, and select the values of Z and t corresponding to the point where the locus pierces the hyperplane. The geometrical ideas relating the optimal control problem to the notion of expanding convex sets are believed to have been first explored by Bellman, Glicksberg, and Gross in a study of the bang-bang problem (20).

Figure 4 illustrates this procedure. The point of intersection of the Z trajectory and the hyperplane will not generally coincide with the desired target; $-x_1(o)$. If this point does correspond to $-x_1(o)$ then the optimal initial conditions of the adjoint have been found and are equal to the components of the normal to the hyperplane. A series of corrections to the η-vector is usually necessary to find the optimal initial conditions. A steepest descent procedure is effective. The changes in $\overline{\eta}(\eta^{\nu+1} = \eta^{\nu} + \delta\eta)$ are given by a result of Neustadt, namely $\delta\overline{\eta} = K[x(o) + Z(\eta,t)]$. The correction to η^{ν} should lie along the "error vector" $(\overline{x}(o) + \overline{Z}(\eta,t))$ in the hyperplane determined by η^{ν}.

A starting value for the $\overline{\eta}$ vector must be obtained. The results of previous studies have shown that there is a relation between the optimal isochrones and Liapunov functions (27). This observation leads to a means of approximating the optimal isochrones by quadratic forms. These quadratic forms are then used to obtain the starting values for the iteration. These studies used

$$\eta(o) = - \frac{x(o)}{|| x(o) ||} .$$

The initial studies were based on simple systems. Using the ARAP analog-digital computer, a second order undamped oscillatory system with one control force was simulated and regulated in a time optimal fashion.

The equations are:

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

$$\dot{x}_2 = -\omega^2 x_1 + u \tag{1'}$$

$$u^o = \text{sgn}\left[-\eta_1 \frac{\sin \omega t}{\omega} + \eta_2 \cos \omega t\right] \tag{2'}$$

$$Z_1(t) = -\int_o^t \frac{1}{\omega} \sin \omega\tau[u]d\tau \tag{4a'}$$

$$Z_2(t) = \int_o^t \cos \omega\tau[u]d\tau \tag{4b'}$$

$$\eta_1[Z_1(t) + x_1(o)] + \eta_2[Z_2(t) + x_2(o)] \leq 0 \tag{5}$$

$$\eta_{i_n} = \frac{\eta_{i_{n-1}} - K_1[Z_i(T) + x_i(o)] - K_2[Z_i(T) + x_i(o)]^2}{\sum_{i=1}^{2} \left\| \eta_{i_{n-1}} - K_1[Z_i(T) + x_i(o)] - K_2[Z_i(T) + x_i(o)]^2 \right\|} \tag{6a}$$

The analog computer is started with a first guess of the adjoint system initial conditions η_1, η_2

$$\eta_i = -\frac{x_i(o)}{\| x_i(o) \|}$$

This initial guess is obtained from a Liapunov function. The computer then runs until $\eta_i[Z_i(t) + x_i(o)]$ becomes positive. The analog machine is then transferred into the hold mode; Z_1, Z_2 sampled; and a new initial condition vector η_1, η_2 found for the adjoint system. This cycle is repeated until the terminal point in Z-space is within distance ϵ of $-x_i(o)$.

For the examples shown here, the system parameters are: $\omega^2 = .1$ and $F = \pm .5$, giving steady state rest points of $x_1 = \pm 5$. In all tests, the effects of the gain constants K_1, K_2 in (6a) on the convergence characteristics were noted. The general trend observed was that increasing the K's resulted in fewer iterations required per solution with five iterations being the minimum number. Further increases then resulted in large and undesirable oscillations about the correct solution. The values of the gain constants at which the cycling occurs depends upon the system in question and its initial conditions. For the time optimal regulation of the harmonic oscillator values of $K_1 = 1$; $K_2 = 1$ were satisfactory under all conditions.

Using these nominal values, the iterations shown in Figures 5 and 6 were obtained. Each of the trajectories shown in the Z-phase space represent one of the iterations in the search for the correct initial conditions for the adjoint system. The trajectory of the physical system in x-phase space corresponding to the correct η

vector is also shown. Figure 6 illustrates the same data in time history form.

In addition to oscillators, the regulation of double integrators was also investigated ($\omega^2 = o$). A sample case is shown in Figure 7. The hyperplane is almost parallel to the trajectory of the system in the region of the solution. Under these conditions, a small angular change in the adjoint initial vector, (producing an equal angular change in the trial tangent line) will result in a large displacement of the point of intersection. It is necessary to reduce the steepest descent gains if oscillations are to be avoided. In the case shown, $K_1 = .1$, $K_2 = .1$ were used in the region sufficiently near to the correct solution (use of these gains throughout the problem results in an excessive number of iterations). It will generally be necessary to program controller gain variations.

In addition to regulation (returning the system to its null point) it was decided to investigate the capabilities of this technique in following command signals. Referring back to Equation 4, it can be seen that the ability to follow variable input signals is obtained by retaining the terms $\xi_i(t)$; i.e., the target point coordinates, $\xi_i(t)$, are no longer located at the origin. For the second-order example, the equations used by the controller are

$$\lambda_1(t) = x_1(o) - \int_o^t \frac{1}{\omega} \sin \omega\tau[u]d\tau - \xi_1(t) \cos \omega t + \xi_2(t) \frac{\sin \omega t}{\omega} \tag{7a}$$

$$\lambda_2(t) = x_2(o) + \int_o^t \cos \omega\tau[u]d\tau - \xi_1(t)\omega \sin \omega t - \xi_2(t) \cos \omega t \tag{7b}$$

where $\lambda_i(t) = x_i(o) + Z_i(t)$ terms resulting from $\xi_i(t) \neq o$.

$$\eta_1[\lambda_1(T)] + \eta_2[\lambda_2(T)] \leq 0 \quad \text{---} \tag{5'}$$

$$\eta_{i_n} = \frac{\eta_{i_{n-1}} - K_1[\lambda_i(T)] - K_2[\lambda_i(T)]^2}{\sum || \eta_{i_{n-1}} - K_1[\lambda_i(T)] - K_2[\lambda_i(T)]^2 ||} \tag{8}$$

These equations reduce to those given previously by setting $\xi_1(t) = \xi_2(t) = 0$ and transposing the $x_i(0)$ terms to the left hand side of the equations.

The two forms of input signals investigated in this portion of the study were:

(a) Arbitrary stationary points in the phase space, i.e. going to a prescribed position and attaining a prescribed velocity

at that point.

(b) Step and ramp signals. In phase space, the target starts with initial horizontal and vertical displacements from the origin and moves horizontally at a rate proportional to its vertical displacement. Figure 8 shows the iteration process for the first type of input.

It is important to remember that Neustadt's procedure is based on finding the maximum time for which the condition $\eta \cdot [Z(T) + x(0)] = 0$ can occur. The error vector, $[Z(t) + x(0)]$ need not decrease at each step of the steepest descent and may become very large even as the time T approaches the optimal time. In order to speed up the convergence of the iterations, a change is made in the procedure if successive changes in T are observed to be growing small without reduction in the size of $Z(T) + x(0)$. There are several other ways to use the Z-equations and two of these have been employed during the terminal phase of the iterations. Gamkrelidze suggests a method (also described by Krasovskii) for a steepest descent procedure, keeping T fixed and varying η to make $Z(T)$ parallel to $x(0)$. The final time T is to be increased from something small until $Z(T) = -x(0)$. In this method, as in the previous one, there is still the question of determining the maximum T (i.e. distinguishing local minima) in the case of moving targets.

It is also possible to seek the value of η that minimizes the error at the time $\eta \cdot [Z(T) + x(0)] = 0$. This is a modification of a search method investigated by Paiewonsky (27). The latter method usually requires an experimental determination of the gradient in practice, although a closed form solution is sometimes available. The results obtained to date on these techniques are not sufficient to allow a complete comparison with Neustadt's method regarding speed of convergence, practical difficulties, etc. It is hoped that it will be possible to do this soon. It is possible, however, to state that the problem of the vanishing of the gradient of the error (as a function of the initial η) as described by Paiewonsky in reference (27) can occur for the latter method. This does not occur with Neustadt's method as the gradient there is always given by the error vector.

C. Applications. The applicability of this technique to problems of flight mechanics depends upon a linearization of some kind. Many possibilities suggest themselves; one is to find a problem that can be linearized from the start. If this cannot be done, it may be possible to apply a method of successive approximations to the equations solving a succession of linear problems. The following example was prepared in order to study the properties of a closed loop optimal controller using a repetitive computer. Consider first the problem of determining the optimal steering program for the upper stages of a booster rocket. If the rocket burns continuously then the minimum fuel problem is a minimum time problem. The gravitational field will be assumed to be uniform in this example. The equa-

tions of motion are

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= A \cos u \\ \dot{x}_3 &= x_4 \\ \dot{x}_4 &= A \sin u - g \end{aligned} \tag{9}$$

The convexity of the reachable sets can be verified directly. The adjoint equations are:

$$\begin{aligned} \dot{P}_1 &= 0 \qquad \dot{P}_2 = - P_1 \\ \dot{P}_3 &= 0 \qquad \dot{P}_4 = - P_3 \end{aligned} \tag{10}$$

The optimal control is the well-known bi-linear tangent law:

$$\text{Tan } u^{\circ} = \frac{P_4(t)}{P_2(t)} = \frac{P_4(0) - P_3(0)t}{P_2(0) - P_1(0)t}$$

In this example, the equations for the Z's are

$$\begin{aligned} \dot{Z}_1 &= - t A \cos u \\ \dot{Z}_2 &= A \cos u \\ \dot{Z}_3 &= - t(A \sin u - g) \\ \dot{Z}_4 &= A \sin u - g \end{aligned} \tag{11}$$

The particular numerical example chosen is the optimal steering of a nuclear powered upper stage rocket. The initial conditions are

$$\begin{aligned} V_X(0) &= 10{,}330 \text{ ft/sec} \\ V_Y(0) &= 5{,}500 \text{ ft/sec} \\ y(0) &= 20 \text{ miles} \end{aligned}$$

and the desired terminal conditions are:

$$\begin{aligned} V_X &= 25{,}300 \text{ ft/sec} \\ V_Y &= 0 \text{ ft/sec} \\ y &= 200 \text{ miles} \end{aligned}$$

The horizontal range was not prescribed. These equations were set up on the ARAP analog-digital computer at first for the

special case of $A = F/m(t) = 5$. The objectives of the initial study were to determine the sensitivity of the convergence of the iteration procedure to computational errors and to determine the number of iterations required as a function of the initial starting value of the adjoint. These studies showed that convergence of the iterations depended to a great extent on the accuracy with which the stopping time could be determined. For systems of order higher than two, $Z(T,\eta)$ proved to be a highly sensitive function of η. The maximum tolerable error in determining the time at which $\eta \cdot (Z(T,\eta) - x(o))$ vanished was of the order of 10^{-4} seconds. This is illustrated by Table 1, which represents a typical case.

Components of η			$F(\eta)$	Components of $Z(F(\eta),\eta)-x(o)$		
.54580027	.01316405	1.0729829	148.56863	.830	22,700	-7.373
.54580027	.01316405	1.0729839	148.56906	.397	52,705	-16.369
.54580027	.01316405	1.0734830	148.56824	1.259	-7,410	1.640
.54580027	.01314405	1.0734830	148.56763	10.711	-121,610	31.768

TABLE I

This extreme sensitivity makes it difficult to reduce the error, $||Z(T,\eta) - x(o)||$ when the time T is close to the optimal time. Two steps were taken to overcome this difficulty. First, a special program for the step size was developed. Second, when successive values of stopping time, T, differed by a preselected amount of different computational scheme was used to obtain reduced values of the error $||Z(T,\eta) - x(o)||$ and precise estimates of η^*. The details of these computational studies will be contained in a forthcoming ASD report.

The next step was to program a computational model of closed-loop optimal controller for simulation on an IBM 7090. This is shown schematically in fig. 9. The controller samples the output of the system and determines the value of the control, $u^o(t)$ on the basis of the noisy measurements. The possibility of filtering the noisy measurements was included.

It takes a finite amount of time for the computer to determine the control and a variable length delay was incorporated to study the effect of this time lag. Provision was also made for a calculation in the controller to simulate compensation for the delay. The control computer requires an estimate of the system parameters, in this case the initial mass $m(o)$, the propellant mass flow rate β, and the thrust F or the effective exhaust velocity c. The control computer does not

know the exact values of these parameters and a study was made to determine the effect of uncertainties in system parameters on overall performance. The mass flow rate, β, was allowed to have random variations corresponding to rough burning, or in extreme cases, a "chamber out" condition. No adaptive features were included in the controller although it is clear that the system parameters needed are capable of being estimated in flight.

In order to reduce the number of iterations required during the computation of the control, a "nominal" trajectory was introduced. The nominal trajectory used was the optimal path from the initial point to the terminal point. That is, once the optimal initial values for the adjoint are found by the computer, after the very first sample period, the nominal values for the state variables and adjoint can be obtained by a faster than real time integration. These values of the adjoint are stored and used as starting values for the subsequent iterations at later sampling periods.

This work is not yet complete. However, some preliminary results of the computer studies are available. The objective in the perturbation studies is to observe the behavior of the iteration scheme as the initial conditions are perturbed away from the starting values. Two types of investigations were made. In a typical run of the first series, the initial point in phase space is fixed, the starting value of the adjoint is varied and the number of iterations required to converge to the optimal value of the adjoint is obtained. The second series of runs is similar but the starting value of the adjoint is fixed and the initial state of the system is varied. The number of iterations required to find the optimal state of the adjoint is obtained here also.

A series of runs to observe the closed loop system behavior in the presence of noise and parameter uncertainties is also being carried out at the present time as part of the computer studies. The results of these studies will be reported in detail at a later date.

REFERENCES

1. Blackman, N. M., "On the Effect of Noise in a Non-Linear Control System," paper presented at the IFAC Conference, Moscow, 1961.

2. Hopkin, A. M. and Wang, P. K. C. "A Relay Type Feedback Control System Design for Random Inputs," Trans. AIEE, Vol. 78, Part II, p. 228, Sept., 1959.

3. Barret, J. F., "Application of Kolmogorov's Equation to Randomly Disturbed Automatic Control Systems," paper presented at the IFAC Conference, Moscow, 1960.

4. Selected Papers on Noise and Stochastic Processes, Edited by Nelson Wax, Dover, 1954.

5. Tsien, H. S. Engineering Cybernetics, McGraw Hill, 1954.

6. Laning, J. H., Jr. and Battin, R. H. Random Processes in Automatic Control, McGraw Hill, 1956.

7. Peterson, E. L., Statistical Analysis and Optimization of Systems, John Wiley and Sons, Inc., 1961.

8. Aoki, Stochastic-Time Optimal-Control Systems, Applications and Industry, May 1961, 60-1018.

9. Bellman, R., Dynamic Programming and Stochastic Control Processes, Information and Control, New York, N. Y., Vol. 1, No. 3, 1958, pp. 228-39.

10. Adorno, D. S., Studies in the Asymptotic Theory of Control Systems; I. Stochastic and Deterministic N-Stage Processes, Jet Propulsion Laboratory, Technical Report No. 32-21, May, 1960.

11. Zadeh, L. and Eaton, J. H., "Optimal Pursuit Strategies in Discrete-State Probabilistic Systems," ASME Paper 61-JAC-11.

12. Graham, Dunstan, and R. C. Lathrop, "The Influence of Time Scale and Gain on Criteria for Servomechanism Performance," Trans. AIEE, Vol. 73, Part II, 1954, p. 153.

13. Performance Criteria for Linear Constant Coefficient Systems with Deterministic Inputs, ASD-TR-61-501, Wright-Patterson Air Force Base, Ohio, February 1962.

14. Bliss, Gilbert Ames, Lectures on the Calculus of Variations, University of Chicago Press, p. 296, 1946.

REFERENCES (Continued)

15. Hestenes, Magnus R., A General Problem in the Calculus of Variations with Applications to Paths of Least Time, Rand RM 100, 1 March 1949, Rand Corp., Santa Monica, California.

16. Kalman, R. E., The Theory of Optimal Control and the Calculus of Variations, RIAS Tech. Rept. 61-3.

17. Markus, L., and E. B. Lee, "On the Existence of Optimal Controls," J. Basic Engr. (Trans. ASME, Series D) Vol. 84, pp. 13-20, March, 1961.

18. Breakwell, John V., The Optimization of Trajectories, AL-2706 NAA, Adv. Engrg. p. VI+ 37+A1 - A6, 29 August 1957; ASTIA No. AD-209 092; Journal Soc. Indust. and Appl. Math., Vol 7 No. 2, June 1959, p. 215-247; Chapter 12, General Research in Flight Science, January 1959 - January 1960, Vol III; and Flight Dynamics and Space Mechanics, LMSD - 288139, January 1960.

19. Berkovitz, L. D., "Variational Methods in Problems of Control and Programming," Journal of Mathematical Analysis and Applications, Vol. 3, pp. 145-169, August 1961.

20. Bellman, R., Glicksberg, I., and Gross, O., On the Bang-Bang Control Problem, Quart. Appl. Math. 14 (1956) 11-18.

21. Dreyfus, S., Dynamic Programming and the Calculus of Variations, Rand Rept. P-1464, Feb. 1960.

22. LaSalle, J. P., The Time Optimal Control Problem, Contributions to the Theory of Nonlinear Oscillations V, Princeton Univ. Press. 1960.

23. Pontriagin, L. S., Boltjanski, V. G., Gamkrelidze, R. V., "Optimal Processes," Proc. 1st IFAC Conf., Moscow, 1959.

24. Kipiniak, W., "A Variational Approach to Dynamic Optimization and Control System Synthesis," May 1961. MIT Monograph.

25. Chang, S. S. L., "Digitized Maximum Principle," Proceedings of IRE, Vol. 48, pp. 2030-2031, Dec. 1960.

26. Desoer, C. A., "The Bang-Bang Servo Problem Treated by Variational Technique," Information and Control, Vol. 2, No. 4, pp. 333-348, December 1959.

27. Paiewonsky, B. H., A Study of Time Optimal Control, ARAP Report No. 33, 1961, Also: Contributions to Differential Equations, Vol III, Academic Press, New York, 1962.

28. Halkin, H., The Principle of Optimal Evolution, Contributions to the Theory of Differential Equations, Vol. III, Academic Press, New York, 1962.

29. Roxin, E., A Geometric Interpretation of Pontriagin's Maximum Principle, RIAS TR 61-15, Dec. 1961.

30. Anderson, C. A., "Switching Criteria for Certain Contractor Servos with Arbitrary Inputs," Univ. of Michigan, PH.D. Thesis, 1958.

31. Neustadt, L. W., "Synthesizing Time Optimal Control Systems," Journal Math. Analysis and Applications, Vol. 1, No. 3, 4, December 1960.

32. Kelley, Henry J., "Gradient Theory of Optimal Flight Paths," ARS Paper 1230-60, 9-12 May 1960, New York, N. Y., presented at the ARS Semi-Annual Meeting, Los Angeles, Calif., 9-12 May 1960. Also: J. ARS, October 1960, p. 947-954 (Vol. 30, No. 10).

33. Bryson, A. E. and W. F. Denham, "A Steepest Descent Procedure for Problems in Optimal Control and Programming," Journal of Applied Mechanics ASME Series E, to be published.

34. Swanlund, G., "Analysis Techniques to Determine Guidance Computation Requirements for Space Vehicles," ARS Journal, May 1962, p. 755.

35. Ho, Y. C., "A Successive Approximation Technique for Optimal Control Systems Subject to Input Saturation," Journal of Basic Engineering (Trans. ASME, Series D) Vol. 84, pp. 33-40, March 1962.

36. Dreyfus, S., Variational Problems with Inequality Constraints, RAND P-2357, July 1961, rev. Aug. 1961.

37. Meer, A., Synthesis of an Optimum Nonlinear Controller as a Spatial Analog, S. M. Thesis, MIT Electrical Engineering Department, Sept. 1961.

38. Desoer, C. A. and J. Wing., Minimal Time Regulator Problems for Linear Sampled-Data Systems (General Theory), California, Univ. Berkeley, Electronics Res. Lab., Sci. Rep. 6, Inst. Eng. Res., Ser. 60/I-346 (AFCRL 149, AFOSR 424), February 14, 1961, 40 p. 18 refs.

39. Tou, J. T. and Boonyok Vadhanaphuti, "Optimum Control of Nonlinear Discrete-Data Systems," (AIEE, Winter Gen. Meeting, N. Y., Jan. 29-Feb. 3, 1961.) Applications and Industry, Sept. 1961, p. 166-170; Discussion, p. 170, 171. 12 refs.

40. Kulikowski, R., "On the Synthesis of Optimum Sampled-Data Control Systems," Bull. Pol. Acad. Scie., Vol. VIII, No. 11-12, 1960.

41. Neustadt, L. W., "Discrete Time Optimal Control Systems," OSR-RIAS Symp., Colorado Springs, July-Aug., 1961.

42. Bryson, A. E., Denham, W. F., Carroll, F. J., and Mikami, K., "Determination of Lift or Drag Programs to Minimize Re-Entry Heating," Jour. Aerospace Sci., April 1962, Vol. 29, No. 4.

43. Kelley, H. J., "Guidance Theory and Extremal Fields." Proceedings Nat. Aerospace Electronics Convention, Dayton, Ohio, May 1962.

44. Dow, P. D., Fields, D. P., and Scammell, F. H., Automatic Re-Entry Guidance at Escape Velocity, Progress in Astronautics and Rocketry, Vol. 8 "Guidance and Controls," Academic Press, N. Y., 1962.

45. Leitmann, George, On a Class of Variational Problems in Rocket Flight, Rept. No. LMSD-5067, 29 p., Sept., 1958, Lockheed Missile Systems Div., Sunnyvale, Calif. Also: Journal Aero/Space Sci., Vol. 26, No. 9, p. 586-591, Sept. 1959.

46. Behrbohm, Hermann, Optimal Trajectories in the Horizontal Plane, SAAB TN 33, 25 p., 15 March 1955, SAAB Aircraft Corp., Linkoping, Sweden; AD-104 957.

47. Miele, Angelo, Mathematical Theory of the Optimum Trajectories of a Rocket--Final Report, Purdue University, School Aero. Engng., A-58-10, 35 p. Nov. 1958, AD-206 361. AFOSR TR 58-154.

48. Melbourne, W. G., Richardson, D. E., and Sauer, C. G., Interplanetary Trajectory Optimization with Power-Limited Propuslion Systems, GPL Tech. Rept. 32-173, Feb. 1962.

49. Edelbaum, T., Optimum Low Thrust Transfer Between Circular and Elliptic Orbits, United Aircraft Corp. Res. Labs. Rept. A-1100 58-1, Jan. 1962.

50. Lawden, Derek F., "Minimal Trajectories," Jour. British Interplanetary Soc., Vol. 9, No. 4, p. 179-186, July 1950.

51. Bellman, R. and S. Dreyfus, "An Application of Dynamic Programming to the Determination of Optimal Satellite Trajectories," Jour. British Interplanetary Soc., No. 3-4, Vol. 17, May-Aug. 1959.

52. Cartaino, John F. and S. E. Dreyfus, Applications of Dynamic Programming to the Airplane Minimum Time-to-Climb Problem, RAND RM 1710, 17 p., 22 March 1956. Also: RAND Corp., Rept. P-834, p. 74-77, June 1957; AD-101 000.

53. Smith, F. T., "Optimization of Multistage Orbit Transfer Processes by Dynamic Programming," ARS Journal, Vol. 31, Nov. 1961, p. 1553-1559, 13 refs.

54. Bryson, A. E. and W. F. Denham, "A Guidance Scheme for Supercircular Re-Entry of a Lifting Vehicle," ARS Journal, June 1962.

55. Kelley, H. J., Falco, M., and Ball, D. J., Air Vehicle Flight Path Optimization, AFOSR TR/DRA-62-4, Holloman AFB, New Mexico, Feb. 1962.

56. Friedland, B., "The Structure of Optimum Control Systems," Journal of Basic Engineering, (Trans. ASME Series D), Vol. 84, pp. 1-12, March 1962.

57. Breakwell, John V. and A. E. Bryson, "Neighboring-Optimum Terminal Control for Multivariable Non-Linear Systems." (To appear in SIAM publication.)

58. Single Axis Attitude Regulation of Extra Atmospheric Vehicles, ASD TR 61-129, WPAFB, Ohio, Feb. 1962.

59. Chestnut, Sollecito, Troutman, "Predictive-Control Application 61-12," Applications and Industry, July 1961.

60. Morris, R. V. and C. W. Steeg, "Multicondition Terminal Control Systems for Aircraft," Jour. Aerospace Sci., p. 712, Sept. 1960.

61. Young, J. W., A Method for Longitudinal and Lateral Range Control for a High-Drag Low-Lift Vehicle Entering the Atmosphere of a Rotating Earth, U. S., NASA TN D-954, Sept. 1961, 26 p.

62. Foudriat, E. C., and R. C. Wingrove, Guidance and Control During Direct-Descent Parabolic Reentry, NASA-Industry Apollo Tech. Conf., Wash., D. C., July 18-20, 1961. U. S., NASA TN D-979, Nov 1961. 24 p. 17 refs.

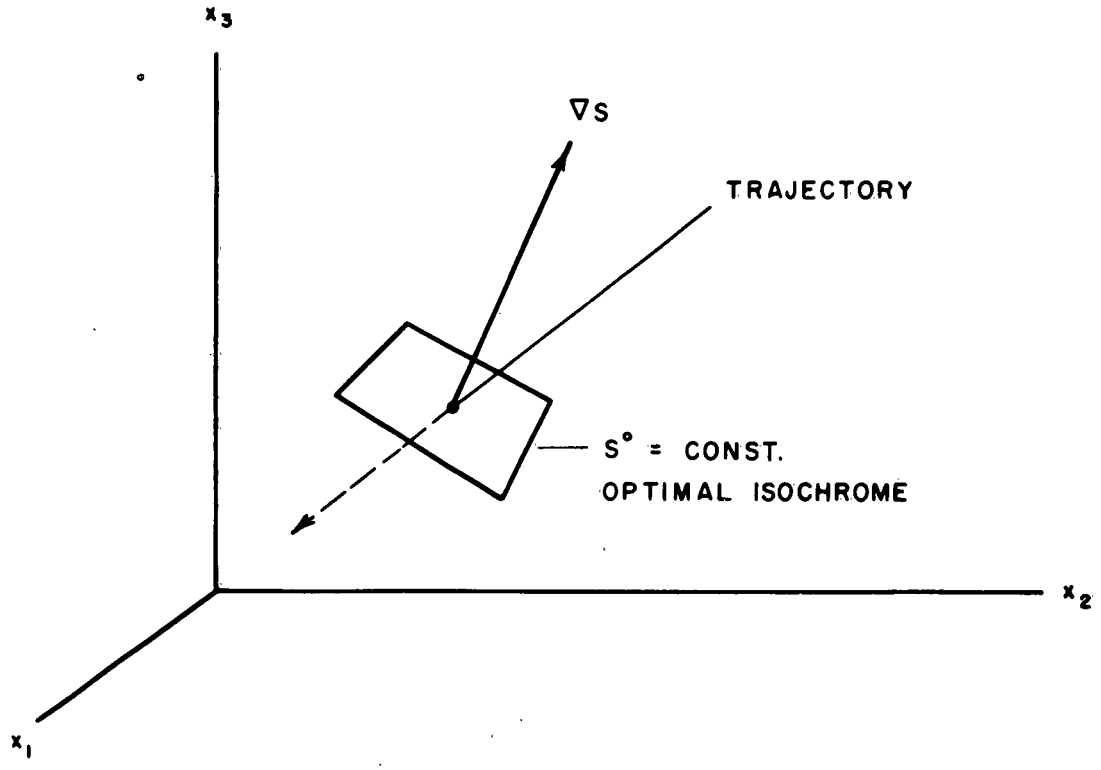


Figure 2. Surface of Constant Optimal Time

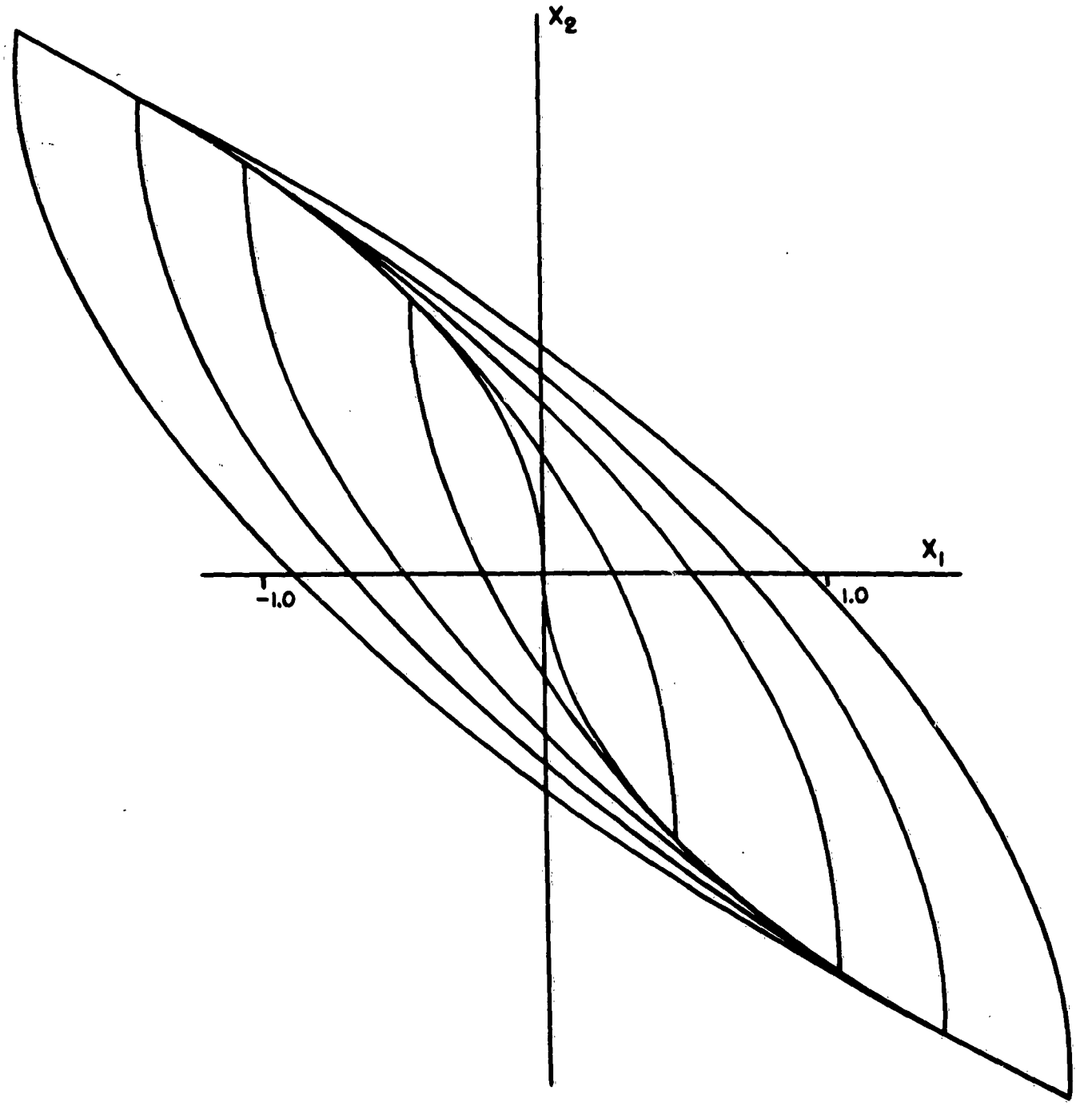


Figure 3. Optimal Isochrones, $\dot{x}_1 = x_2$
$\dot{x}_2 = u(t)$

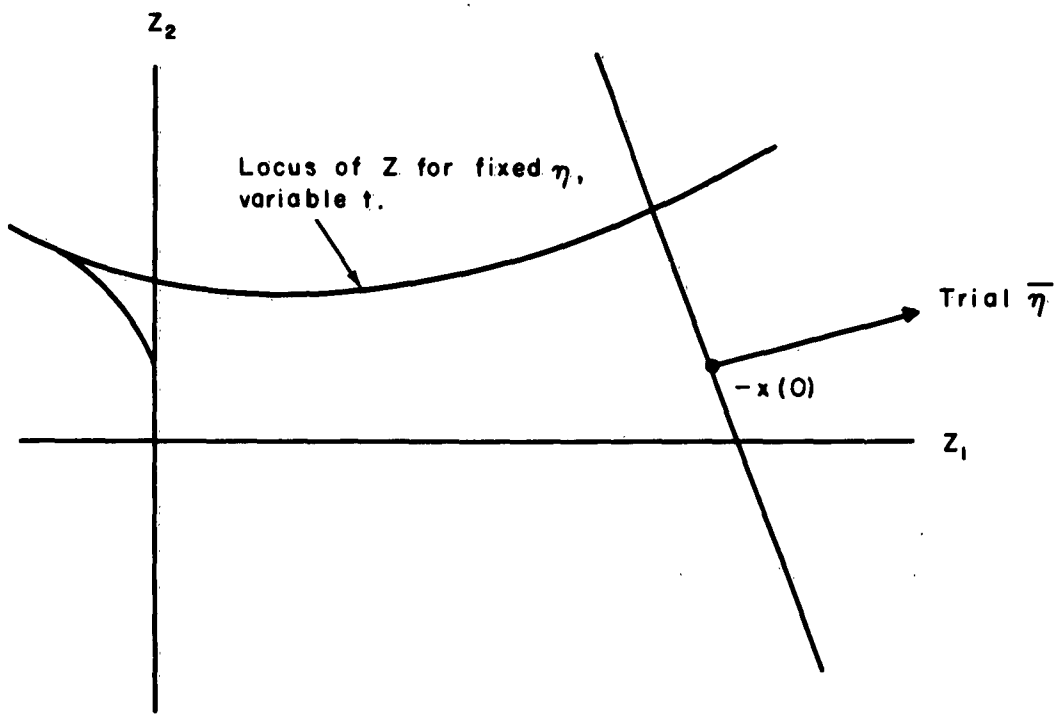


Figure 4. Optimal Hyperplane

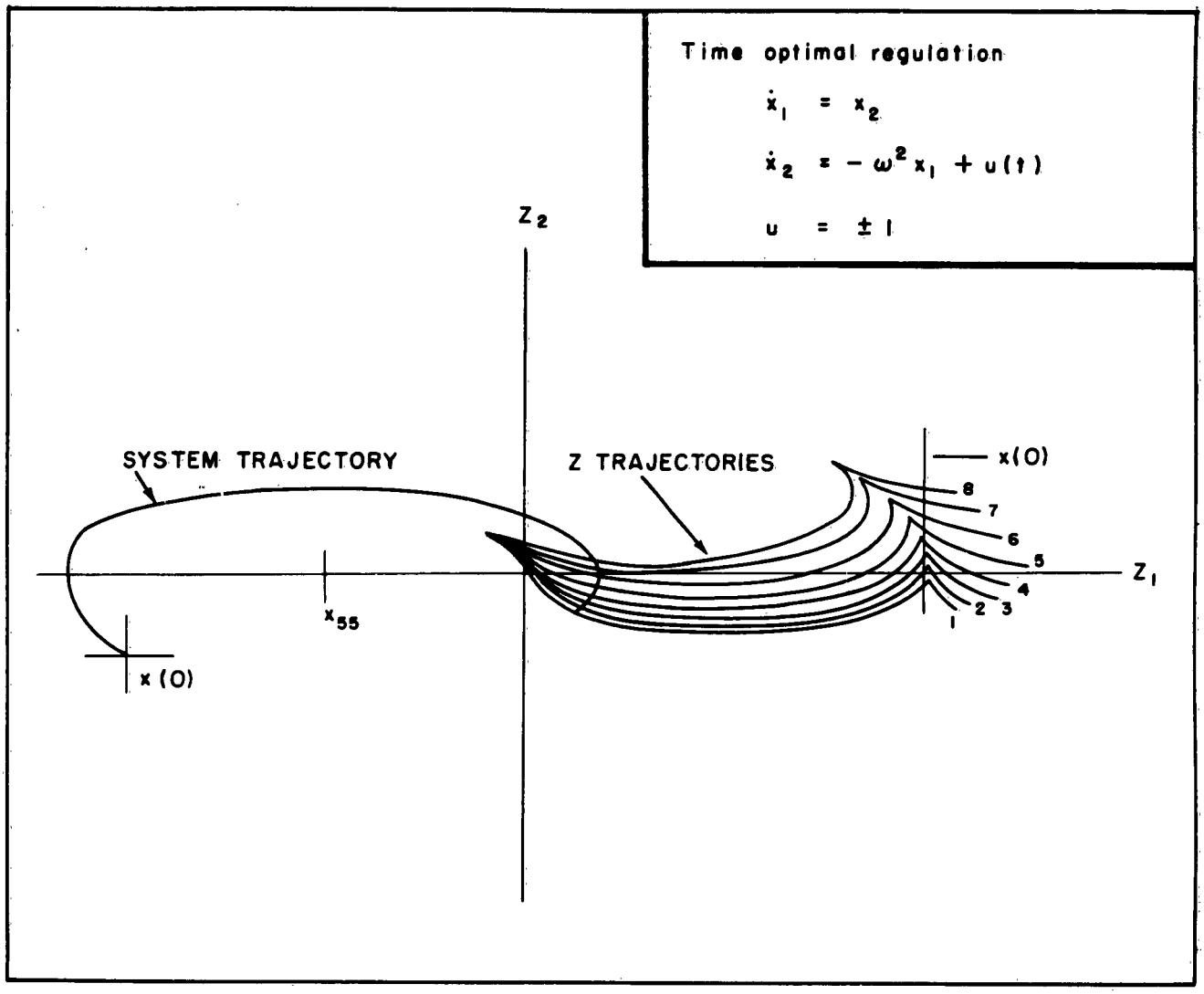


Figure 5. Phase Diagram - Time Optimal Regulation

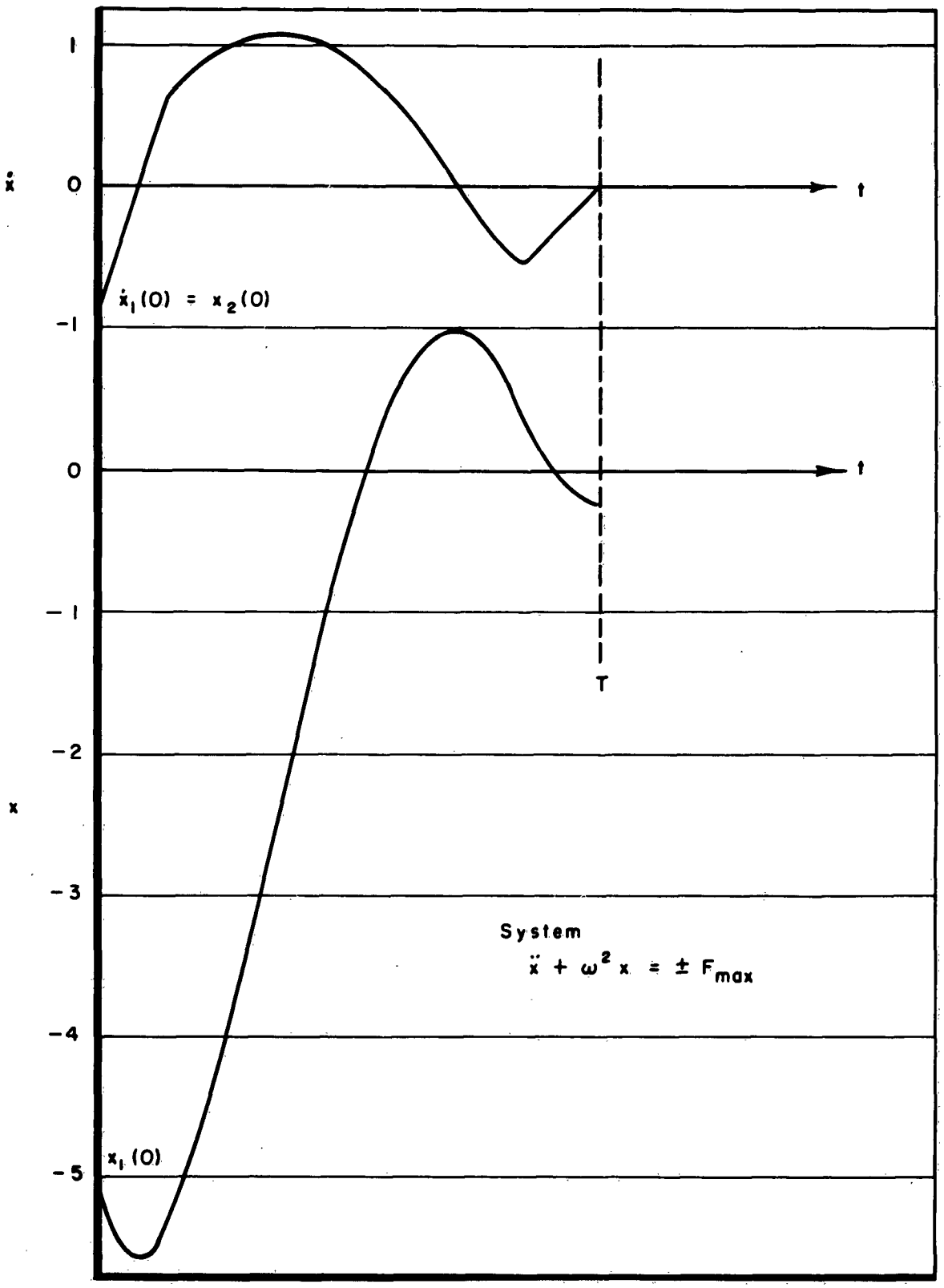


Figure 6. Time History of Optimal Regulation

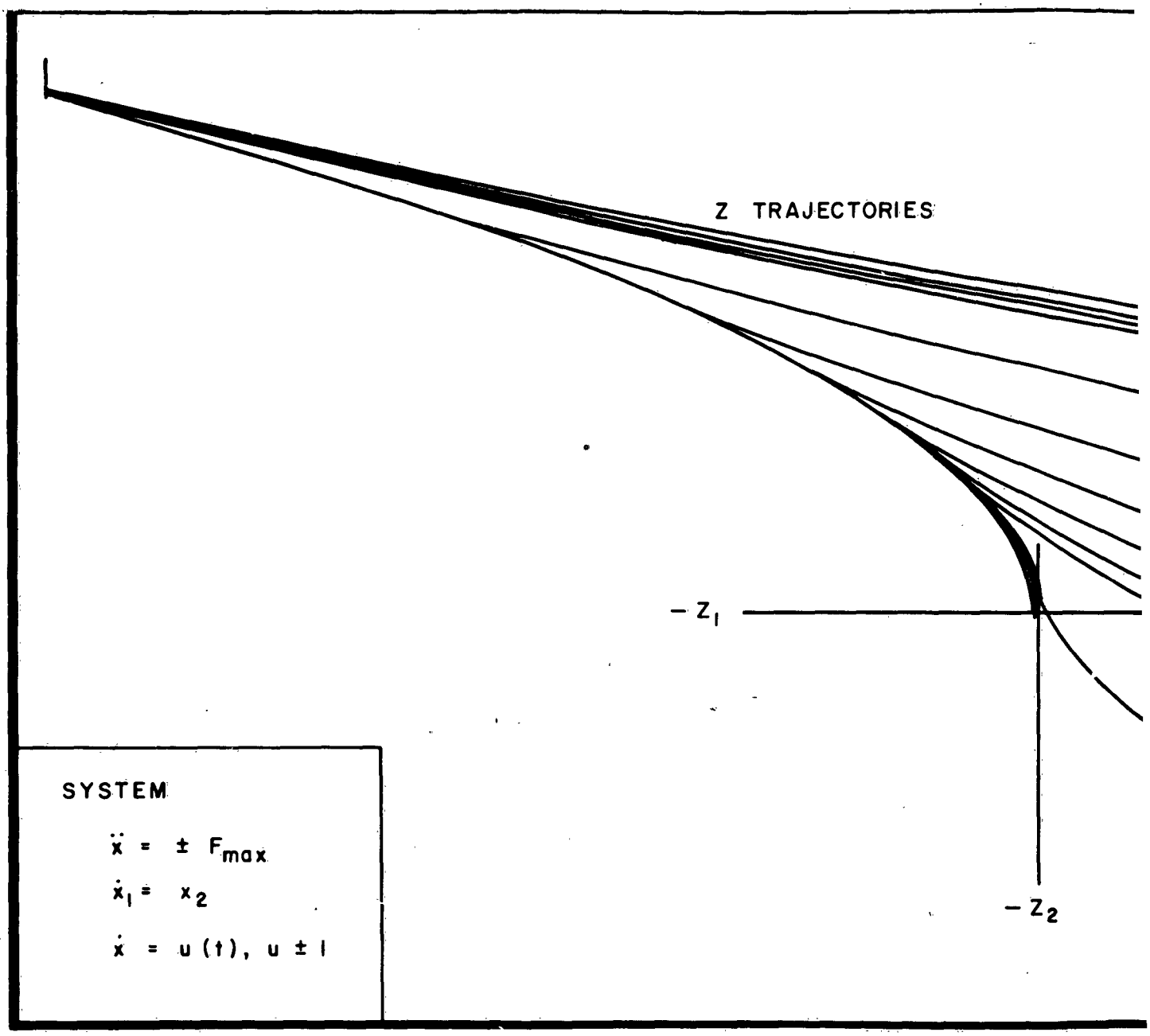
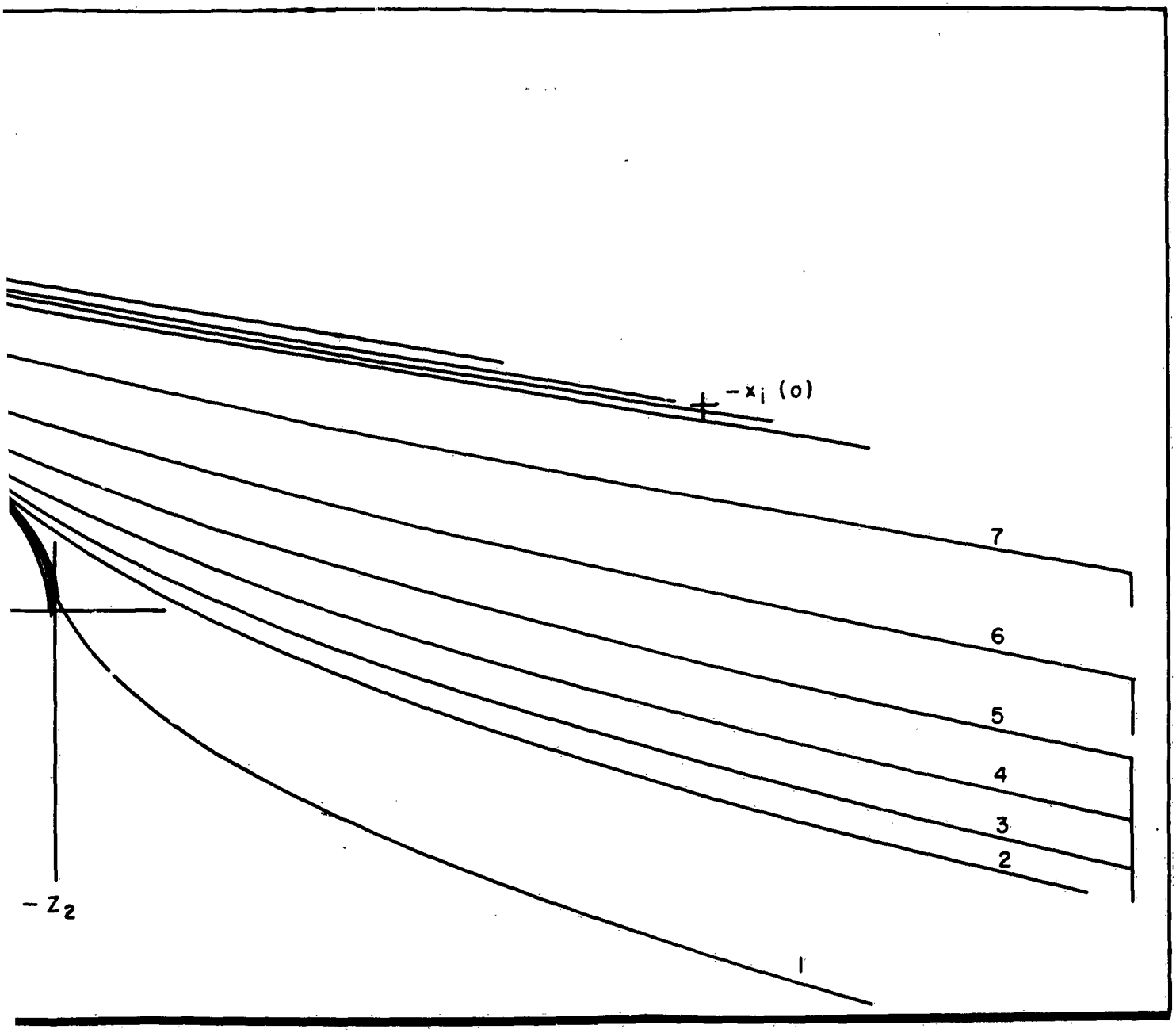
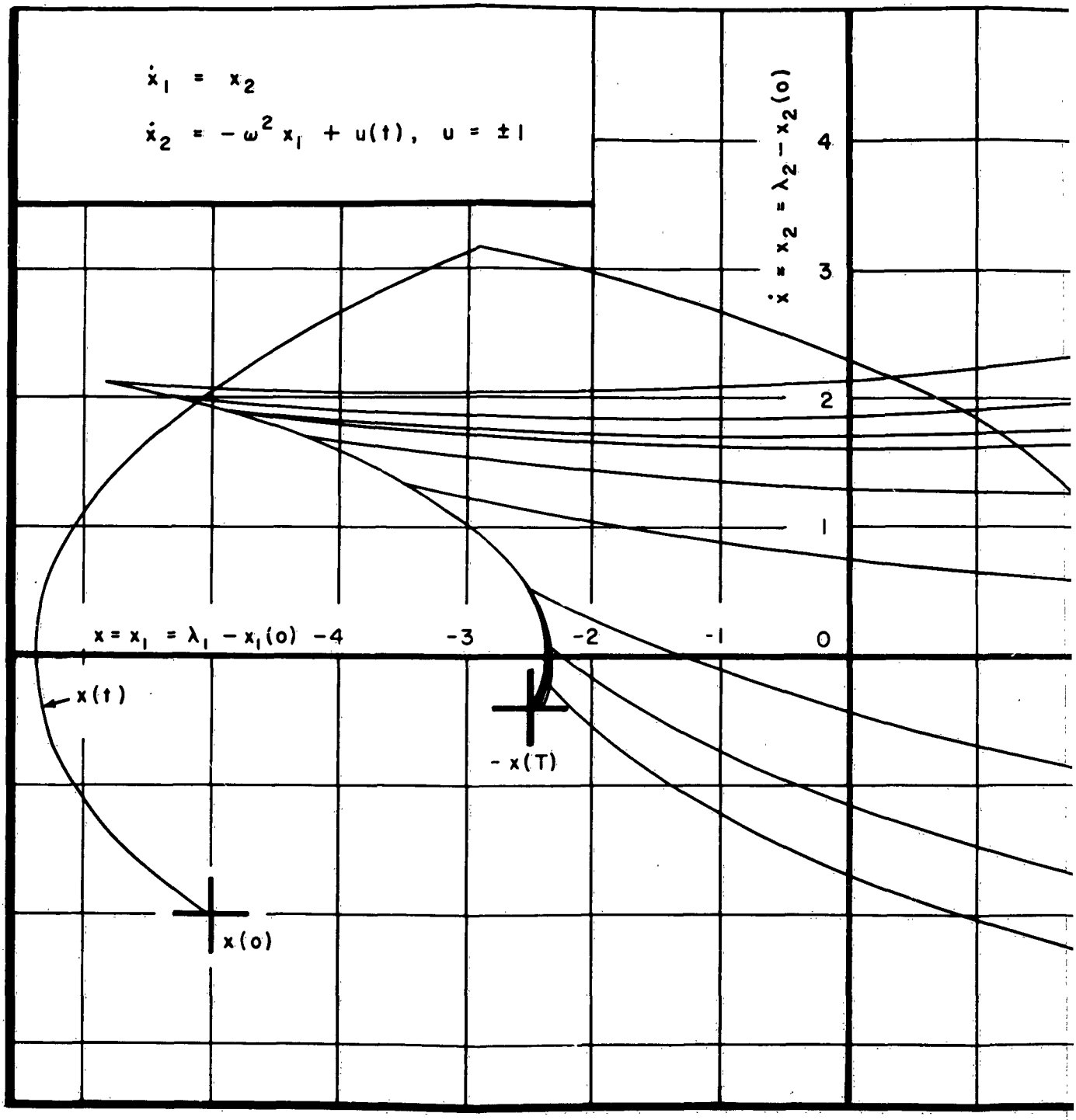


FIGURE 7. PHASE DIAGRAM TIME



AGRAM TIME OPTIMAL REGULATION

2



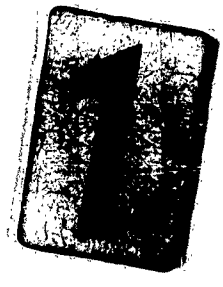
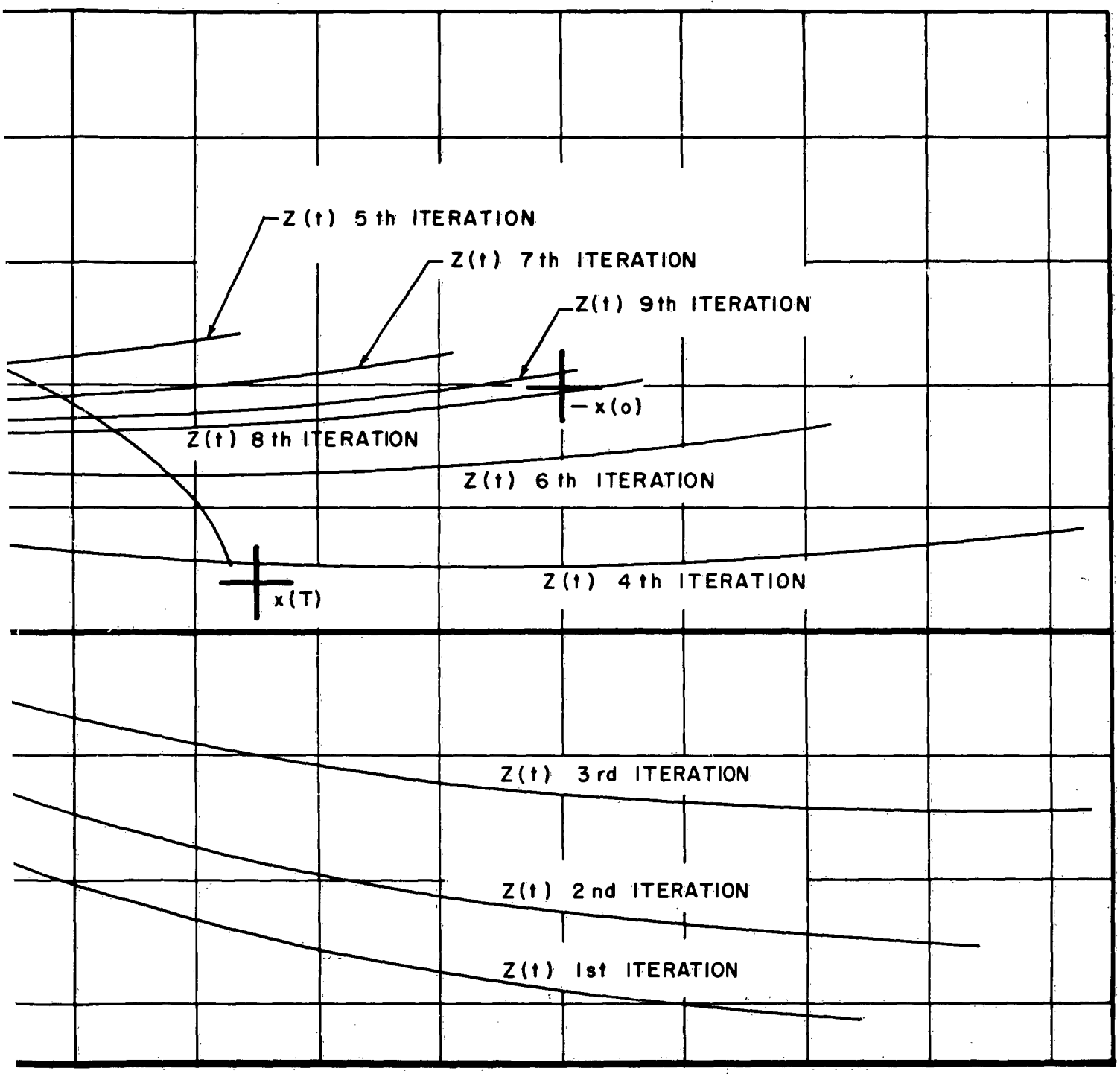


FIGURE 8. TIME OPTIMAL STEERING



IAL STEERING TO A POINT IN PHASE SPACE

2

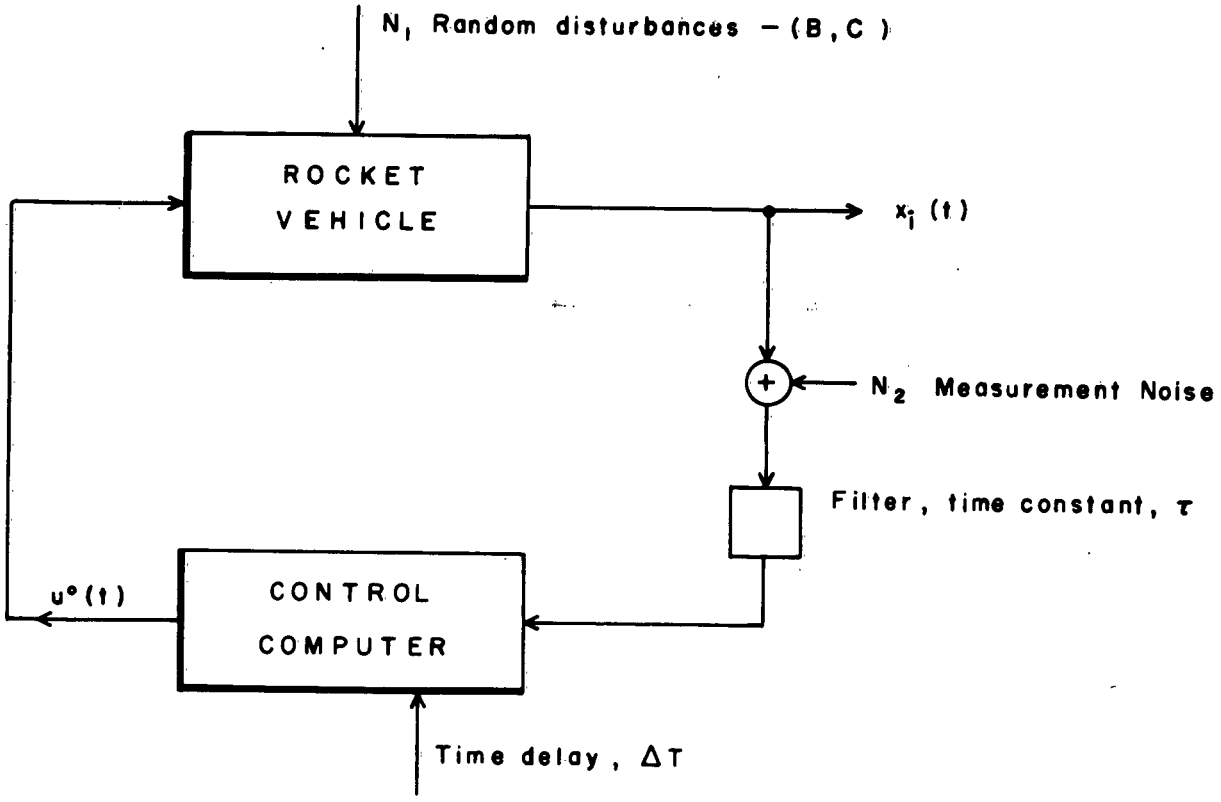


Figure 9. Closed-Loop Optimal Controller

ASD-TDR-63-119

Applications of Optimization Theory

- SECTION I - AN APPLICATION OF TIME-OPTICAL CONTROL THEORY TO LAUNCH VEHICLE REGULATION

By

F. B. Smith, Jr.
J. A. Lovingood

- SECTION II - THE EQUIVALENT MINIMIZATION PROBLEM AND THE NEWTON-RAPHSON OPTIMIZATION METHOD

By

D. K. Scharmack

NASA-Marshall Space Flight Center
Huntsville, Alabama

and

Minneapolis-Honeywell Regulator Co.
Military Products Group
Minneapolis, Minnesota

APPLICATIONS OF OPTIMIZATION THEORY

Section I - "An Application of Time-Optimal Control Theory to Launch Vehicle Regulation"

F. B. Smith, Jr., Minneapolis-Honeywell Reg. Co.
Military Products Group
Minneapolis, Minnesota

J. A. Lovingood NASA
Marshall Space Flight Center
Huntsville, Alabama

Section II- "The Equivalent Minimization Problem and the Newton-Raphson Optimization Method"

D. K. Scharmack Minneapolis-Honeywell Reg. Co.
Military Products Group
Minneapolis, Minnesota

ABSTRACT

Section I considers the practical problem of closed-loop control of linear plants. A minimum response time criterion is used in the design of a pitch attitude controller for a flexible launch vehicle. The criterion is applied to a fourth order model containing the primary dynamics of the thirteenth order vehicle. There are tried and proven methods for obtaining the control variable as a function of time which takes the system from an initial condition to the target set in minimum time. It is shown that a suitable collection of these open-loop trajectories can be used to define a closed-loop control law. Results of an analog simulation are presented which show that this control law properly applied to the flexible vehicle results in good control.

Section II considers finding the optimum path from an initial condition to a target set. The problem is reduced to an initial value problem in which the minimizing initial values of the adjoint variables or multipliers are sought. (Inequality constraints are included in the formulation.) The problem then becomes one of minimizing a function of several variables subject to constraint equations in those variables. This is a problem for which necessary and sufficient conditions for a strong relative minimum are well known. A second order Newton-Raphson iteration procedure for numerically finding the minimum is described. Finally, experiences in the use of the Newton-Raphson method are described.

SECTION I

AN APPLICATION OF TIME-OPTIMAL CONTROL THEORY TO LAUNCH VEHICLE REGULATION*

In spite of the relatively large effort which has gone into the study of optimization during the last few years, there have been few applications to significant, practical closed-loop control problems. This is true in spite of the fact that theoretical developments promise solutions or potential solutions to control problems for which conventional synthesis procedures are not completely satisfactory. Among the difficulties which have hindered practical applications are: adequate description of real plants often requires differential equations of quite high order, the control law is usually a non-linear function of many variables and difficult to implement, and the fact that the theoretical solution of the optimization problem most often yields the open-loop control law $u(t, x(o))$ rather than the required closed-loop law, $u(x)$. The flexible launch vehicle is used in this paper to illustrate these problems and to demonstrate the use of some techniques to overcome them.

1.1 Equations of Motion The assumed equations of a typical 250,000 pound flexible launch vehicle are given in Table 1. Poles and zeroes of the $\frac{\theta_R}{u}$ transfer function are listed in Table 2. Airframe coefficients are taken at the maximum dynamic pressure flight condition with flight speed assumed constant. The equations include dynamics of the rigid body, three body flexure modes, tail-wags dog, actuator, rate servo and an integration of pitch rate for control of pitch attitude. A single control variable is assumed available from gimbaling of the engine. A maximum gimbal rate of 0.2 rad/sec is commanded at all times.

1.2 Specification of the Controller In applying optimal control theory to the synthesis of controllers for practical plants it is necessary to specify both the optimization criterion and what is to be controlled. With the criterion used here, minimum response time, it has been common to apply the criterion to the state vector y, of a plant in the form,

$$\dot{y} = Ay + Bu \qquad (1.2.1)$$

However, when this is done the resulting response in multi-degree of freedom systems may be entirely unacceptable. This is forcefully demonstrated by time optimally regulating the state vector of the rigid launch vehicle given in Figure 1. When all components of the state vector, pitch attitude, pitch rate, angle of attack and gimbal deflection are brought to zero in minimum time from an initial displacement in pitch attitude of 0.01 radian, displacements of attitude and angle of attack greater than 0.15 radian occur. Although this is the

*Work reported in Section I was accomplished under NASA Contract NASr-27.

time-optimal response for regulation of the state vector, it is certainly not acceptable since it would literally destroy the vehicle. On the other hand, if the problem posed is that of bringing the single component, pitch attitude, to zero in minimum time and holding it there then the deadbeat response to step input of attitude is obtained (Figure 1). In this case angle of attack and gimbal deflection are not zero at the response time (time when θ and $\dot{\theta}$ are first zero) but decay with a 21.7 second time constant characteristic of the plant. It has been shown that single component control can be described as motion to a region in the n-dimensional space. The target region is determined as that region in n-space where the component being controlled is zero and is capable of being held there with a bounded control variable. (Reference 1, 2) The necessary and sufficient conditions for minimum time motion to such a region have been obtained (Reference 3).

In the work presented in this paper, optimum control synthesis techniques are demonstrated for control of pitch attitude. The controller obtained is fourth order, one dimensional. That is, the control variable is a function of four variables, and the target set is a line segment in this four-space. Choice of pitch attitude was arbitrary. The techniques apply equally as well to control of other components of the state vector or to control of a linear combination of them such as minimum drift.

1.3 A Truncated Model Although time-optimal control theory applies in principle to regulation of plants of any order, it is not desirable nor necessary to apply it in controller design to the complete plant representation when the motion of the variable being controlled is primarily influenced by relatively few variables. In the launch vehicle considered, the flexure mode frequencies are quite high and aero-dynamic coupling small so flexure has only minor effects on rigid body pitching motion. The same is true of the actuator dynamics. Consequently there is a natural division of the plant into a set of dominant and a set of secondary dynamics. Time-optimal synthesis is applied to control the dominant modes only, and conceptually the secondary dynamics act as a filter on the primary modes. This is shown in Figure 2. The transfer function $\frac{\theta_R}{u}$, for the entire plant of Table 1, has been divided into two parts

$$\frac{\theta_R(s)}{u(s)} = G_1 \cdot G_2 \qquad (1.3.1)$$

Primary dynamics are contained in

$$G_1 = \frac{0.8808\ (s + 0.0478)}{s(s + 0.02)\ (s - 1.4296)\ (s + 1.4964)} \qquad (1.3.2)$$

and secondary dynamics in G_2. Feedback of the fictitious output of G_1 is used for controller design. The partial principle coordinate methods of Reference 4 permit one to derive the

linear transformation relating the y coordinates to the state of the system, x. The transformation

$$y = Lx \tag{1.3.3}$$

where y is an m-vector, L an mxm-matrix and x an n-vector, in general, then permits the fictitious control loop of Figure 2 to be changed to the one which is physically realizable in Figure 3.

A plant in state vector form which gives the transfer function of equation (1.3.2) is,

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -0.0394 & 2.1403 & -4.404 \\ 0 & 1.00 & -0.02738 & -0.04213 \\ 0 & 0 & 0 & -0.02 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ .2 \end{bmatrix} u \tag{1.3.4}$$

This was obtained by deriving $\frac{y_1(s)}{u(s)}$ and $\frac{y_3(s)}{u(s)}$ transfer functions from a set of equations of this form but with unknown coefficients and then adjusting coefficients to give the proper poles, zeros and gains. A similar set of equations could be obtained directly from the transfer functions of equation (1.3.2) and the transformation to continuous coordinates of Reference 5.

The transformation matrix L, which relates the output of the flexible vehicle to the y variables contains many elements which are very small. It is possible to neglect these. The transformation used in the analog simulation was,

$$\begin{bmatrix} \theta_F \\ \dot{\theta}_F \\ \alpha_F \\ \delta_F \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0.0341 & 0 & 0 & 0 \\ 0 & 0.999 & 0.0729 & 0 & -0.15 \\ 0 & 0.0341 & 0.999 & 0 & -0.0016 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_R \\ \dot{\theta}_R \\ \alpha_R \\ \delta_c \\ \delta_e \end{bmatrix} \tag{1.3.5}$$

It is seen that y_1, corresponds very closely with θ_R, y_2 with $\dot{\theta}_R$, y_3 with α_R and y_4 with δ_c. Motion of θ_F corresponds very closely with that of θ_R, so it is reasonable to take equation (1.3.4) as the truncated model of the full system.

Two points should be emphasized in the choosing of a truncated model for controller design. First, division of the plant into primary and secondary dynamics cannot be made until the variable to be controlled has been specified. This variable may be one of the physical variables appearing in the state vector x or may be a linear combination of them. Second,

even if the secondary dynamics are a result of a limited number of physical variables in the equations of motion (equation 1.2.1), the primary dynamics cannot be obtained by simply neglecting these variables and equations. For example, if the equations for the flexure modes and actuator were omitted in truncating to a fourth order model corresponding to G_1, the poles at -1.4962 and 1.4296 would be at -1.47 and 1.403.

1.4 Closed-Loop Time Optimal Control Law The next step in the synthesis procedure is to derive a closed-loop controller for the model of equation (1.3.4). The criterion for design is time-optimal regulation of pitch attitude; that is y_1 is to be brought to zero from an initial condition in minimum time subject to a bounded control variable, and then held at zero. This corresponds to motion to a one dimensional line segment in the four dimensional space of θ_F, $\dot{\theta}_F$, α_F, δ_F.

There is no known method for obtaining a useful closed form expression for the closed-loop control law u(x) which moves the plant to the desired line segment optimally. However, it is possible to compute open-loop solutions u(t,x(0)) for any initial condition x(0) using the computational techniques described in Reference 6. These techniques solve a set of transcendental equations for a control variable u(t,x(0)) which is constrained to satisfy the maximum principle. Since the maximum principle has been shown to be a necessary and sufficient condition for the optimum solution, the u(t,x(0)) obtained is the optimal one. It is not practical to solve the required equations on line to achieve effective closed-loop control. Instead a collection of open-loop optimum trajectories from a set of initial conditions distributed evenly throughout the phase space region of interest is used to define a closed-loop control law by the method described in Reference 7. Each of the variables θ_F, $\dot{\theta}_F$, α_F, and δ_F is divided up into 32 regions called quanta. A Boolean variable X_j^i, is defined for each quantum (i= 1, 2, 3, 4, j=1, 2,...32). The variable X_j^i takes a value one if the measured magnitude of the i^{th} variable is within the j^{th} region and takes the value zero if the magnitude is within any other region. A logic form,

$$u(x) = \text{sign}\left[\sum_{i=1}^{4}\sum_{j=1}^{32} X_i^j \lambda_i^j\right] \qquad (1.4.1)$$

is assumed capable of mechanizing the control law and the 128 constants, λ_i^j, are experimentally adjusted to make u(x) agree with the optimum control at discrete points on the optimum trajectory. This adjustment or training procedure is shown in Figure 4. Switch S is opened at t = 0 and the open-loop optimal solution u(t) applied to the simulated plant. Output of the plant x(t), is the input to the logical net and the output of the net u(x(t)), is compared with the optimum control variable u(t,x(0)) at discrete intervals of time. If the control variables are different λ_i^j corresponding to

the X_i^j s which are one for that x(t) are incremented in the direction to make the sign of their sum the same as the sign of u(t,x(0)). If u(x(t) and u(t,x(0)) are the same then no adjustment is made. This procedure was carried out on a general purpose digital computer using a set of 198 optimum trajectories for the plant of equation (1.3.4), distributed in the space,

$$\begin{aligned} 0 &\leq \theta_F \leq 0.1 \\ -0.12 &\leq \dot{\theta}_F \leq 0.12 \\ -0.1 &\leq \alpha_F \leq 0.1 \\ -0.12 &\leq \delta_F \leq 0.12 \end{aligned} \qquad (1.4.2)$$

Control variable comparison points were at intervals of 0.1 second. As the adjustment is carried out, the number of differences (called errors) between u(x(t)) and u(t,x(0)) is an indication of the convergence of the procedure. The per cent errors, $100 \frac{\text{No. of errors in N points}}{N}$, is plotted as a function of the number of trajectories in Figure 5. First switch points are those between t = 0 and the first switch time, second switch points between the first switch time and the second, etc. Initial λ_i^j were all taken to be zero. It is seen that errors drop very rapidly at first, being less than 10 per cent after only 100 trajectories. At 5000 and 7500 trajectories the resolution of the logic of equation (1.4.1) is artifically increased by multiplying all λ_i^j's by two. At 11,000 trajectories the λ_i^j's are multiplied by a factor of ten. Typical closed-loop control responses using the logic at the stages of training shown in Figure 5 are presented in Figure 6. At 198 trajectories the controller has not yet stabilized the statically unstable vehicle. At 2100 trajectories the closed loop is apparently stable but responses are poor. At 11,000 trajectories responses closely approximate optimum. (Limited hardware did not permit evaluation of closed-loop responses at 13,500 trajectories). The logic of equation (1.4.1) with constants at 11,000 trajectories is taken as the closed-loop controller for the plant of Table 1.

1.5 Control of the Flexible Vehicle A block diagram of the control system is given in Figure 7. Mechanization of the logical net for this optimal control of the fourth order plant was accomplished using standard, commercial analog to digital converters for quantization and diode-transistor logic in conjunction with standard ladder networks to form the logic of equation (1.4.1)(Reference 7). A linear switching mode of the control variable was used when the plant output was within approximately one quantum of the target set. This reduced residual errors due to switching on a quantized switching surface and held the plant within the target set. The linear switching used in this mode was,

$$U = \text{sign}\left[\theta_F + 1.25\,\dot{\theta}_F + 0.65\,\ddot{\theta}_F\right].$$

No attempt was made to minimize the steady state limit cycle with the control variable in this mode.

Two schemes for measurement of the variables fed back to the controller were investigated. The first measured the state of the system using the method of Reference 8 which uses a complement of n sensors in measuring the state of an n^{th} order system. In the second, a rigid body pitch rate signal was derived using the phase blending technique of Reference 9. This provided a signal which could be freed of first mode influence, however, in this case a slight amount of first mode feedback was included in the signal to damp the first mode bending.

Typical analog responses are shown in Figure 8, 9, 10 and 11. Rigid body pitch attitude responses are quite similar for rigid body feedback and for blender feedback of pitch rate. The small amount of first mode feedback (blender gain k_1=0.9) causes the first mode to damp out with the blender system whereas with rigid body feedback there is a sustained oscillation. When the blender gain K_1 was set to cancel all first mode feedback (K_1 = 1.0), the blender system also exhibited a sustained oscillation of the first mode. Responses to 40-fps sharp-edged gusts are shown in Figure 10. The single component attitude regulator essentially ignores the gust disturbance and maintains the desired attitude. Figure 11 illustrates response to various command inputs. Although the system was designed to approximate time optimal regulation, it exhibits a very good following capability.

1.6 Conclusion It has been shown that the collection of experimental procedures and theoretical knowledge is sufficient to use a time-optimal regulation criterion for rational design of controllers for a high order plant with known coefficients. The synthesis procedure includes obtaining a representative set of open-loop optimum trajectories for a truncated model which is based on the dominant dynamics of the plant. The set of open-loop trajectories is used to define a closed-loop control law for the model. When this controller is applied to the full plant, the output is effectively that of the optimally controlled model filtered by the secondary dynamics of the plant. The resulting controller is relatively simple. However measurement requirements are severe in that the entire state of the system must be measured. This is feasible using the methods of Reference 8 but undesirable because of the large number of sensors required. Such schemes as the gyro blender give promise of relaxing these requirements.

REFERENCES

1. Lee, E. B., "On the Time-Optimal Regulation of Plants with Numerator Dynamics," IRE Transactions on Automatic Control, Vol AC-6 No. 3, p. 351, September 1961.

2. Harvey, C. A., "On Determining the Switching Criterion for Time-Optimal Control," International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Colorado Springs, Colorado, July 1961.

3. Harvey, C. A. and Lee, E. B., "On the Uniqueness of Time-Optimal Control for Linear Processes," Journal of Mathematical Analysis and Application, 1962.

4. Lovingood, J. A., Approximations to State Vector Control, Adaptive State Vector Control, MH-MPG Report 1529-TR7, 31 May 1962 (Available from NASA).

5. Laning, J. H. and Battin, R. H., Random Processes in Automatic Control, McGraw-Hill Book Co., Inc., New York, 1956.

6. Stone, C. R., Smith F. B., and Lovingood, J. A., An Introduction to Self-Evaluating State Vector Control of Linear Systems, Adaptive State Vector Control, MH-MPG Report 1529-TR1, 13 June 1962 (Available from NASA).

7. Smith, F. B., A Logical Net Mechanization for Time-Optimal Regulation, Adaptive State Vector Control, MH-MPG Report 1529-TR8, 15 August 1962 (Available from NASA).

8. Harvey, C. A., Measurement of the State Vector, Adaptive State Vector Control, MH-MPG Report 1529-TR3, 15 March 1962 (To be published as a NASA tech note).

9. Lee, R. C. K. and Falkner, V. L., "Adaptive Control Systems for Large Elastic Boosters," 1960 IRE Symposiums on Adaptive Control Systems, Garden City, Long Island, October 17-19, 1960.

Table 1. Equations of Plant in Standard Form

$$
\begin{bmatrix}
\dot{\theta} \\ \dot{\beta} \\ \dot{\alpha} \\ \dot{\delta}_c \\ \dot{\delta}_e \\ \ddot{\delta}_e \\ \dddot{\delta}_e \\ \dot{\xi}_1 \\ \ddot{\xi}_1 \\ \dot{\xi}_2 \\ \ddot{\xi}_2 \\ \dot{\xi}_3 \\ \ddot{\xi}_3
\end{bmatrix}
=
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-0.04119 & 2.044 & 0 & -4.296 & 0 & -0.00132 & -0.007378 & -0.4351(10^{-4}) & 0.07056 & 0.9747(10^{-5}) & -0.14405 & 0.1619(10^{-4}) \\
+1.00003 & -0.02484 & 0 & -0.044529 & 0 & -0.00001234 & -0.11798(10^{-3}) & -0.49797(10^{-6}) & +0.4372(10^{-3}) & 0.22946(10^{-6}) & -0.00155 & -0.56710(10^{-6}) \\
0 & 0 & -0.02 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 128000 & -128000 & -4360 & -43.6 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
16.5905 & 1036.0 & 0 & 4794.7 & 0 & 1.6409 & -322.8 & -0.19485 & -65.777 & -0.022306 & 185.61 & 0.037698 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
-1.52804 & 231.44 & 0 & -4448.5 & 0 & -1.6509 & -2.9269 & -0.014518 & -2151 & -0.68564 & -177.74 & -0.025685 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
75.0402 & -3744.6 & 0 & 5214.5 & 0 & 2.409 & -8.3915 & +0.07525 & -118.32 & -0.013664 & -8589.2 & -0.95553
\end{bmatrix}
\begin{bmatrix}
\theta \\ \beta \\ \alpha \\ \delta_c \\ \delta_e \\ \dot{\delta}_e \\ \ddot{\delta}_e \\ \xi_1 \\ \dot{\xi}_1 \\ \xi_2 \\ \dot{\xi}_2 \\ \xi_3 \\ \dot{\xi}_3
\end{bmatrix}
+
\begin{bmatrix}
0 \\ 0 \\ 0 \\ +0.2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
u
$$

Table 2. Critical Frequencies of $G(s) = \frac{\theta_R(s)}{U(s)}$

Static gain of airframe = 0.0985 rad/rad

Source	Poles	Zeros
Pseudo - Integrator	-0.02	--
Actuator	-30 $-6.5 \pm j\,65\sqrt{0.99}$	--
Rigid Body	1.4296 -1.4964	-0.0478
Tail Wags Dog	--	$-0.0233 \pm j\,57.00$
First Bending Mode	$-0.0962 \pm j\,18.00$	$-0.120 \pm j\,17.77$
Second Bending Mode	$-0.233 \pm j\,46.34$	$-0.225 \pm j\,46.81$
Third Bending Mode	$-0.479 \pm j\,92.69$	$-0.454 \pm j\,94.35$

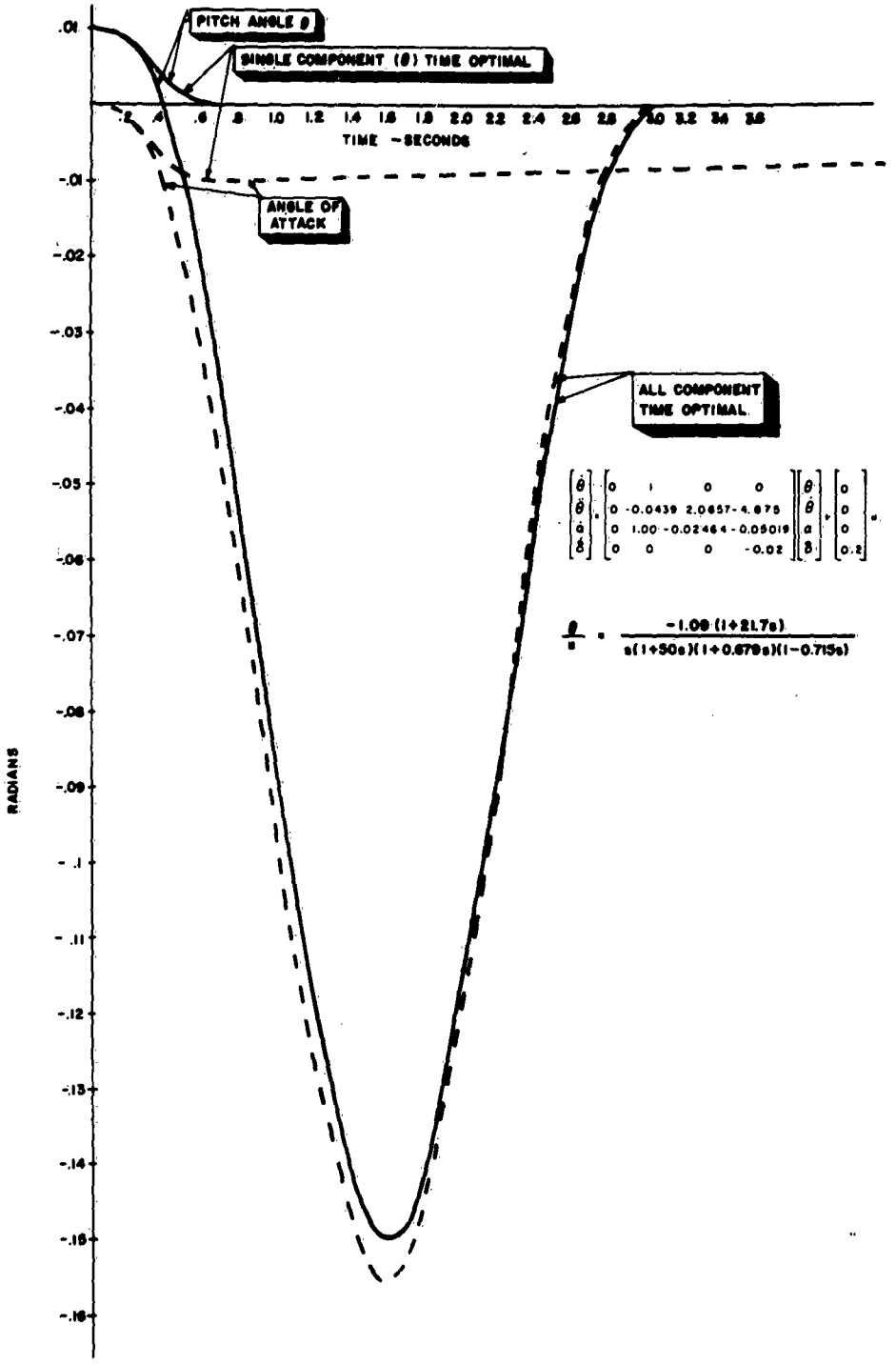


Figure 1. Time Optimal Control of a Rigid Launch Vehicle

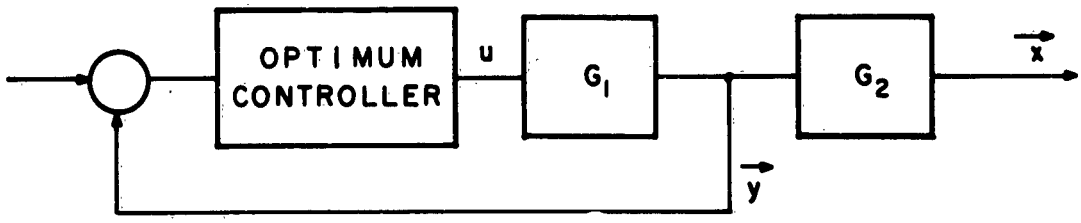


Figure 2. Conceptual Feedback for Controller Synthesis

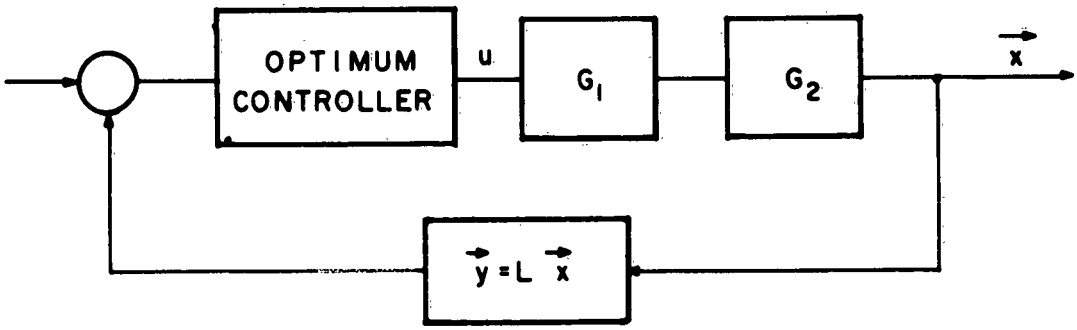


Figure 3. Actual Feedback for Controller Mechanization

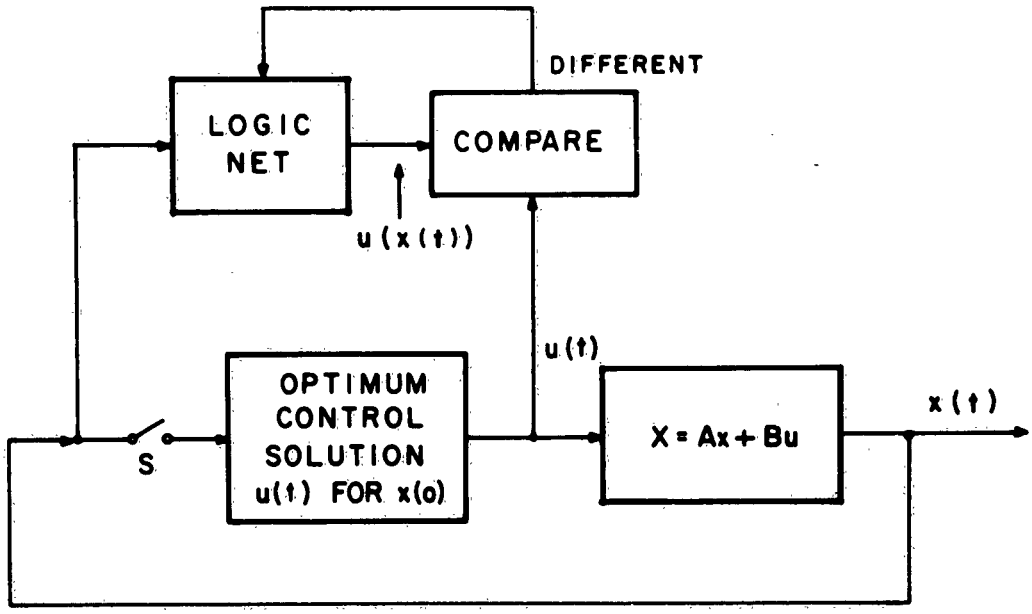


Figure 4. Logic Adjustment Procedure

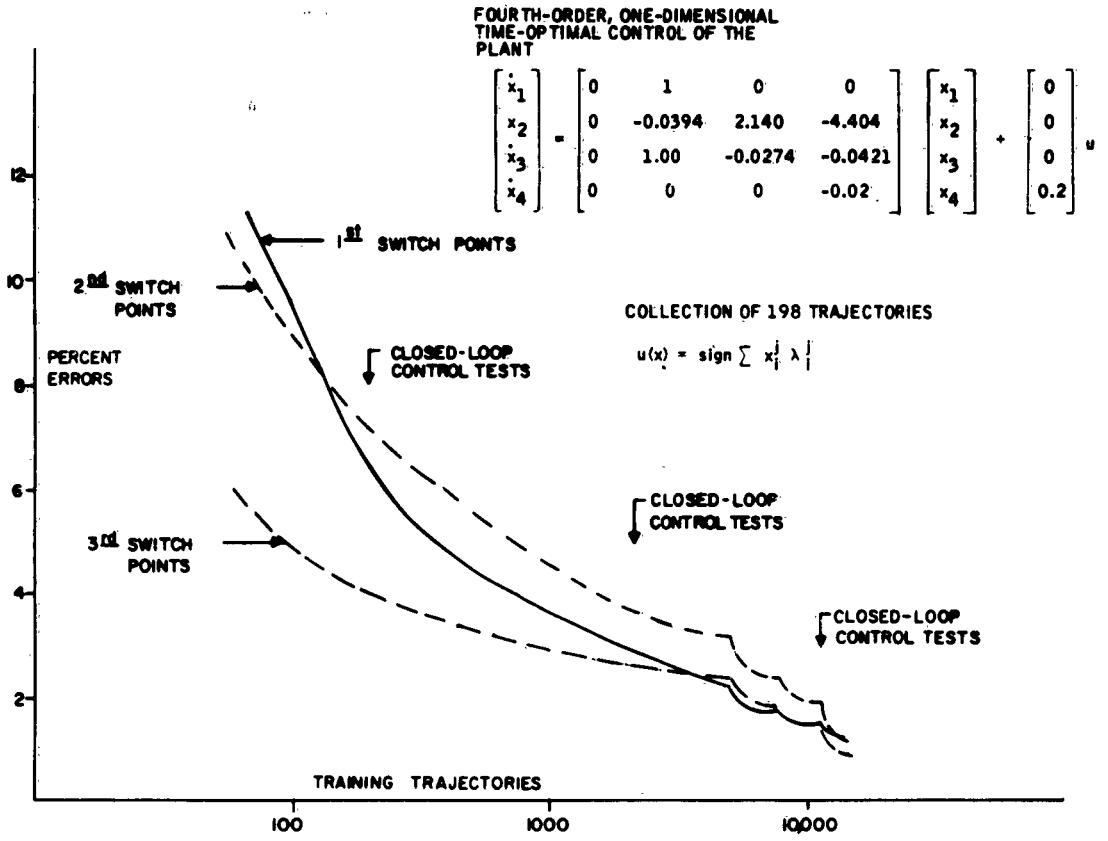


Figure 5. Fourth-order Training Curve

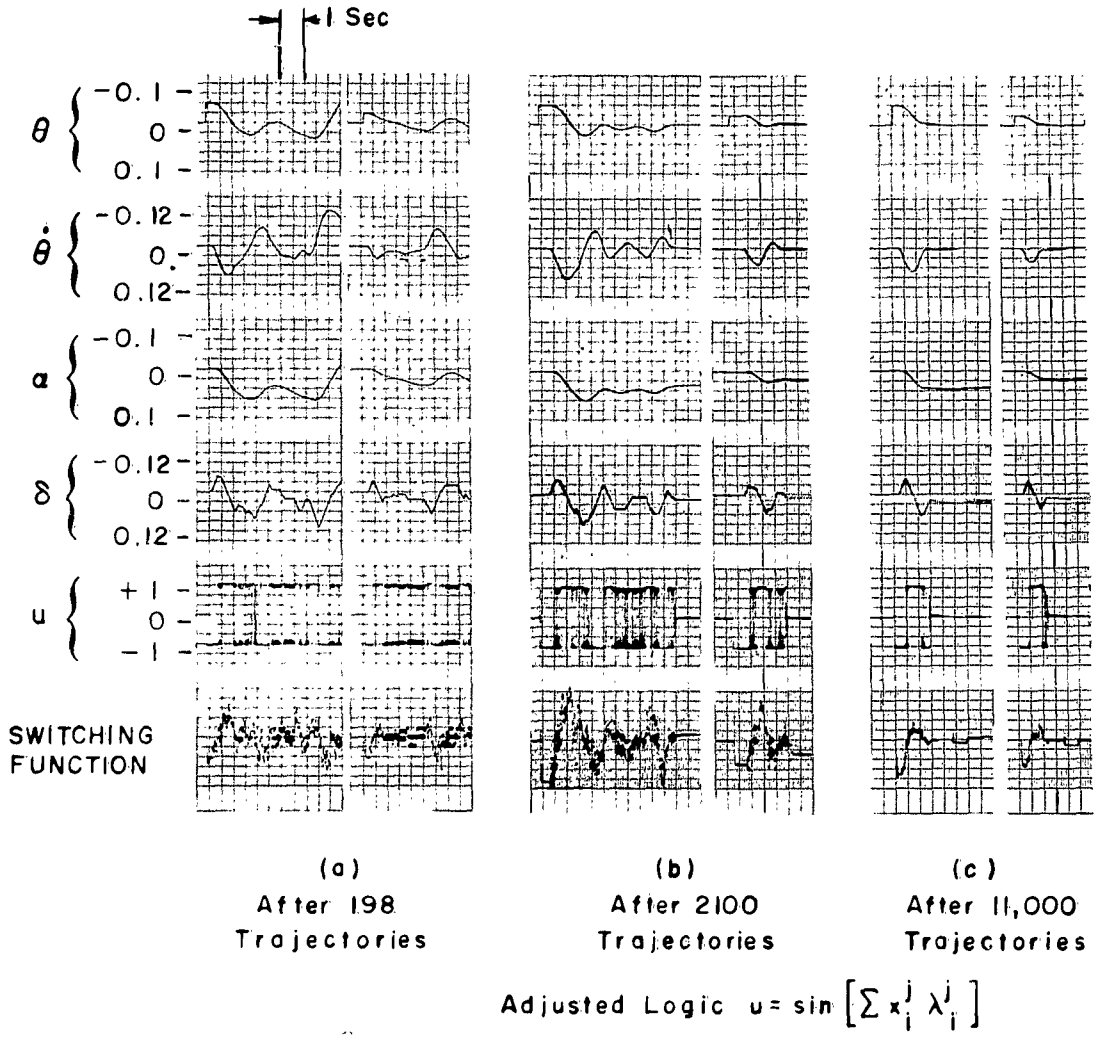


Figure 6. Closed-loop Responses at Three Stages of Training

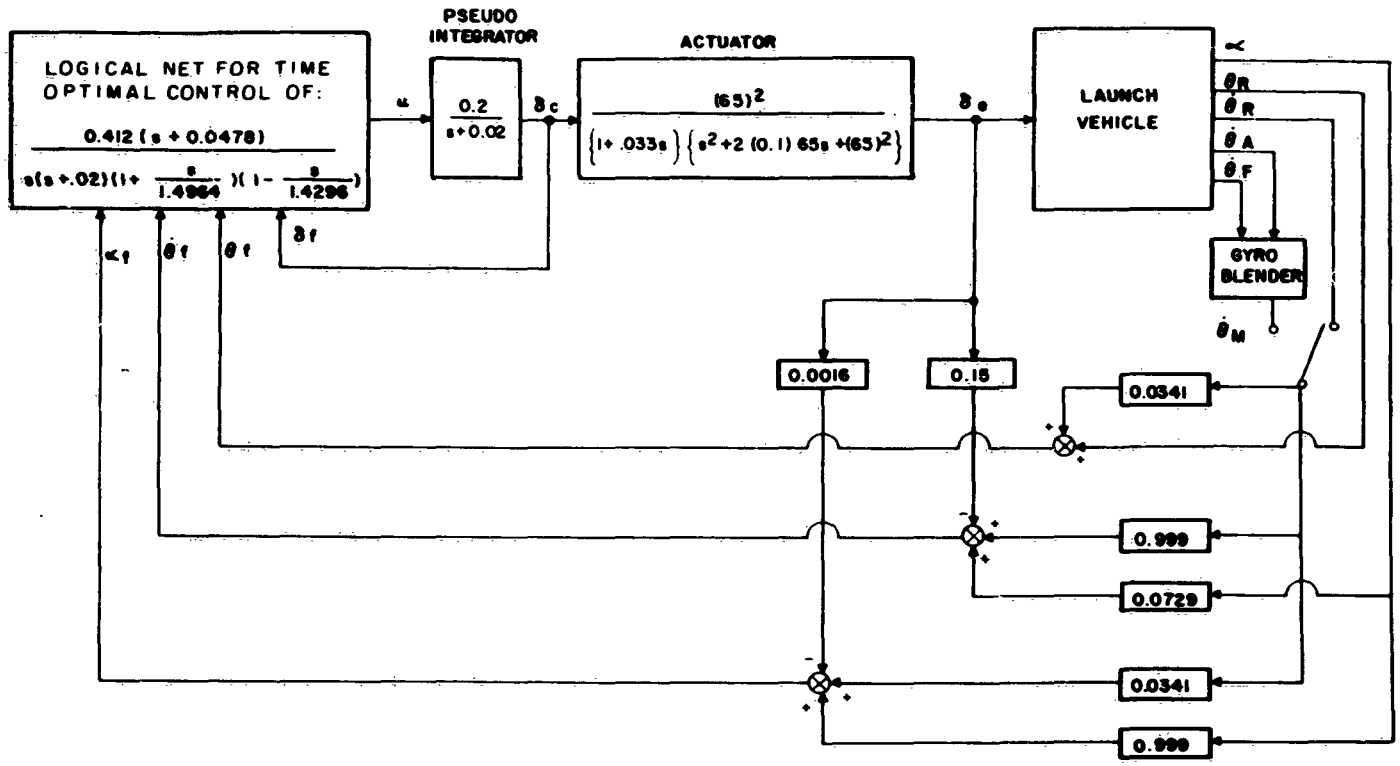


Figure 7. Block Diagram - Fourth-Order Attitude Regulation

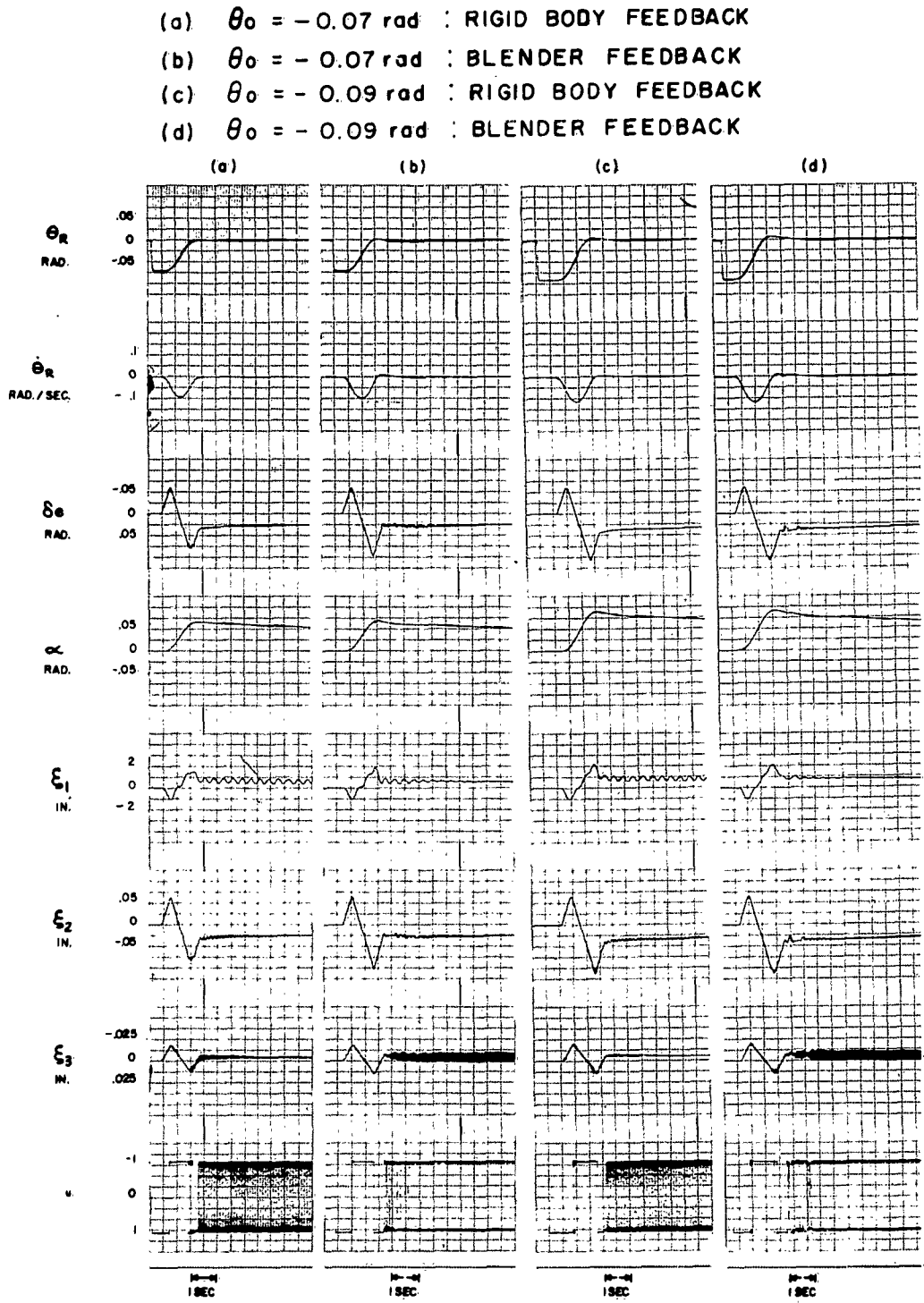


Figure 8. Attitude Regulation - Response to Initial Displacement Errors

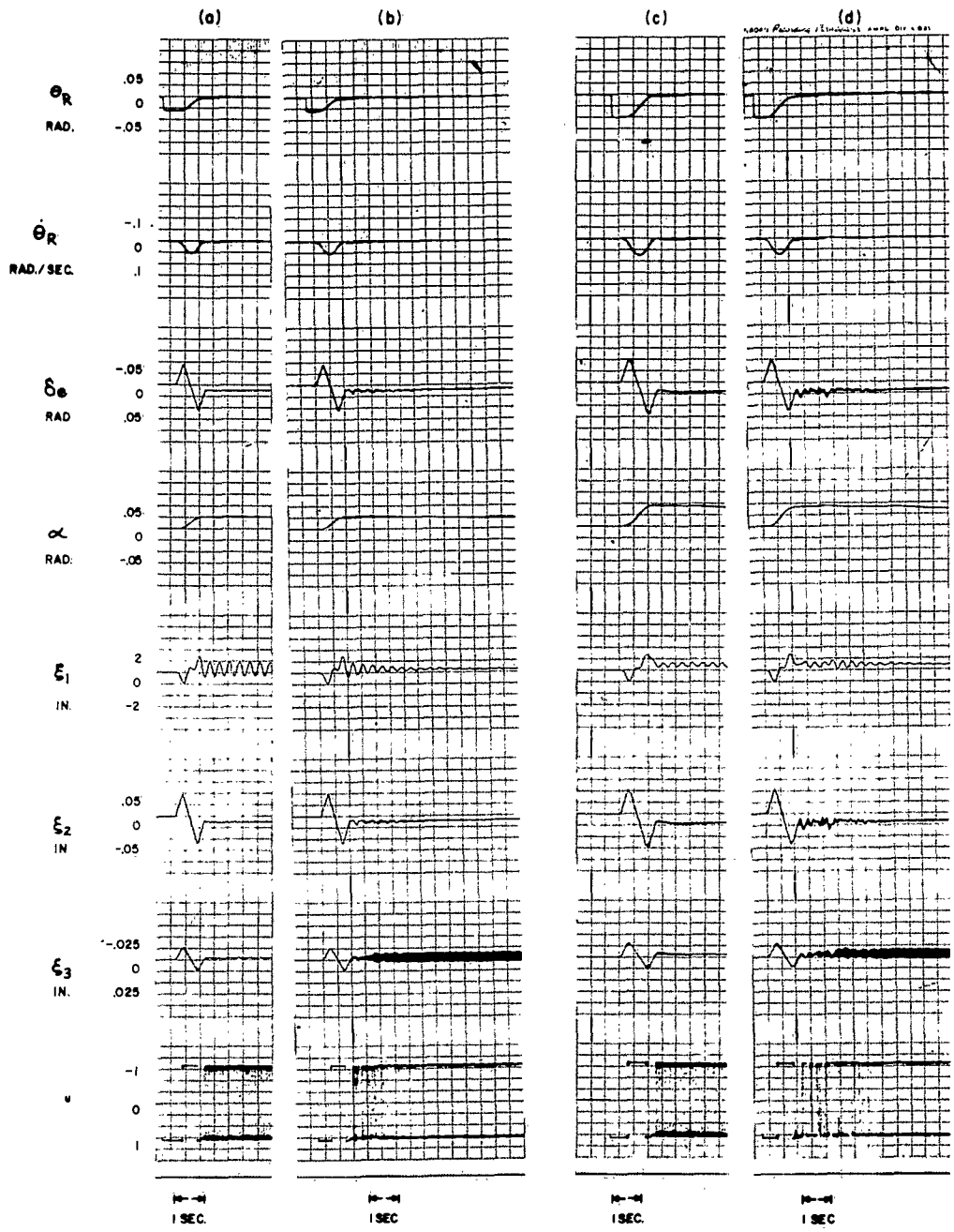


Figure 9. Attitude Regulation - Response to Initial Displacement Errors

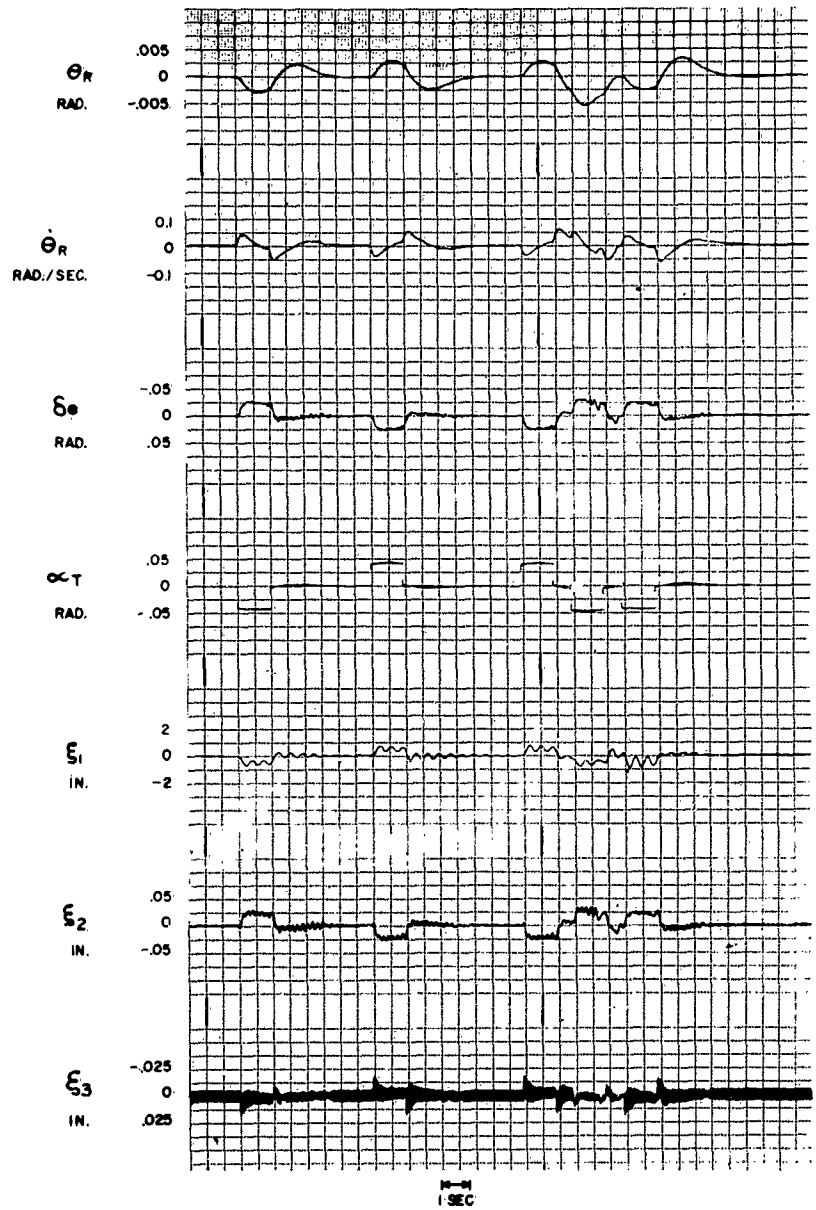


Figure 10. Attitude Regulation - Response to 40 fps Sharp-edge Gust

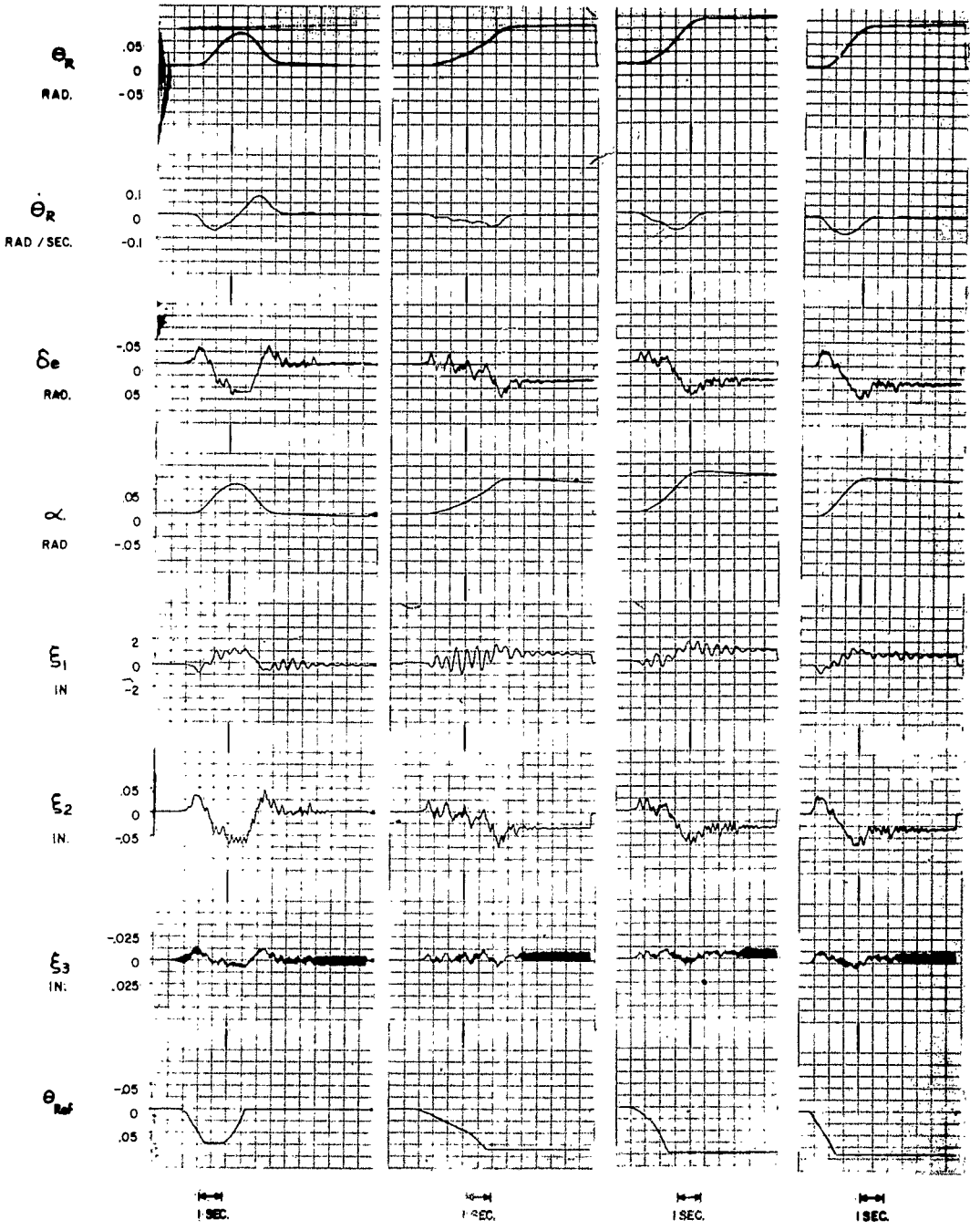


Figure 11. Attitude Control - Response to Command Inputs

SECTION II

THE EQUIVALENT MINIMIZATION PROBLEM AND THE NEWTON-RAPHSON OPTIMIZATION METHOD*

2.1 Introduction The purpose of this section is to develop a second-order optimization scheme for the control optimization problem. The particular form of the problem of Bolza which is considered is first stated. This is followed by the three necessary conditions for the given form of the problem. These are used to develop the reduced differential equations of the extremals, which must be satisfied by the optimum path. It is then shown that the solutions of these equations are functions of the independent variable and the initial conditions for the problem. It is further shown that the solution has continuous partial derivatives of at least second order in these variables. This immediately suggests the simpler equivalent minimization problem; that of minimizing a function of several variables subject to constraint equations in those variables. This problem has well known necessary and sufficient conditions for a relative minimum. These are then used to develop the second-order Newton-Raphson optimization method. The sufficiency condition for a relative minimum is incorporated in this method. The Newton-Raphson method for the fixed end-point problem is then presented, followed by the method of steepest descent.

Three example problems are presented. The first of these, in which the integral of the square of the control function is minimized, is simple enough to allow an analytical solution. It is shown that a relative minimum does exist, and furthermore, that it is a global minimum. The presented concepts are also illustrated by this problem. The second example is the time optimal bang-bang problem. This is presented as an example of a problem with corners which can be treated by the Newton-Raphson optimizing scheme. The third example is a re-entry problem which must be solved on a computer. The deviation of heating rate from a given constant is to be minimized. An inequality constraint on the sensed acceleration is included in the problem formulation. The reduced differential equations of the extremals are derived for this problem.

A special form for the Newton-Raphson method is then developed. This assumes that the terminal surface is described by a given terminal time alone. Conclusions from the application of this method to the third example problem are summarized.

2.2 Problem Formulation The problem considered here is a special form of the problem of Bolza as formulated by Bliss[2]

---

* Work in Section II was supported under U. S. Air Force Contract AF 33(657)-7383, ASD.

and extended to include inequality constraints by Valentine.[7] It is desired to find that path which minimizes the function

$$J = g(\bar{x}(T), T) + \int_{t_o}^{T} f_o(\bar{x}, \bar{u})d\tau \qquad (2.2.1)$$

subject to differential equations of the form

$$\dot{\bar{x}} = \bar{f}(\bar{x},\bar{u}), \qquad \bar{x}(t_o) = \bar{x}_o \qquad (2.2.2)$$

inequality constraints

$$\bar{G}(\bar{x},\bar{u}) \geq \bar{Z} \qquad (2.2.3)$$

and terminal surface equations

$$\bar{\Psi}(T, \bar{x}(T)\ ) = \bar{Z}. \qquad (2.2.4)$$

In the above, $\bar{x}$ and $\bar{u}$ are n and m dimensional state and control column vectors respectively, and $(\cdot)$ represents differentiation with respect to the independent variable, t. The dimension of vector equation (2.2.3) is q, and that of (2.2.4) is r, where $r \leq n + 1$. $\bar{Z}$ represents the zero vector of appropriate dimension.

Following Valentine,[7] equations (2.2.3) are rewritten in the form

$$\dot{\bar{\sigma}}^2 = \bar{G}(\bar{x},\bar{u}) \qquad (2.2.5)$$

where the components of $\dot{\bar{\sigma}}^2$ are $\dot{\sigma}_j^2$, $j = 1,\ldots,q$. The $\dot{\sigma}$'s are real slack variable derivatives, introduced to complete the set of differential equations for the problem of Bolza. No more than m of the $\dot{\sigma}$'s may be zero at any point on the path.

It is assumed that $f_o(\bar{x},\bar{u})$ and the differential equations (2.2.2) and (2.2.5) have continuous partial derivatives of at least third order in all variables in an open region S_1 about the minimizing path. Furthermore, the matrix made up of the partial derivatives of the differential equations with respect to all derivatives and the control functions must have rank n+q at each point of the minimizing path. This insures that the differential equations are independent. The matrix has the form

$$\begin{bmatrix} -I & \nabla_u \bar{f} & Z \\ Z' & \nabla_u \bar{G} & -\Sigma \end{bmatrix} \qquad (2.2.6)$$

where I is the nxn identity matrix, Z is an nxq zero matrix Z' is the transpose of Z, and Σ is a qxq diagonal matrix whose elements are $2\dot{\sigma}_j$. In addition,

$$\nabla_u \bar{f} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial u_m} \\ \vdots & & \\ \frac{\partial f_n}{\partial u_1} & \cdots & \frac{\partial f_n}{\partial u_m} \end{bmatrix} \tag{2.2.7}$$

$$\nabla_u \bar{G} = \begin{bmatrix} \frac{\partial G_1}{\partial u_1} & \cdots & \frac{\partial G_1}{\partial u_m} \\ \vdots & & \vdots \\ \frac{\partial G_q}{\partial u_1} & \cdots & \frac{\partial G_q}{\partial u_m} \end{bmatrix} \tag{2.2.8}$$

The solution, $\bar{x}(t)$, of the differential equations (2.2.2) and (2.2.5) is supposed to be continuous, with at least absolutely continuous first derivatives. In some instances it will be possible to consider corners, i.e., points at which the derivatives are discontinuous. The control functions are treated as derivatives in (2.2.6) so that they can be discontinuous, as in the bang-bang problem. Potential corners are those points at which inequality constraints change from greater than to equality states or vice versa. They may also be defined by switching points, as in the bang-bang problem where the constraint is always zero. A subarc is defined as that part of the path between corners or potential corners.

Finally, the functions $g(T, \bar{x}(T))$ and (2.2.4) are assumed to have continuous partial derivatives of at least third order in an open set S_2 of points $(T, \bar{x}(T))$, and the matrix

$$\nabla \bar{\Psi} = \begin{bmatrix} \frac{\partial \Psi_1}{\partial t} & \frac{\partial \Psi_1}{\partial x_1} & \cdots & \frac{\partial \Psi_1}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \Psi_r}{\partial t} & \frac{\partial \Psi_r}{\partial x_1} & \cdots & \frac{\partial \Psi_r}{\partial x_n} \end{bmatrix} \tag{2.2.9}$$

is assumed to have rank r. An admissible arc is defined to be a path, where all of its elements $(t,\bar{x},\dot{\bar{x}},\bar{u})$ and $(T, \bar{x}(T))$ lie in S_1 and S_2 respectively.

It might be remarked that differential equations and inequality constraints containing t explicitly can easily be brought to the form of equations (2.2.2) and (2.2.5). The independent variable is changed from t to s by adding the differential equation

$$\frac{dt}{ds} = 1, \qquad s_o = t_o \tag{2.2.10}$$

and noting that

$$\frac{d\bar{x}}{dt} = \frac{d\bar{x}}{ds}\,\frac{ds}{dt} = \frac{d\bar{x}}{ds} \quad . \tag{2.2.11}$$

Then t becomes a dependent variable, and the resulting n+q+1 differential equations are in the desired form. It should be noted that the order of the system has been increased by one.

2.3 The Necessary Conditions and the Initial Value Problem
The multiplier rule and the necessary conditions of Wierstrass and Clebsch are given here for this problem without proof. The reader is referred to references (1), (2) and (7) for their derivation. It is then shown that if these necessary conditions provide unique descriptions of the control functions over every subarc of the path, then the optimization problem is an initial value problem.

The first necessary condition for this problem is The Multiplier Rule. An admissible arc E, defined on an interval $[t_o,T]$ is said to satisfy the multiplier rule if there exist constants $p_o = 1$, $\bar{e} = [e_1,\ldots,e_r]$, not all zero and a function

$$F(t,\bar{x},\bar{u},\dot{\bar{x}},\bar{p},\bar{\mu},\dot{\bar{\sigma}}) = f_o + \bar{p}(\bar{f}-\dot{\bar{x}}) + \bar{\mu}(\bar{G} - \dot{\bar{\sigma}}^2) \tag{2.3.1}$$

with multipliers $\bar{p}(t) = [p_1(t),\ldots,p_n(t)]$, continuous on $[t_o,T]$ and $\bar{\mu}(t) = [\mu_1(t),\ldots,\mu_q(t)]$ continuous on $[t_o,T]$ except possibly at corners of E where unique right and left hand limits exist, and satisfying the Euler-Lagrange equations

$$\dot{\bar{p}} = -\left[\nabla_x f_o + \bar{p}\,\nabla_x \bar{f} + \bar{\mu}\,\nabla_x \bar{G}\right] \tag{2.3.2}$$

$$\bar{Z} = \nabla_u f_o + \bar{p}\,\nabla_u \bar{f} + \bar{\mu}\,\nabla_u \bar{G} \tag{2.3.3}$$

$$0 = \mu_j G_j, \qquad j = 1,\ldots,q \tag{2.3.4}$$

where

$$\bar{Z} \geq \bar{\mu}, \tag{2.3.5}$$

and differential equations

$$\dot{\bar{x}} = \bar{f}; \qquad \bar{x}(t_o) = \bar{x}_o \tag{2.3.6}$$

$$\dot{\bar{\sigma}}^2 = \bar{G} \geq \bar{Z} \tag{2.3.7}$$

along E, and furthermore, such that the equations

$$\left(f_o + \bar{p}\bar{f} + \frac{\partial g}{\partial t} + \bar{e}\,\frac{\partial \bar{\psi}}{\partial t}\right)\Big|_T \, dT + \tag{2.3.8}$$

$$+ (\nabla_x g + \bar{e}\nabla_x \bar{Y} - \bar{p})\Big|_T \quad d\bar{x}(T) = 0, \quad \bar{Y} = \bar{Z}$$

hold for every choice of the differentials dT, $d\bar{x}(T)$. The multipliers p_o, $\bar{p}$, $\bar{\mu}$ do not vanish simultaneously at any point of the interval $[t_o\ T]$ for an arc E satisfying the multiplier rule. Furthermore, the function

$$H \equiv f_o + \bar{p}\bar{f} \tag{2.3.9}$$

is a constant on E. Every minimizing arc E for the given form of the problem of Bolza must satisfy the multiplier rule.

In the above

$$\nabla_x f_o = \left[\frac{\partial f_o}{\partial x_1}, \ldots, \frac{\partial f_o}{\partial x_n}\right], \quad \nabla_u f_o = \left[\frac{\partial f_o}{\partial u_1}, \ldots, \frac{\partial f_o}{\partial u_m}\right] \tag{2.3.10}$$

$$\nabla_x g = \left[\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_n}\right]$$

$$\nabla_x \bar{f} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}, \quad \nabla_x \bar{G} = \begin{bmatrix} \frac{\partial G_1}{\partial x_1} & \cdots & \frac{\partial G_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial G_q}{\partial x_1} & \cdots & \frac{\partial G_q}{\partial x_n} \end{bmatrix} \tag{2.3.11}$$

with $\nabla_u \bar{f}$ and $\nabla_u \bar{G}$ defined by equations (2.2.7) and (2.2.8) respectively. The vector $\frac{\partial \bar{Y}}{\partial t}\Big|_T$ is the first column of the matrix (2.2.9) and $\nabla_x \bar{Y}\Big|_T$ is the remainder of the matrix.

The vanishing of the coefficients of dT and $d\bar{x}(T)$ in equation (2.3.8) constitutes the transversality condition on the arc E. Thus,

$$H = (f_o + \bar{p}\bar{f})_t = (f_o + \bar{p}\bar{f})_T = -\left(\frac{\partial g}{\partial t} + \bar{e}\,\frac{\partial \bar{Y}}{\partial t}\right)\Big|_T \tag{2.3.12}$$

$$\bar{p}(T) = (\nabla_x g + \bar{e}\nabla_x \bar{Y})\Big|_T. \tag{2.3.13}$$

The Hamiltonian, equation (2.3.12), is often called a first integral for this problem. It has further significance in the second necessary condition.

The Necessary Condition of Weierstrass (The Minimum Principle). An admissible arc $\bar{E}$ satisfying the multiplier rule with multipliers $p_o = 1$, $\bar{p}$, $\bar{\mu}$, is said to satisfy the necessary condition of Weierstrass with these multipliers if the condition

$$H(t,\bar{x},\bar{p},\bar{u}) \leq H(t,\bar{x},\bar{p},\bar{U}) \qquad (2.3.14)$$

is valid at every element $(t,\bar{x},\dot{\bar{x}},\bar{u})$ of E for all admissible points $(t,x,\dot{x},U) \neq (t,x,\dot{x},u)$ satisfying the equations (2.2.2) and (2.2.3). Every minimizing arc E for the given form of the problem of Bolza must satisfy this condition.

A consequence of the Weierstrass condition is

The Necessary Condition of Clebsch. At each point of E let $\tilde{G}$ be a vector whose components are those components of $\bar{G}$ which vanish, and $\tilde{\mu}$ the corresponding multipliers. Then the inequality

$$\bar{\pi}' \nabla_u^2 (H + \tilde{\mu}\tilde{G}) \bar{\pi} \geq 0 \qquad (2.3.15)$$

must be satisfied for every vector $\bar{\pi} \neq \bar{Z}$ where $\bar{\pi}' = [\pi_1,\ldots,\pi_m]$ and satisfies the equations

$$\nabla_u \tilde{G} \bar{\pi} = \bar{Z} . \qquad (2.3.16)$$

Every minimizing arc for the given form of the problem of Bolza must satisfy this condition.

In equation (2.3.15) the operator ∇_u^2 is the matrix

$$\partial_u^2 = \begin{bmatrix} \frac{\partial^2}{\partial u_1^2},\ldots, & \frac{\partial^2}{\partial u_1 \partial u_m} \\ \vdots & \\ \frac{\partial^2}{\partial u_m \partial u_1},\ldots, & \frac{\partial^2}{\partial u_m^2} \end{bmatrix}$$

It will now be shown that the problem can be considered to be an initial value problem. For this purpose, the arc is split up into subarcs. On each subarc a given subset of the inequality constraints are equality constraints, and all the rest are greater than zero, except possibly at a finite number of points. It is supposed that there are no corner points anywhere on the path. For simplicity, it will be assumed that all the constraints are greater than zero along the first subarc. Then according to Valentine,[7] the inequality constraints may be neglected over this subarc. The function (2.3.1) may thus be written

$$F_1 = f_o + \bar{p}(\bar{f} - \dot{\bar{x}}) \qquad (2.3.17)$$

The corresponding Euler-Lagrange equations (2.3.2) and (2.3.3) are

$$\dot{\bar{p}} = -\left[\nabla_x f_o + \bar{p}\, \nabla_x \bar{f}\right] = \nabla_x F_1 \qquad (2.3.18)$$

$$\bar{Z} = \nabla_u f_o + \bar{p}\, \nabla_u \bar{f} = \nabla_u F_1 . \qquad (2.3.19)$$

It is well known from implicit function theory that equations (2.3.19) can be solved (at least in principle) in the form

$$\bar{u} = \bar{u}(\bar{x},\bar{p}) \tag{2.3.20}$$

if the determinant

$$R_1 = \det \nabla_u^2 F_1 = \det \begin{bmatrix} \frac{\partial^2 F_1}{\partial u_1^2}, & \cdots, & \frac{\partial^2 F_1}{\partial u_1 \partial u_m} \\ \vdots & & \vdots \\ \frac{\partial^2 F_1}{\partial u_m \partial u_1}, & \cdots, & \frac{\partial^2 F_1}{\partial u_m^2} \end{bmatrix} \tag{2.3.21}$$

is different from zero. Equation (2.3.21) is the determinant of the Hilbert differentiability condition for this problem. Substitution of (2.3.20) into (2.3.18) and (2.2.2) thus produces a 2n set of differential equations

$$\dot{\bar{x}} = \bar{f}_1(\bar{x}, \bar{p}), \qquad \bar{x}(t_o) = \bar{x}_o \tag{2.3.22}$$

$$\dot{\bar{p}} = \nabla_x F_1(\bar{x}, \bar{p}) \tag{2.3.23}$$

These will be called the reduced differential equations of the extremals. It is well known from the theory of differential equations that (2.3.22) and (2.3.23) have a unique solution for a given set of initial conditions. But half of these, $\bar{x}_o$, are given for the problem. The solution is thus a function of the n initial conditions, $\bar{p}_o$, and is said to be imbedded in an n-parameter family of extremals. Over the first subarc, then the solution has the form

$$\bar{x} = \bar{x}(t,\bar{x}_o,\bar{p}_o) = \bar{x}(t,\bar{p}_o) \tag{2.3.24}$$

$$\bar{p} = \bar{p}(t,\bar{x}_o,\bar{p}_o) = \bar{p}(t,\bar{p}_o) \tag{2.3.25}$$

$$\bar{u} = \bar{u}(t,\bar{x}_o,\bar{p}_o) = \bar{u}(t,\bar{p}_o) \tag{2.3.26}$$

It is further known from the theory of differential equations that equations (2.3.24) and (2.3.25), and consequently (2.3.26), have continuous partial derivatives in the variables, t, $\bar{x}_o$ and $\bar{p}_o$ of at least second order.

Before continuing, it should be remarked that the above results are of course obtainable from the general form of the problem of Bolza. It is known that the arc can be imbedded in a 2n + 2m parameter family of arcs, and that there are 2n + 2m differential equations of the extremals. The 2m

differential equations become the m algebraic equations (2.3.19) and their time derivatives (which introduce no new information) when initial conditions are imposed. Furthermore, the determinant in the Hilbert differentiability condition reduces to the form (2.3.21).

Over the second subarc it is assumed that one of the inequality constraints, say G_1, is equal to zero. The function (2.3.1) is written, following Valentine,[7] as

$$F_2 = f_o + \bar{p}\,(\bar{f} - \dot{\bar{x}}) + \mu_1 G_1 \tag{2.3.27}$$

and the Euler-Lagrange equations as

$$\dot{\bar{p}} = -\left[\nabla_x f_o + \bar{p}\,\nabla_x \bar{f} + \mu_1\, G_1\right] = \nabla_x F_2 \tag{2.3.28}$$

$$\bar{Z} = \nabla_u f_o + \bar{p}\,\nabla_u \bar{f} + \mu_1\,\nabla_u G_1 = \nabla_u F_2 \tag{2.3.29}$$

To this is added the equality constraint

$$0 = G_1 \tag{2.3.30}$$

If the m+1 equations (2.3.29) and (2.3.30) are to be solved for the m+1 variables $\bar{u}$ and μ_1, the determinant

$$R_2 = \det \begin{bmatrix} \nabla_u^2 F_2 & \nabla_u' G_1 \\ \nabla_u G_1 & 0 \end{bmatrix} \tag{2.3.31}$$

must not be zero. Here, ∇_u' is the transpose of the operator ∇_u. Equation (2.3.31) is the determinant of the Hilbert differentiability condition for this arc. It is then found that

$$\bar{u} = \bar{u}(\bar{x},\ \bar{p}) \tag{2.3.32}$$

$$\mu_1 = \mu_1(\bar{x},\ \bar{p}) \tag{2.3.33}$$

and consequently that

$$\dot{\bar{x}} = \bar{f}_2(\bar{x},\ \bar{p}) \tag{2.3.34}$$

$$\dot{\bar{p}} = \nabla_x F_2(\bar{x},\ \bar{p}), \tag{2.3.35}$$

the reduced differential equations of the extremals for this arc. These equations again have a unique solution for a given set of initial conditions. Furthermore, the solutions possess continuous partial derivatives of at least second order with respect to t and the initial conditions $\bar{x}_1$ and $\bar{p}_1$. It remains to be shown that the solutions over this arc are continuous functions of the initial conditions, $\bar{p}_o$, and that partial derivatives of at least second order exist.

The terminal point of the first subarc is defined by the equation

$$G_1(\bar{x}\ (t_1,\bar{p}_o),\ \bar{p}(t_1,\bar{p}_o)) = 0. \tag{2.3.36}$$

This may be solved for $t_1(\bar{p}_o)$ if the derivative

$$\dot{G}_1 = \nabla_x G_1\ \dot{\bar{x}} + \nabla_p G_1\ \dot{\bar{p}},$$

is non-zero. Equation (2.3.36) represents the equation for the terminal surface of the family of extremals which are solutions of equations (2.3.22) and (2.3.23). This terminal surface represents the initial surface for the family of extremals over the second subarc which are solutions of equations (2.3.34) and (2.3.35). For a given set $\bar{p}_o$, the terminal values $\bar{x}(t_1,\ \bar{p}_o)$, $\bar{p}(t_1,\ \bar{p}_o)$ are the initial values for the differential equations (2.3.34) and (2.3.35). This follows from the continuity of $\bar{x}(t)$ and $\bar{p}(t)$. Then since the solutions

$$\bar{x}(t) = \bar{x}(t,\bar{x}_1,\bar{p}_1), \qquad t \geq t_1 \tag{2.3.37}$$

$$\bar{p}(t) = \bar{p}(t,\bar{x}_1,\bar{p}_1), \qquad t \geq t_1 \tag{2.3.38}$$

are continuous functions of $\bar{x}_1$ and $\bar{p}_1$, and since

$$\bar{x}_1 = \bar{x}(t_1,\ \bar{p}_o) = \bar{x}_1(\bar{p}_o) \tag{2.3.39}$$

$$\bar{p}_1 = \bar{p}(t_1,\ \bar{p}_o) = \bar{p}_1(\bar{p}_o) \tag{2.3.40}$$

are continuous functions of $\bar{p}_o$ alone, it follows that

$$\bar{x}(t) = \bar{x}(t,\ \bar{p}_o), \quad t \geq t_1 \tag{2.3.41}$$

$$\bar{p}(t) = \bar{p}(t,\ \bar{p}_o), \quad t \geq t_1. \tag{2.3.42}$$

Furthermore, since continuous partial derivatives of at least

second order exist for equations (2.3.37) - (2.3.40) in the indicated variables, it follows that (2.3.41) and (2.3.42) possess partial derivatives of at least second order in t, $\bar{p}_o$. Since there are no corners, μ_1 is at least absolutely continuous.

Further subarcs may now be added, so long as the number of equality constraints does not exceed m, the dimension of the control vector. Each subarc possesses control functions and multipliers of the form (2.3.32) and (2.3.33), and furthermore, differential equations similar to (2.3.34) and (2.3.35). It is readily verified that the solutions are continuous functions of the initial conditions, $\bar{p}_o$, and that continuous partial derivatives of at least second order exist.

Points where equality constraints change to the "greater-than" state will now be examined. For this purpose it will be assumed that only one of the constraints, say G_1, is zero over the first subarc, and that it is greater than zero over the second. Equations (2.3.32) - (2.3.35) hold over the first subarc, and (2.3.20), (2.3.22) and (2.3.23) over the second subarc. The terminal surface, (2.3.36), is replaced by the equation

$$\mu_1(\bar{x}\ (t_1,\bar{p}_o),\ \bar{p}(t_1,\bar{p}_o)) = 0 \qquad (2.3.43)$$

since μ_1 is a continuous function, and must go to zero before G_1 can be greater than zero. It is then seen that the arguments follow through as before, provided the determinants R_1 and R_2 are different from zero, and that the derivative

$$\dot{\mu}_1 = \nabla_x \mu_1\ \dot{\bar{x}} + \nabla_p \mu_1\ \dot{\bar{p}}'$$

is non-zero on the terminal surface of the first subarc. On the second subarc, then, the solution is a continuous function of t and $\bar{p}_o$, and partial derivatives in these variables of at least second order exist. The result is easily generalized to several equality constraints going to the greater than state over a series of subarcs.

To sum up, a path is composed of a finite number of subarcs, each of which has its own set of differential equations and terminal surfaces. The first subarc has a specified set of initial conditions, $\bar{x}_o$, and the last one has a specified terminal surface, equations (2.2.4). The equations of this terminal surface have continuous partial derivatives of at least second order with respect to the initial conditions, $\bar{p}_o$, and the terminal value of the independent variable, T.

2.4 The Equivalent Minimization Problem It will be shown here that the minimization problem can be reduced to the more familiar one of minimizing a function of several

variables subject to constraint equations in those variables. For this purpose, the integral term of equation (2.2.1) is written as a differential equation

$$\dot{x}_{n+1} = f_o, \qquad x_{n+1}(t_o) = 0 \tag{2.4.1}$$

and added to the differential equations of the extremals. The solution to this is easily seen to be of the form

$$x_{n+1} = x_{n+1}(t,\bar{p}_o) = \int_{t_o}^{t} f_o(\bar{x},\bar{p})d\tau \tag{2.4.2}$$

over the path, and on the terminal surface,

$$x_{n+1}(T) = x_{n+1}(T,\bar{p}_o). \tag{2.4.3}$$

Furthermore, since g of equation (2.2.1) is a functional of continuous functions,

$$J(T) = J(T,\bar{p}_o) = g(T, \bar{x}(T,\bar{p}_o)) + x_{n+1}(T,\bar{p}_o) \tag{2.4.4}$$

The terminal surface equations (2.2.4) may be expressed in the form

$$\bar{\Psi}(T) = \bar{\Psi}(T,\bar{p}_o) = \bar{\Psi}(\bar{T}, \bar{x}(T,\bar{p}_o)) = \bar{Z}. \tag{2.4.5}$$

It is assumed for the present that the dimension, r, of (2.4.5) is $\leq n$. The minimizing arc must satisfy the three necessary conditions, and hence the differential equations of the extremals. Since the solutions are functions of the initial conditions, $\bar{p}_o$ and t, the problem is thus one of minimizing (2.4.4) in the variables T, $\bar{p}_o$, subject to the constraint equations (2.4.5). This is a problem for which necessary and sufficient conditions for a relative minimum are well known.* The first necessary condition requires that the partial derivatives of the function

$$F = J(T, \bar{p}_o) + \bar{\lambda}\ \bar{\Psi}(T, \bar{p}_o) \tag{2.4.6}$$

with respect to T, $\bar{p}_o$, all vanish at the minimizing point. Here $\bar{\lambda} = [\lambda_1,\ldots,\lambda_r]$ is a set of constant undetermined Lagrange multipliers. One thus finds

---

* for example, see Bliss, reference (2), pp. 210-211.

$$\frac{\partial F}{\partial T} = f_o + \left(\frac{\partial g}{\partial T} + \nabla_x g\ \dot{\bar{x}}\right) + \bar{\lambda}\left(\frac{\partial \bar{\Psi}}{\partial T} + \nabla_x \bar{\Psi}\dot{x}\right) = 0 \qquad (2.4.7)$$

$$\nabla_{p_o} F = \nabla_{p_o} x_{n+1} + \nabla_x g\ \nabla_{p_o} \bar{x} + \bar{\lambda}\nabla_x \bar{\Psi}\ \nabla_{p_o} \bar{x} = \bar{Z} \qquad (2.4.8)$$

where $\nabla_{p_o} \bar{x}$ is the matrix composed of partial derivatives of the functions

$$\bar{x}(T) = \bar{x}(T, \bar{p}_o) \qquad (2.4.9)$$

with respect to the initial conditions $\bar{p}_o$. This nxn matrix is non-singular.

Equations (2.4.7) and (2.4.8) must, of course, be related to the boundary condition (2.3.8) of the multiplier rule. The relationship can be found by substituting

$$d\bar{x}(T) = \dot{\bar{x}}\, dT + \nabla_{p_o} \bar{x}\, d\bar{p}_o \qquad (2.4.10)$$

into the boundary condition (2.3.8) and rearranging. Here, $d\bar{p}_o$ is a column vector. This gives

$$\left[f_o + \left(\frac{\partial g}{\partial T} + \nabla_x g\dot{\bar{x}}\right) + \bar{e}\left(\frac{\partial \bar{\Psi}}{\partial T} + \nabla_x \bar{\Psi}\dot{x}\right)\right] dT + \left[-\bar{p}(T)\nabla_{p_o} \bar{x}\right.$$

$$\left. + \nabla_x g\ \nabla_{p_o} \bar{x} + \bar{e}\ \nabla_x \bar{\Psi}\ \nabla_{p_o} \bar{x}\right] d\bar{p}_o = 0. \qquad (2.4.11)$$

Vanishing of the coefficients of dT and $d\bar{p}_o$ in this equation is equivalent to the transversality condition. Comparison of (2.4.7) with the coefficient of dT in (2.4.11) then shows that

$$\bar{\lambda} = \bar{e}, \qquad (2.4.12)$$

i.e., the undetermined Lagrange multipliers of the boundary condition for the problem of Bolza are the same as those for the equivalent problem. Furthermore, from (2.4.8) and the coefficients of dp_o of (2.4.11), one finds

$$\nabla_{p_o} x_{n+1} = -\bar{p}(T)\ \nabla_{p_o} \bar{x}. \qquad (2.4.13)$$

If x_{n+1} is assumed to be a function of the state of the system, i.e.,

$$x_{n+1} = x_{n+1}(T, \bar{x}(T)), \tag{2.4.14}$$

then equation (2.4.13) can be rewritten in the form

$$\nabla_x x_{n+1} \nabla_{p_o} \bar{x} = -\bar{p}(T) \nabla_{p_o} \bar{x}. \tag{2.4.15}$$

Since $\nabla_{p_o} \bar{x}$ is non-singular, it is seen that

$$\nabla_x x_{n+1} = -\bar{p}(T). \tag{2.4.16}$$

Equation (2.4.16) is a well-known result[5] which finds use in dynamic programming. It is further known that (2.4.16) is valid for all values of the independent variable, t. Thus, equation (2.4.13) reduces to an identity.

Equations (2.4.6) - (2.4.8) will now be written in simplified form. Using equation (2.4.12) and the chain rule for partial derivatives, it is found that

$$F = J(T, \bar{p}_o) + \bar{e}\, \bar{\Psi}(T, \bar{p}_o) \tag{2.4.17}$$

$$\frac{\partial F}{\partial T} = f_o + \dot{g} + \bar{e}\, \dot{\bar{\Psi}} = 0 \tag{2.4.18}$$

$$\nabla_{p_o} F = \nabla_{p_o} x_{n+1} + \nabla_{p_o} g + \bar{e} \nabla_{p_o} \bar{\Psi} = \bar{Z} \tag{2.4.19}$$

It will further be assumed that the problem is normal, i.e., that the matrix

$$\begin{bmatrix} \dot{\Psi}_1, & \frac{\partial \Psi_1}{\partial p_{o_1}}, \ \dots, & \frac{\partial \Psi_1}{\partial p_{o_n}} \\ \vdots & \vdots & \vdots \\ \dot{\Psi}_r, & \frac{\partial \Psi_r}{\partial p_{o_1}}, \ \dots, & \frac{\partial \Psi_r}{\partial p_{o_n}} \end{bmatrix} \tag{2.4.20}$$

is non-singular. This is equivalent to stating that the solution is normal for the problem of Bolza. Then the sufficiency condition for a relative minimum for the equivalent problem is that the quadratic form

$$\left[\eta_o,\ \bar{\eta}'\right]\begin{bmatrix}\ddot{F} & \nabla_{p_o}\dot{F}\\ \nabla'_{p_o}\dot{F} & \nabla^2_{p_o}F\end{bmatrix}\begin{bmatrix}\eta_o\\ \bar{\eta}\end{bmatrix}=\left[\eta_o,\bar{\eta}'\right]\nabla^2 F\begin{bmatrix}\eta_o\\ \bar{\eta}\end{bmatrix} \qquad (2.4.21)$$

be greater than zero for all non-zero vectors $\left[\eta_o,\ \bar{\eta}'\right] = \left[\eta_o,\ \eta_1,\ldots,\eta_n\right]$ which satisfy the equations

$$\dot{\Psi}\,\eta_o + \nabla_{p_o}\Psi\,\bar{\eta} = \bar{Z}. \qquad (2.4.22)$$

In the above,

$$\ddot{F} = \dot{f}_o + \ddot{g} + \bar{e}\,\ddot{\Psi}$$

$$\nabla_{p_o}\dot{F} = \nabla_{p_o}\left(f_o + \dot{g} + \bar{e}\,\dot{\Psi}\right).$$

This form of the sufficiency condition is rather awkward to satisfy. An equivalent condition, which reduces to showing that a matrix is positive-definite, will now be developed.

The set (2.4.5) is extended to n+1 equations by adding equations of the form

$$\Psi_j(T,\ \bar{p}_o) = \Psi_j(\hat{T},\ \hat{p}_o) + b\zeta_j,\quad j=r+1,\ldots,n+1 \qquad (2.4.23)$$

where b is a parameter which is zero at the minimizing point, $\hat{T}$, $\hat{p}_o$. These equations are to be chosen such that the $(n+1)\times(n+1)$ matrix

$$\begin{bmatrix}\dot{\Psi} & \nabla_{p_o}\Psi\\ \dot{\Psi}_{r+1} & \nabla_{p_o}\Psi_{r+1}\\ \vdots & \vdots\\ \dot{\Psi}_{n+1} & \nabla_{p_o}\Psi_{n+1}\end{bmatrix} \qquad (2.4.24)$$

is non-singular. This is equivalent to completing the set (2.2.4) with the equations

$$\Psi_j(T,x(T)) = \Psi_j(\hat{T}, \hat{x}(T)) + b\zeta_j, \quad j=r+1,\ldots,n+1 \qquad (2.4.25)$$

since the transformation

$$\begin{bmatrix} dT \\ d\bar{x}(T) \end{bmatrix} = \begin{bmatrix} 1 & \bar{Z} \\ \dot{\hat{x}}(T) & \nabla_{p_o}\, \bar{x}(T) \end{bmatrix} \begin{bmatrix} dT \\ d\bar{p}_o \end{bmatrix} \qquad (2.4.26)$$

must be non-singular. Bliss[2] shows that

$$\frac{dT}{db} = \eta_o \qquad (2.4.27)$$

$$\frac{d\bar{p}_o}{db} = \bar{\eta} \quad . \qquad (2.4.28)$$

Then differentiation of equations (2.4.5) and (2.4.23) with respect to the parameter b gives

$$\Phi \begin{bmatrix} \eta_o \\ \bar{\eta} \end{bmatrix} = \begin{bmatrix} \bar{Z} \\ \bar{\zeta} \end{bmatrix} \qquad (2.4.29)$$

where $\bar{Z}$ is the rxl zero vector and $\bar{\zeta}$ is an (n+1-r)xl column vector whose components are ζ_j. Φ is the matrix (2.4.24). Then since this is non-singular, (2.4.29) may be rewritten

$$\begin{bmatrix} \eta_o \\ \bar{\eta} \end{bmatrix} = \Phi^{-1} \begin{bmatrix} \bar{Z} \\ \bar{\zeta} \end{bmatrix} \qquad (2.4.30)$$

Equation (2.4.30) may be thought of as a transformation which relates the η's to arbitrary $\bar{\zeta}$. Substitution of (2.4.30) into (2.4.21) then gives

$$\begin{bmatrix} \bar{Z}' & \bar{\zeta}' \end{bmatrix} (\Phi^{-1})' \nabla^2 F\, \Phi^{-1} \begin{bmatrix} \bar{Z} \\ \bar{\zeta} \end{bmatrix} > 0. \qquad (2.4.31)$$

Call the product matrix of (2.4.31) A, and partition it into four sub-matrices, A_1-A_4. Then (2.4.31) becomes

$$\begin{bmatrix} \bar{Z}' & \bar{\zeta}' \end{bmatrix} \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \begin{bmatrix} \bar{Z} \\ \bar{\zeta} \end{bmatrix} > 0 \qquad (2.4.32)$$

Upon performing the matrix multiplication, it is found that (2.4.32) reduces to

$$\bar{\zeta}' A_4 \bar{\zeta} > 0 \qquad (2.4.33)$$

Since $\bar{\zeta}$ is arbitrary, equation (2.4.33) is true if and only if the matrix A_4 is positive-definite. This is a very useful form of the sufficiency condition for the equivalent minimization problem.

2.5 The Newton-Raphson Scheme A second order scheme, based upon the equivalent minimization problem, is presented here. It is suitable for computer solution of the optimization problem, and has the advantage that the optimization process can be made to be automatic. The reason for this is that both the direction and the amount by which to move the solution can be computed. Gradient schemes, on the other hand, provide only the direction of changes. Human intervention is generally required to assertain the amount of change from step to step.

At this point, the subscript on $\bar{p}_0$ will be dropped, since only initial values of $\bar{p}$ will henceforth be considered. It will be assumed that a non-optimal solution for the reduced differential equations of the extremals has been found for the initial values $\bar{p}$. One of the components of $\bar{\Psi}$, say Ψ_1, has been used as the stopping condition on this path. Thus, the condition

$$\Psi_1 = 0 \qquad (2.5.1)$$

is satisfied, and furthermore, the terminal value of the independent parameter, T, is known. The other components of $\bar{\Psi}$ are not necessarily zero, but their values are known. It is further assumed that the partial derivatives in (2.4.20) and (2.4.21) are known. A method for estimating these for the general problem is given below. It is desired to compute new initial values, $\hat{p}$, which will simultaneously satisfy the terminal conditions (2.4.5) and minimize the function J(T) of equation (2.4.4). The terminal value of the independent variable for this path will be designated $\hat{T}$. Taylor's series expansion of the functions J and $\bar{\Psi}$ through second order terms then gives the approximate equations

$$\begin{aligned} J(\hat{T},\hat{p}) = {} & J(T,\bar{p})+\dot{J}(T,\bar{p})(\hat{T}-T)+\nabla_p J(T,\bar{p})(\hat{p}-\bar{p}) \\ & + (\hat{T}-T)\nabla_p \dot{J}(T,\bar{p})(\hat{p}-p)+\frac{1}{2}\ddot{J}(T,\bar{p})(\hat{T}-T)^2 \\ & + \frac{1}{2}(\hat{p}-\bar{p})'\nabla_p^2 J(T,\bar{p})(\hat{p}-\bar{p}) \end{aligned} \qquad (2.5.2)$$

$$\Psi_i(\hat{T},\hat{p}) = 0 = \Psi_i(T,\bar{p})+\dot{\Psi}_i(T,\bar{p})(\hat{T}-T)+\nabla_p\Psi_i(T,\bar{p})(\hat{p}-\bar{p})$$

$$+(\hat{T}-T)\nabla_p\dot{\Psi}_i(T,\bar{p})(\hat{p}-\bar{p})+\tfrac{1}{2}\ddot{\Psi}_i(T,\bar{p})(\hat{T}-T)^2$$

$$+\tfrac{1}{2}(\hat{p}-\bar{p})'\nabla_p^2\Psi_i(T,\bar{p})(\hat{p}-\bar{p}),\quad i=1,\ldots,r. \qquad (2.5.3)$$

Here, $\hat{p}$, $\bar{p}$ are column vectors. These quadratic equations approximate conditions at the minimizing point. One can thus write the function F of (2.4.6) and apply the necessary and sufficient conditions of the equivalent problem. Thus

$$F = J(\hat{T},\hat{p}) + \bar{e}\,\bar{\Psi}\,(\hat{T},\hat{p}). \qquad (2.5.4)$$

It is convenient, at this point, to introduce the function

$$F_1 = J(T,\bar{p}) + \bar{e}\,\bar{\Psi}\,(T,\bar{p}). \qquad (2.5.5)$$

F_1 has partial derivatives made up of the partial derivatives in equations (2.5.2) and (2.5.3). Thus,

$$\dot{F}_1 = \dot{J}(T,\bar{p}) + \bar{e}\,\dot{\bar{\Psi}}\,(T,\bar{p}) \qquad (2.5.6)$$

$$\nabla_p F_1 = \nabla_p J(T,\bar{p}) + \bar{e}\nabla_p\bar{\Psi}\,(T,\bar{p}) \qquad (2.5.7)$$

and so forth. Then the first necessary condition for equation (2.5.4) gives the equations

$$\frac{\partial F}{\partial T} = 0 = \dot{F}_1 + \ddot{F}_1(\hat{T}-T) + \nabla_p\dot{F}_1(\hat{p}-\bar{p}) \qquad (2.5.8)$$

$$\nabla_p F = \bar{Z} = \nabla_p F_1+\nabla_p\dot{F}_1(\hat{T}-T) + \nabla_p^2 F_1(\hat{p}-\bar{p}) \qquad (2.5.9)$$

These equations, together with (2.5.3), constitute n+r+1 equations in the n+r+1 unknowns, $\hat{p}$, $\bar{e}$, $\hat{T}$. The interated solution of these equations then constitutes the Newton-Raphson method, provided that, for the matrix

$$\nabla^2 F = \nabla^2 F_1 \qquad (2.5.10)$$

the matrix A4 of (2.4.33) is positive-definite. Otherwise, the solution would be driven toward a saddle point, or worse,

a maximizing point.

If the dimension of $\bar{\Psi}$ is n+1, the method reduces to the simpler and more familiar form of the Newton-Raphson method. In this case, one is trying to determine the minimizing path which passes through given initial and terminal points. Satisfaction of the three necessary conditions for the problem of Bolza, and hence, the reduced differential equations of the extremals, insures that the path will indeed be the minimizing path. The problem then reduces to one of finding those initial values, $\hat{p}$, which satisfy the n+1 terminal conditions (2.4.5). The Newton-Raphson equations, to be solved iteratively for $\hat{p}$, $\hat{T}$, are then

$$\Psi_i(\hat{T},\hat{p}) = 0 = \Psi_i(T,\bar{p}) + \dot{\Psi}_i(T,\bar{p})(\hat{T}-T) + \nabla_p \Psi_i(T,\bar{p})(\hat{p}-\bar{p}),\ i = 1,\ldots,n+1 \qquad (2.5.11)$$

Truncation of these equations after first-order terms in the expansion is a consequence of seeking a root, rather than a minimizing point. The matrix (2.4.20) must be non-singular for the method to work at all.

It might be remarked that some pre-digesting of the problem can be done to eliminate the variable T. If equation (2.5.1) is the stopping condition on the problem, then the equation (2.5.3) for Ψ_1 can be used to perform the elimination. The result is an n+r system of equations, (2.5.8) and (2.5.9), in the n+r unknowns $\hat{p}$, $\bar{e}$. The elimination is more readily performed for the fixed terminal point problem and the method of steepest descent which follows.

2.6 The Method of Steepest Descent If the original guess on the initial conditions, $\bar{p}$, is too far from the minimizing values, $\hat{p}$, equations (2.5.2) and (2.5.3) will not apply. In this case, the matrix A_4 of equation (2.4.33) will not be positive-definite. The simpler method of steepest descent may then be used to drive toward what is hopefully the minimizing point. Equations (2.5.2) and (2.5.3) are truncated after first order terms, and written in the form

$$dJ = \dot{J}(T,\bar{p})\, dT + \nabla_p J(T,\bar{p})\, d\bar{p} \qquad (2.6.1)$$

$$d\bar{\Psi} = \dot{\bar{\Psi}}(T,\bar{p})\, dT + \nabla_p \bar{\Psi}(T,\bar{p})\, d\bar{p} \qquad (2.6.2)$$

where

$$dT = \tilde{T}-T \qquad (2.6.3)$$

$$d\bar{p} = \tilde{p} - \bar{p} \qquad (2.6.4)$$

$$dJ = J(\tilde{T},\tilde{p}) - J(T,\bar{p}) \qquad (2.6.5)$$

$$d\bar{\Psi} = \bar{\Psi}(\tilde{T},\tilde{p}) - \bar{\Psi}(T,\bar{p}) \qquad (2.6.6)$$

and $\tilde{T}$, $\tilde{p}$ are better estimates of the minimizing point which is to be determined. Equations (2.6.1) and (2.6.2) may be written as a vector equation

$$d\bar{\Phi} = \nabla\bar{\Phi}\, d\bar{y} \tag{2.6.7}$$

where

$$d\bar{\Phi} = \begin{bmatrix} dJ \\ d\bar{\Psi} \end{bmatrix}, \quad d\bar{y} = \begin{bmatrix} dT \\ d\bar{p} \end{bmatrix} \quad \nabla\bar{\Phi} = \begin{bmatrix} \dot{J} & \nabla_p J \\ \dot{\bar{\Psi}} & \nabla_p \bar{\Psi} \end{bmatrix} \tag{2.6.8}$$

then the change, $d\bar{y}$, is chosen in the gradient direction as

$$d\bar{y} = \nabla'\bar{\Phi}\bar{K} \tag{2.6.9}$$

where $\nabla'\bar{\Phi}$ is the transpose of $\nabla\bar{\Phi}$, and $\bar{K}$ is an $(r+1)x1$ constant vector to be determined. Substitution of $d\bar{y}$ from (2.6.9 into (2.6.7) then gives

$$d\bar{\Phi} = \nabla\bar{\Phi}\ \nabla'\bar{\Phi}\ \bar{K}. \tag{2.6.10}$$

The $(r+1)x(r+1)$ symmetrical matrix, $\nabla\bar{\Phi}\nabla'\bar{\Phi}$, is in general non-singular, but is singular at the minimizing point, $\hat{p},\hat{T}$. Then excluding singular points,

$$\bar{K} = (\nabla\bar{\Phi}\nabla'\bar{\Phi})^{-1}\, d\bar{\Phi} \tag{2.6.11}$$

The required change, from (2.6.9), is thus

$$d\bar{y} = \nabla'\bar{\Phi}(\nabla\bar{\Phi}\ \nabla'\bar{\Phi})^{-1}\, d\bar{\Phi}. \tag{2.6.12}$$

In using this method, appropriate changes, $d\bar{\Phi}$, are selected. This is generally done manually. Then the new estimates of the initial values, $\tilde{p}$, are computed from equations (2.6.12), the second of (2.6.8) and (2.6.4). It should be noted, however, that the method blows up as the minimum point is approached.

2.7 The Partial Derivatives For only the simplest of problems is it possible to compute all of the required partial derivatives explicitly. Those which can always be computed are J, $\dot{J}$, $\bar{\Psi}$ and $\dot{\bar{\Psi}}$, since analytical expressions for these are available. The other partials will, in general, have to be approximated. There may be several ways of doing this. The method given here is suitable for computer solution, and corresponds to the results of differentiating a Lagrange three point interpolation formula.

It is assumed that a solution of the reduced differential equations of the extremals for the initial conditions, $\bar{p}$, is known, which satisfies the stopping condition, equation (2.5.1). This establishes a terminal value of the independent variable, T, as well as values for J, $\dot{J}$, $\bar{\Psi}$ and $\dot{\bar{\Psi}}$. These will be denoted T_o, J_o, $\dot{J}_o$, $\bar{\Psi}_o$ and $\dot{\bar{\Psi}}_o$ respectively. Now suppose that a new solution of the reduced differential equation of the extremals is found for the initial conditions $[p_1 + \delta p_1, p_2, \ldots, p_n]$, where δp_1 is a small positive increment. The stopping condition for this path is

$$t = T_o; \tag{2.7.1}$$

and the value of J, $\dot{J}$, $\bar{\Psi}$ and $\dot{\bar{\Psi}}$ will be denoted $J(p_1 + \delta p_1)$, $\dot{J}(p_1 + \delta p_1)$, $\bar{\Psi}(p_1 + \delta p_1)$ and $\dot{\bar{\Psi}}(p_1 + \delta p_1)$. It is noted that of the n+1 variables T, $\bar{p}$, only p_1 has been changed to obtain these values. A third solution may now be obtained for the initial conditions $[p_1 - \delta p_1, p_2, \ldots, p_n]$ and stopping condition (2.7.1). The values for this path are denoted $J(p_1 - \delta p_1)$, $\dot{J}(p_1 - \delta p_1)$, $\bar{\Psi}(p_1 - \delta p_1)$ and $\dot{\bar{\Psi}}(p_1 - \delta p_1)$. These values will be used to determine the partial derivatives.

Now consider the function J. It may be expanded in a Taylor's series through second order terms to give

$$J(p_1 + \delta p_1) = J_o + \frac{\partial J}{\partial p_1} \delta p_1 + \frac{1}{2} \frac{\partial^2 J}{\partial p_1^2} (\delta p_1)^2 \tag{2.7.2}$$

Similarly,

$$J(p_1 - \delta p_1) = J_o - \frac{\partial J}{\partial p_1} \delta p_1 + \frac{1}{2} \frac{\partial^2 J}{\partial p_1^2} (\delta p_1)^2. \tag{2.7.3}$$

Solution of these equations gives

$$\frac{\partial J}{\partial p_1} = \frac{J(p_1+\delta p_1) - J(p_1-\delta p_1)}{2\delta p_1} \tag{2.7.4}$$

$$\frac{\partial^2 J}{\partial p_1^2} = \frac{J(p_1+\delta p_1) + J(p_1-\delta p_1) - 2J_o}{(\delta p_1)^2} \tag{2.7.5}$$

These equations are generalized to

$$\frac{\partial J}{\partial p_i} = \frac{J(p_i+\delta p_i) - J(p_i-\delta p_i)}{2\delta p_i}, \qquad i=1,\ldots,n \tag{2.7.6}$$

$$\frac{\partial^2 J}{\partial p_i^2} = \frac{J(p_i+\delta p_i) + J(p_i-\delta p_i) - 2J_o}{(\delta p_i)^2}, \qquad i=1,\ldots,n \tag{2.7.7}$$

by perturbing the ith element of $\bar{p}$ by the amount $\pm \delta p_i$. Similar equations can be written for $\dot{J}$, $\bar{\Psi}$ and $\dot{\bar{\Psi}}$. Equations (2.7.6) and their counterparts then establish the required first partial derivatives, $\nabla_p J$, $\nabla_p \bar{\Psi}$. Equations (2.7.7) are used to determine the diagonal elements of $\nabla_p^2 J$ and $\nabla_p^2 \Psi_i$, i=1,...,r. Some of the cross partials, $\nabla_p \dot{J}$, $\nabla_p \dot{\bar{\Psi}}$, are determined by applying equations (2.7.6) to the function $\dot{J}$ and $\dot{\bar{\Psi}}$. The others, of the form $\frac{\partial^2 J}{\partial p_i \partial p_j}$, are determined from the following procedure. If two of the elements of $\bar{p}$, say p_i and p_j, are perturbed by the small positive amounts δp_i, δp_j respectively, the corresponding solu-

tion of the reduced differential equations of the extremals will be $J(p_i+\delta p_i,\ p_j+\delta p_j)$. Taylor's series expansion then gives

$$J(p_i+\delta p_i,\ p_j+\delta p_j) = J_o + \frac{\partial J}{\partial p_i}\delta p_i + \frac{\partial J}{\partial p_j}\delta p_j + \frac{\partial^2 J}{\partial p_i \partial p_j}\delta p_i\,\delta p_j + \frac{1}{2}\frac{\partial^2 J}{\partial p_i^2}(\delta p_i)^2 + \frac{1}{2}\frac{\partial^2 J}{\partial p_j^2}(\delta p_j)^2 \quad (2.7.8)$$

Substitution of equations (2.7.6) and (2.7.7) into (2.7.8) and rearrangement then gives

$$\frac{\partial^2 J}{\partial p_i \partial p_j} = \frac{J(p_i+\delta p_i, p_j+\delta p_j)+J_o-J(p_i+\delta p_i)-J(p_j+\delta p_j)}{\delta p_i\ \delta p_j} \quad (2.7.9)$$

Equations similar to (2.7.9) can be found for either or both of δp_i and δp_j considered negative. It is recommended that cross partials be computed for the perturbations δp_i, δp_j taken in the direction of expected change in the solution. This can be found from the direction of steepest descent.

Since one solution of the reduced differential equations of the extremals is required to determine values for the set J, J_i, J_j, J_{ij}, J_{ii}, J_{jj}, one might ask about the number of solutions required per iteration step. For the method of steepest descent and the fixed terminal point Newton-Raphson method this turns out to be 2n, not counting the initial solution. The general Newton-Raphson method requires $n(\frac{n+3}{2})$ solutions. For small n, say 2 or 3, and moderate computer solution time, this is not unreasonable for the general Newton-Raphson method. For larger n the computer time might become excessive. Under these conditions it might be wise to use the method of steepest descent initially and finish off with the general Newton-Raphson method.

Two tricks might help when n is large. If the second partial derivatives do not change rapidly, they need not be recalculated every step. This is especially true in the vicinity of the optimal point. The required number of solutions then drops to 2n. Secondly, one might try a shift of coordinates to principle axes. This can be accomplished by a translation, a rotation, and finally a stretch to reduce the matrix of second partial derivatives to the identity matrix. Then the cross partials disappear and the required number of solutions is again 2n. An iterative method can be constructed for this process.

2.8 Example Problem I The following simple analytical example illustrates the form (2.3.24)-(2.3.26) of the solution. It

also shows the necessary and sufficient conditions for the equivalent minimization problem. The problem chosen is the two dimensional harmonic oscillator with a single control function. The problem was found in reference (5). The equations of motion are:

$$\frac{dx_1}{dt} = x_2, \qquad x_1(t_o) = x_1^o \tag{2.8.1}$$

$$\frac{dx_2}{dt} = -x_1 + u, \quad x_2(t_o) = x_2^o \tag{2.8.2}$$

It is desired to minimize the control effort over the path. This may be expressed in integral form as:

$$J = \int_{t_o}^{T} u^2 \, d\tau \tag{2.8.3}$$

The terminal surface is specified by a given terminal time and by $x_1(T) = 0$. Thus,

$$\psi_1 = T-K = 0 \tag{2.8.4}$$

$$\psi_2 = x_1(T) = 0 \tag{2.8.5}$$

where T is the value of t on the terminal surface. The Hamiltonian for this system is

$$H = p_1 x_2 + p_2(-x_1 + u) + u^2. \tag{2.8.6}$$

The Euler-Lagrange equations are thus

$$\dot{p}_1 = p_2 \tag{2.8.7}$$

$$\dot{p}_2 = -p_1 \tag{2.8.8}$$

$$p_2 + 2u = 0. \tag{2.8.9}$$

The optimal control function, from (2.8.9), is seen to be

$$u = -\frac{p_2}{2} \tag{2.8.10}$$

This control function automatically satisfies the minimum principle and the Clebsch necessary condition. Equations (2.8.7) and (2.8.8) are linear equations with constant coefficients, and may be solved directly. In terms of initial conditions,

$$\begin{bmatrix} p_1(t) \\ p_2(t) \end{bmatrix} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} p_1^o \\ p_2^o \end{bmatrix} \tag{2.8.11}$$

where p_1^o and p_2^o are the initial conditions on p_1 and p_2.

The homogeneous part of equations (2.8.1) and (2.8.2) is the same as that for equations (2.8.7) and (2.8.8). The system is thus self-adjoint. The solution of equations (2.8.1) and (2.8.2) can then be written as:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} x_1^o \\ x_2^o \end{bmatrix} - \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \int_{t_o}^{t} \begin{bmatrix} \cos \tau & -\sin \tau \\ \sin \tau & \cos \tau \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \frac{p_2(\tau)d\tau}{2} \tag{2.8.12}$$

After substitution and integration, one finds

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} x_1^o \\ x_2^o \end{bmatrix} - \frac{1}{4} \begin{bmatrix} (t\cos t - \sin t), & t \sin t \\ -t \sin t, & (\sin t + t \cos t) \end{bmatrix} \begin{bmatrix} p_1^o \\ p_2^o \end{bmatrix} \tag{2.8.13}$$

where t_o has been taken as zero. Finally, the solution of equation (2.8.3) is

$$J(t) = \frac{1}{8}\left[p_1^{o2}(t - \sin t \cos t) - 2\, p_1^o p_2^o \sin^2 t + p_2^{o2}(t + \sin t \cos t)\right]. \tag{2.8.14}$$

Equations (2.8.11), (2.8.13), and (2.8.14) are the solutions referred to in equations (2.3.24) and (2.3.25).

On the terminal surface, then,

$$J(T) = \frac{1}{8}\left[p_1^{o2}(T - \sin T \cos T) - 2\, p_1^o p_2^o \sin^2 T + p_2^{o2}(T + \sin T \cos T)\right] \tag{2.8.15}$$

$$\begin{bmatrix} x_1(T) \\ x_2(T) \end{bmatrix} = \begin{bmatrix} \cos T & \sin T \\ -\sin T & \cos T \end{bmatrix} \begin{bmatrix} x_1^o \\ x_2^o \end{bmatrix} - \frac{1}{4}\begin{bmatrix} (T\cos T - \sin T), & T \sin T \\ -T \sin T, & (\sin T + T \cos T) \end{bmatrix} \begin{bmatrix} p_1^o \\ p_2^o \end{bmatrix} \tag{2.8.16}$$

The function F of equation (2.4.6) for the equivalent minimization problem may be written, from (2.8.15), (2.8.4) and (2.8.5) as

$$F = J(T) + e_1 (T-K) + e_2 x_1 (T). \tag{2.8.17}$$

The first necessary condition then gives

$$\frac{\partial F}{\partial T} = 0 = \frac{1}{4}(p_1^o \sin T - p_2^o \cos T)^2 + e_1 + e_2 \left\{-x_1^o \sin T + x_2^o \cos T + \frac{1}{4}\left[p_1^o T \sin T - p_2^o (\sin T + T \cos T)\right]\right\} \tag{2.8.18}$$

$$\frac{\partial F}{\partial p_1^o} = 0 = \frac{1}{4}\left[p_1^o (T - \sin T \cos T) - p_2^o \sin^2 T\right] - \frac{e_2}{4}(T \cos T - \sin T) \tag{2.8.19}$$

$$\frac{\partial F}{\partial p_2^o} = 0 = \frac{1}{4}\left[-p_1^o \sin^2 T + p_2^o (T + \sin T \cos T)\right] - \frac{e_2}{4} T \sin T \tag{2.8.20}$$

These, together with (2.8.4) and (2.8.5) are to be solved for T, p_1^o, p_2^o, e_1 and e_2. The solutions are

$$T = K \tag{2.8.21}$$

$$p_1^o = A \cos K \tag{2.8.22}$$

$$p_2^o = A \sin K \tag{2.8.23}$$

$$e_1 = \frac{A\left[x_1^o K \sin K - x_2^o (K \cos K - \sin K)\right]}{(K - \sin K \cos K)} \tag{2.8.24}$$

$$e_2 = A = \frac{4\left[x_1^o \cos K + x_2^o \sin K\right]}{(K - \sin K \cos K)} \qquad (2.8.25)$$

provided $K \neq 0$.

The point defined by equations (2.8.21)-(2.8.24) is truly a minimizing point. This will now be shown by constructing the matrix A_4 of equation (2.4.33). The set (2.8.4) and (2.8.5) is first completed by adding the equation

$$\Psi_3 = x_2 + b\,\zeta. \qquad (2.8.26)$$

Then the equation (2.4.29) may be written, for the critical point as

$$\begin{bmatrix} 1 & 0 & 0 \\ \dot{x}_1 & x_{11} & x_{12} \\ \dot{x}_2 & x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} \eta_o \\ \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \zeta \end{bmatrix} \qquad (2.8.27)$$

where

$$x_{11} = \frac{\partial x_1}{\partial p_1^o}, \quad x_{12} = \frac{\partial x_1}{\partial p_2^o}, \quad x_{21} = \frac{\partial x_2}{\partial p_1^o}, \quad x_{22} = \frac{\partial x_2}{\partial p_2^o}.$$

Equation (2.4.30) then becomes

$$\begin{bmatrix} \eta_o \\ \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \zeta \end{bmatrix} \qquad (2.8.28)$$

where

$$a_{21} = C(x_{12}\dot{x}_2 - x_{22}\dot{x}_2), \quad a_{22} = Cx_{22}, \quad a_{23} = -Cx_{12}$$

$$a_{31} = C(x_{21}\dot{x}_1 - x_{11}\dot{x}_2), \quad a_{32} = -Cx_{21}, \quad a_{33} = Cx_{11}$$

$$\frac{1}{C} = (x_{11}\,x_{22} - x_{21}\,x_{12}).$$

Also,

$$\nabla^2 F = \begin{bmatrix} 0 & \frac{A}{4} K\sin K & -\frac{A}{4}(\sin K + K\cos K) \\ \frac{A}{4} K\sin K & \frac{1}{4}(K - \sin K\cos K) & -\frac{\sin^2 K}{4} \\ -\frac{A}{4}(\sin K + K\cos K) & -\frac{\sin^2 K}{4} & \frac{1}{4}(K + \sin K\cos K) \end{bmatrix} \qquad (2.8.29)$$

After all the substitutions have been made, it is found that the matrix A_4 of equation (2.4.33) is the scalar

$$A_4 = \frac{(K-\sin K\cos K)}{4(K^2-\sin^2 K)} \tag{2.8.30}$$

Since $K > \sin K$ for all $K > 0$, $A_4 > 0$. It is thus concluded that the solution is indeed a minimizing solution, since there is only one critical point.

As a final comment for this problem, it is noted that the minimizing path is uniquely specified by the constants K, x_1^o, and x_2^o.

2.9 Example Problem II The problem considered here is the time optimal bang-bang problem. The objective is to show that at least some problems with corners can be treated by the methods described above. The problem will not be solved. It will only be shown that the function F of the equivalent minimization problem has partial derivatives of at least second order in the variables T, $\bar{p}_o$. It is thus possible to show sufficiency and to use the Newton-Raphson method.

The differential equations for the time optimal bang-bang problem with a single control function are linear, and may be written

$$\dot{\bar{x}} = B\bar{x} + \bar{b}u, \quad \bar{x}(t_o) = \bar{x}_o \tag{2.9.1}$$

where $\bar{b}$ and $\bar{x}$ are nx1 matrices and B is nxn. The elements of $\bar{b}$ and B are constants. The control function is constrained by the equation

$$|u| \leq 1 \tag{2.9.2}$$

which may be rewritten as

$$1 - u^2 \geq 0 \tag{2.9.3}$$

It is recalled that inequality constraints must be in the form $\bar{G} \geq \bar{Z}$ and that continuous partials of at least third order must exist. It is known that there is no optimal solution when $|u| < 1$. The function F for the problem of Bolza, designated L here, may thus be written

$$L = 1 + \bar{p}(B\bar{x} + \bar{b}u - \dot{\bar{x}}) + \mu\,(1-u^2) \tag{2.9.4}$$

The Euler-Lagrange equations are then

$$-\dot{\bar{p}} = \bar{p}\,B \tag{2.9.5}$$

$$0 = \bar{p}\bar{b} - 2\mu\,u. \tag{2.9.6}$$

Equations (2.9.6) and (2.9.3) can be solved for μ and u provided the determinant R_2 of equation (2.3.31) is non-zero. Thus,

$$R_2 = \det\begin{bmatrix} -2\mu & -2u \\ -2u & 0 \end{bmatrix} = -4u^2 \tag{2.9.7}$$

is always non-zero since $|u| = 1$. The proper sign for the control function is found from application of the minimum prin-

ciple. Thus, it is found that

$$\bar{p}\,\bar{b}u \leq \overline{pb}\,U \tag{2.9.8}$$

where U is any admissible control function. It is concluded that

$$u = -\operatorname{sgn}\overline{pb}. \tag{2.9.9}$$

Then from equations (2.9.9) and 2.9.6),

$$\mu = -\frac{1}{2}\left|\bar{p}\,\bar{b}\right|. \tag{2.9.10}$$

It is noted that μ is always negative.

There are two types of subarcs for this problem, corresponding to $u = +1$ and $u = -1$ respectively. If $u = +1$, then $\overline{pb}$ must be negative. The reduced differential equations for this arc may thus be written,

$$\dot{\bar{x}} = B\bar{x} + \bar{b} \tag{2.9.11}$$

$(u = +1)$

$$\dot{\bar{p}} = -\bar{p}B \tag{2.9.12}$$

Furthermore,

$$\mu = \frac{1}{2}\,\bar{p}\,\bar{b}. \tag{2.9.13}$$

When $u = -1$, the differential equations are

$$\dot{\bar{x}} = B\bar{x} - \bar{b} \tag{2.9.14}$$

$(u = -1)$

$$\dot{\bar{p}} = -\bar{p}\,B \tag{2.9.15}$$

and the multiplier equation is

$$\mu = -\frac{1}{2}\,\bar{p}\,\bar{b}. \tag{2.9.16}$$

Now μ is at least absolutely continuous, since $\bar{p}$ is continuous and $\bar{b}$ is constant. Furthermore, since $\mu \leq 0$, it is concluded that the switching points must be the zeros of μ. Thus, the terminal surface equation for each subarc (except for the last one) is

$$0 = \overline{pb}. \tag{2.9.17}$$

So long as $\dot{\bar{p}}\,\bar{b} \neq 0$ at these points, the arguments of article 2.3 hold. It is then concluded that the solution is a continuous function of the initial conditions, $\bar{p}_0$, and that continuous partial derivatives of at least second order in the variables t, $\bar{p}_0$ exist. The time optimal bang-bang problem is thus one problem with corners which can be treated by the methods given above.

2.10 Example Problem III The problem chosen is that of minimizing the deviation of heating rate from a specified heating rate for a re-entry vehicle. This corresponds physically to a radiating vehicle which radiates heat most efficiently at a specified temperature. Mathematically, the quantity to be minimized may be written as an integral over the path,

$$J(T) = \frac{1}{T-t_o} \int_{t_o}^{T} (\dot{Q}-\dot{q})^2 \, d\tau \qquad (2.10.1)$$

where t_o = initial time
T = terminal time
$\dot{Q}$ = specified constant heating rate
$\dot{q}$ = calculated heating rate, a function of the path.

The unit of $J(T)$ is heating rate squared. Thus, if $J(T)$ is minimized, then so is $\left[J(T)\right]^{\frac{1}{2}}$, the heating rate.

An inequality constraint on the problem is that the sensed acceleration never exceed a specified constant. This may be written as

$$a_p \leq B \qquad (2.10.2)$$

where B = specified constant
a_p = instantaneous sensed acceleration, a function of the path.

The model chosen for this problem is two dimensional with a spherical earth and exponential atmosphere. The equations of motion are:

$$\frac{dv}{dt} = \frac{-S}{2m} \rho v^2 C_D - g_o\left(\frac{1}{1+\xi}\right)^2 \sin\gamma \qquad (2.10.3)$$

$$\frac{dv}{dt} = \frac{v \cos\gamma}{R(1+\xi)} + \frac{S}{2m} \rho v C_L - \frac{g_o}{v}\left(\frac{1}{1+\xi}\right)^2 \cos\gamma \qquad (2.10.4)$$

$$\frac{d\xi}{dt} = \frac{v}{R} \sin\gamma \qquad (2.10.5)$$

$$\frac{d\zeta}{dt} = \frac{v \cos\gamma}{(1+\xi)} \qquad (2.10.6)$$

where:
S = reference area (constant), $(ft)^2$
m = vehicle mass (constant), slugs
g_o = value of gravity at sea level, $ft(sec)^{-2}$
R_o = earth radius, ft.
v = velocity, $ft(sec)^{-1}$
γ = flight path angle, positive up, radians
$\xi = \frac{h}{R}$, dimensionless altitude
h = altitude, ft.
$\zeta = R\theta$, great circle range, ft.
$\rho = \rho_o \exp(-\beta R\xi)$, density in slug $(ft)^{-3}$
ρ_o = sea level density
β = atmospheric exponential constant $(ft)^{-1}$

The aerodynamic coefficients are based upon flat plate Newtonian flow. They may be written as:

$$C_L = C_{LO} \sin\alpha \cos\alpha \left|\sin\alpha\right|$$

$$C_D = C_{DO} + C_{DL}|\sin^3 \alpha|$$

where: C_{LO}, C_{DO} and C_{DL} are constants

α = angle of attack, the control function of the problem.

The heating rate equation is assumed to be that given by Chapman: (4)

$$\dot{q} = c\ \rho^{\frac{1}{2}}\ v^3 \qquad (2.10.7)$$

where: c = constant.

It is realized that equation (2.10.7) is an approximation to the convective heating rate, and that radiative heating plays an important role in the super-circular re-entry region. Equation (2.10.7) was chosen to simplify the overall problem. Finally, the pilots', or sensed, acceleration is:

$$a_p = \frac{S\ \rho}{2mg_o}\ v^2\ \sqrt{C_L^2 + C_D^2} \qquad (2.10.8)$$

To completely specify the nature of the solutions, one must provide initial conditions for the differential equations (2.10.3)-(2.10.6) as well as terminal conditions. It is assumed that initial conditions are known, and that terminal conditions can be written in the form:

$$\Psi_i(\bar{x}(T),\ T) = 0, \qquad i = 1,\ \ldots,\ q \leq n+1 \qquad (2.10.9)$$

where: $\bar{x}(T)$ has components $v(T)$, $\gamma(T)$, $\xi(T)$ and $\zeta(T)$
n = number of components of $\bar{x}(T) = 4$.

Equations (2.10.9) represent surfaces in phase space. The intersection of these surfaces is a surface upon which the terminal point of the optimal path must lie. Hence, equations (2.10.9) are said to describe the terminal surface. If $q=n+1$, then the problem is a fixed end-point problem. If $q = n$ then the terminal surface is a curve. For $q < n$, (2.10.9) describes a surface in phase space.

Initial conditions for this problem are assumed to be in the super-circular velocity region. The terminal surface equations are

$$\Psi_1 = a_1 - T = 0 \qquad (2.10.10)$$

$$\Psi_2 = a_2 - v(T) = 0 \qquad (2.10.11)$$

$$\Psi_3 = a_3 - \gamma(T) = 0 \qquad (2.10.12)$$

$$\Psi_4 = a_4 - \xi(T) = 0 \qquad (2.10.13)$$

$$\Psi_5 = a_5 - \zeta(T) = 0 \qquad (2.10.14)$$

where the constants a_1, a_2, a_3, a_4 and a_5 are specified values

in the sub-circular velocity region. Various combinations of equations (2.10.10)-(2.10.14) can be used to obtain different terminal surfaces, and hence, different optimal solutions.

The problem may now be stated: Find that programming of the control function, $\alpha(t)$, such that the solution of the differential equations (2.10.3)-(2.10.6) for given initial conditions satisfies the inequality constraint (2.10.2), ends on the terminal surface defined by the chosen combination of equations (2.10.10)-(2.10.14), and yields the minimum value for the function (2.10.1) (if a solution exists).

To facilitate the problem solution, a new variable will be defined by the differential equation

$$\dot{x}_5 = (\dot{Q}-\dot{q})^2, \quad x_5(t_o) = 0 \tag{2.10.15}$$

This is added to the set (2.10.3)-(2.10.6). Equation (2.10.1) may then be written

$$J(T) = \frac{x_5(T)}{T-t_o} . \tag{2.10.16}$$

The solution is assumed to consist of two types of subarcs, one for which $a_p < B$ (except for possibly a finite number of points) and the other for which $a_p = B$. The function F of the multiplier rule can then be written in the abbreviated form

$$F = \lambda_1(f_1-\dot{v}) + \lambda_2(f_2-\dot{\gamma}) + \lambda_3(f_3-\dot{\xi}) + \lambda_4(f_4-\dot{\zeta})$$

$$+\lambda_5(f_5-\dot{x}_5) + \mu'(B-a_p), \tag{2.10.17}$$

where it is understood that the multiplier μ' is zero whenever $a_p < B$, and $\leq$ zero when $a_p = B$. The coefficients of the multipliers, λ_i, are the differential equations (2.10.3)-(2.10.6) and (2.10.15). f_i, i=1, ..., 5, represents the right-hand sides of these differential equations. The Euler-Lagrange equations for (2.10.17) can then be written

$$-\frac{d\lambda_1}{dt} = \lambda_1 \frac{\partial f_1}{\partial v} + \lambda_2 \frac{\partial f_2}{\partial v} + \lambda_3 \frac{\partial f_3}{\partial v} + \lambda_4 \frac{\partial f_4}{\partial v}$$

$$+ \lambda_5 \frac{\partial f_5}{\partial v} - \mu' \frac{\partial a_p}{\partial v} \tag{2.10.18}$$

$$-\frac{d\lambda_2}{dt} = \lambda_1 \frac{\partial f_1}{\partial \gamma} + \lambda_2 \frac{\partial f_2}{\partial \gamma} + \lambda_3 \frac{\partial f_3}{\partial \gamma} + \lambda_4 \frac{\partial f_4}{\partial \gamma} \tag{2.10.19}$$

$$-\frac{d\lambda_3}{dt} = \lambda_1 \frac{\partial f_1}{\partial \xi} + \lambda_2 \frac{\partial f_2}{\partial \xi} + \lambda_4 \frac{\partial f_4}{\partial \xi}$$

$$+ \lambda_5 \frac{\partial f_5}{\partial \xi} - \mu' \frac{\partial a_p}{\partial \xi} \tag{2.10.20}$$

$$-\frac{d\lambda_4}{dt} = 0 \qquad \Rightarrow \lambda_4 = C_1 \tag{2.10.21}$$

$$-\frac{d\lambda_5}{dt} = 0 \qquad \Rightarrow \lambda_5 = C_2 \tag{2.10.22}$$

$$\lambda_1 \frac{\partial f_1}{\partial \alpha} + \lambda_2 \frac{\partial f_2}{\partial \alpha} - \mu' \frac{\partial a_p}{\partial \alpha} = 0 \tag{2.10.23}$$

The multiplier λ_4 of equation (2.10.21) is a constant because of the original choice of coordinate system for the differential equations. Polar coordinates always give a cyclic, or ignorable, coordinate. This is the angle θ, which never appears on the right-hand sides of the differential equations. In Cartesian coordinates, there would be four variable multipliers instead of three. It is advisable to use that coordinate system which yields the greatest number of cyclic coordinates. The corresponding multipliers are constants. λ_5 is a constant because of the definition of x_5. Its value, from equation (2.10.16) and the transversality condition (2.3.13), is found to be

$$\lambda_5 = \frac{1}{T-t_o} . \tag{2.10.24}$$

This follows since $x_5(T)$ is not present in the terminal surface equations (2.10.10)-(2.10.14) and since $J \equiv g$ in equation (2.2.1). The Hamiltonian for this problem can be written

$$H = \sum_{i=1}^{5} p_i f_i = \frac{J}{T-t_o} - e_1' . \tag{2.10.25}$$

It is a constant, and its value (the right-hand side of (2.10.25)) is found from equations (2.10.16), (2.10.10) and the transversality conditon (2.3.12). The constant e_1' is zero if equation (2.10.16) is not present for the problem under consideration. Finally, terminal values for the other multipliers are seen to be

$$\lambda_1(T) = e_2' \tag{2.10.26}$$

$$\lambda_2(T) = e_3' \tag{2.10.27}$$

$$\lambda_3(T) = e_4' \tag{2.10.28}$$

$$\lambda_4 = e_5' . \tag{2.10.29}$$

Any one or all of these are zero if the corresponding equations (2.10.11)-(2.10.14) are absent from the problem formulation.

It is convenient to normalize the multipliers at this point. This will be done by multiplying each multiplier by the time interval, $T-t_o$, viewed as a constant. Thus

$$p_i = (T-t_o)\,\lambda_i, \qquad i=1,\ldots,5 \tag{2.10.30}$$

$$e_j = (T-t_o)\,e_j', \qquad j=1,\ldots,r \tag{2.10.31}$$

$$\mu = (T-t_o)\,\mu'. \tag{2.10.32}$$

Then equations (2.10.18)-(2.10.20), (2.10.23), and (2.10.24)-(2.10.29) become

$$-\dot{p}_1 = p_1 \frac{\partial f_1}{\partial v} + p_2 \frac{\partial f_2}{\partial v} + p_3 \frac{\partial f_3}{\partial v} + p_4 \frac{\partial f_4}{\partial v} + \frac{\partial f_5}{\partial v} - \mu \frac{\partial a_p}{\partial v} \tag{2.10.33}$$

$$-\dot{p}_2 = p_1 \frac{\partial f_1}{\partial \gamma} + p_2 \frac{\partial f_2}{\partial \gamma} + p_3 \frac{\partial f_3}{\partial \gamma} + p_4 \frac{\partial f_4}{\partial \gamma} \tag{2.10.34}$$

$$-\dot{p}_3 = p_1 \frac{\partial f_1}{\partial \xi} + p_2 \frac{\partial f_2}{\partial \xi} + p_4 \frac{\partial f_4}{\partial \xi} + \frac{\partial f_5}{\partial \xi} - \mu \frac{\partial a_p}{\partial \xi} \tag{2.10.35}$$

$$0 = p_1 \frac{\partial f_1}{\partial \alpha} + p_2 \frac{\partial f_2}{\partial \alpha} - \mu \frac{\partial a_p}{\partial \alpha} \tag{2.10.36}$$

$$H_1 = \sum_{i=1}^{5} p_i f_i = J + e_1 \tag{2.10.37}$$

$$p_i(T) = e_{i-1}, \qquad i=1,\ldots,4 \tag{2.10.38}$$

$$p_5 = 1. \tag{2.10.39}$$

Computational equations will now be developed for the control function, α, and the multiplier μ. Consider first those subarcs for which $a_p < B$. Over these subarcs $\mu = 0$. Then equation (2.10.36) becomes

$$p_1 \frac{\partial f_1}{\partial \alpha} + p_2 \frac{\partial f_2}{\partial \alpha} = 0, \tag{2.10.40}$$

or

$$p_1 \frac{\partial f_1}{\partial C_D} \frac{\partial C_D}{\partial \alpha} + p_2 \frac{\partial f_2}{\partial C_L} \frac{\partial C_L}{\partial \alpha} = 0. \tag{2.10.41}$$

From the aerodynamic coefficients one finds

$$\frac{\partial C_D}{\partial \alpha} = 3C_{DL} \sin\alpha \cos\alpha |\sin\alpha| \tag{2.10.42}$$

$$\frac{\partial C_L}{\partial \alpha} = C_{LO} |\sin \alpha| (3 \cos^2 \alpha - 1). \qquad (2.10.43)$$

and

$$\frac{\partial f_1}{\partial C_D} = - \frac{S}{2m} \rho v^2 \qquad (2.10.44)$$

$$\frac{\partial f_2}{\partial C_L} = \frac{S}{2m} \rho v \qquad (2.10.45)$$

from equations (2.10.3) and (2.10.4). Substituting these into (2.10.41) and rearranging then gives

$$\frac{S}{2m} \rho v |\sin \alpha| \left[-3 C_{DL} p_1 v \sin \alpha \cos \alpha + C_{LO} p_2 (3 \cos^2 \alpha - 1)\right] = 0. \qquad (2.10.46)$$

Assuming re-entry conditions, the leading terms of (2.10.46) are non-zero if $|\sin \alpha| \neq 0$. The bracketed terms may then be rewritten in the form

$$\tan^2 \alpha - \frac{3C_{DL}}{C_{LO}} \frac{p_1 v}{p_2} \tan \alpha - 2 = 0. \qquad (2.10.47)$$

One is thus led to the equation

$$\tan \alpha = -a \pm \sqrt{a^2 + 2} \qquad (2.10.48)$$

where

$$a = \frac{3C_{DL}}{2C_{LO}} \frac{p_1 v}{p_2} .$$

It is noticed that the sign of $\tan \alpha$ is determined by the choice of the $\pm$ sign in equation (2.10.48).

The lift-drag polar is traversed once as α ranges through π radians. It is thus advisable to limit the control function to a range of π radians to avoid double values. Furthermore, the points $\alpha = 0, \pi$ are singular points since $\frac{\partial^2 F}{\partial \alpha^2}$ (the determinant of equation (2.3.21) is zero at these points. The range $0 < \alpha < \pi$ is chosen here since then $|\sin \alpha| = \sin \alpha$, and the proper sign for equation (2.10.48) is most easily chosen. The singular points are removed by defining $\alpha = 0$ at these points. Thus, the range of α is

$$0 \leq \alpha < \pi . \qquad (2.10.49)$$

Application of the minimum principle gives the equation

$$\sin \alpha \left[-C_{DL} p_1 v \sin^2 \alpha + C_{LO} p_2 \sin \alpha \cos \alpha\right] \leq \sin A \left[-C_{DL} p_1 v \sin^2 A + C_{LO} p_2 \sin A \cos A\right] \qquad (2.10.50)$$

where A is any value of the control function satisfying equa-

tions (2.10.49) and (2.10.2). Substitution of (2.10.47) into (2.10.50) and rearranging then gives

$$p_2 \sin\alpha \tan\alpha \leq p_2 \sin A \tan A \qquad (2.10.51)$$

Since $\sin\alpha \geq 0$ and the sign of $\tan\alpha$ is determined by the $\pm$ sign of equation (2.10.48), it is seen that (2.10.51) is satisfied by the choices

If $p_2 > 0$, choose - sign $(\frac{\pi}{2} < \alpha < \pi)$

If $p_2 < 0$, choose + sign $(0 < \alpha < \frac{\pi}{2})$.

Equation (2.10.48) thus becomes

$$\tan\alpha = -a - (\operatorname{sgn} p_2)\sqrt{a^2 + 2^T} \qquad (2.10.52)$$

The behavior where $p_2 = 0$ will now be examined. This will be done as a limiting process, since p_1, p_2 and v are continuous functions of time. The binomial series expansion, assuming large a, gives

$$\sqrt{a^2 + 2} = |a| + \frac{1}{|a|}$$

It is seen that:

If $a > 0$ and the minus sign were chosen, then

$$\tan\alpha = \lim_{a\to\infty} -2|a|, \qquad \alpha \to \frac{\pi}{2} \text{ (from the left)}.$$

If $a < 0$ and the plus sign chosen,

$$\tan\alpha = \lim_{-a\to\infty} 2|a|, \qquad \alpha \to \frac{\pi}{2} \text{ (from the right)}.$$

On the other hand, if $a < 0$ and the minus sign chosen,

$$\tan\alpha = \lim_{-a\to\infty} -\frac{1}{|a|} = 0, \quad \alpha \to \pi \text{ (from the right)}.$$

Finally, with $a > 0$ and the plus sign,

$$\tan\alpha = \lim_{a\to\infty} \frac{1}{|a|} = 0, \quad \alpha \to 0 \text{ (from the left)}.$$

The last two cases show that it is possible for α to be discontinuous. For example, if p_2 passes through zero and α was near π, then α will jump to zero as p_2 goes through zero. This difficulty is cleared up if the proper interpretation is made. Since $\tan(\pi + \alpha) = \tan\alpha$, α may be thought of as being continuous across jumps, but in the disallowed region $\pi < \alpha < 2\pi$. The control function discontinuity will be retained since computer results can be properly interpreted and since it allows $|\sin\alpha| = \sin\alpha$ in all equations. Finally, although of little practical importance, if p_1 and p_2 are simultaneously zero at a point, the limit ratio p_1/p_2 is $\dot{p}_1/\dot{p}_2$ at that point by L'Hospital's rule.

To sum up, the reduced differential equations of the extremals for subarcs having $a_p < B$ are (2.10.3)-(2.10.6), (2.10.15) and (2.10.33)-(2.10.35) with $\mu \equiv 0$. The control function is computed from equation (2.10.52) in the range $0 \leq \alpha < \pi$. The subarc terminal surface is seen to be $a_p = B$, provided that $\dot{a}_p \neq 0$.

When $a_p = B$, one must calculate both α and μ. The method chosen for doing this here is to find α from $a_p = B$ and μ from (2.10.36). Expansion and rearrangement of the constraint equation $a_p = B$ gives:

$$b^2 - C_{DO}^2 = \sin^3\alpha\left[c_1 \sin^3\alpha + c_2 \sin\alpha + c_3\right] \qquad (2.10.53)$$

where:

$$b = \frac{2mBg_o}{S\rho v^2}$$

$$c_1 = C_{DL}^2 - C_{LO}^2$$

$$c_2 = C_{LO}^2$$

$$c_3 = 2\, C_{DO}\, C_{DL}.$$

The right hand side of (2.10.53) is a function only of the vehicle aerodynamic coefficients (constants) and $\sin\alpha$. It is zero when $\alpha = 0$ or π, and maximum when $\alpha = \frac{\pi}{2}$. From this it follows that b must satisfy

$$C_{DO} \leq b \leq (C_{DL} + C_{DO}) \qquad (2.10.54)$$

Physically speaking, if $C_{DO} > b$ at any point of the path, then α has gone to zero in the futile attempt to keep $a_p = B$. The path must then be disqualified as a candidate for the optimum path.

Newton's method was chosen to extract α from (2.10.53). The iteration equation is

$$\Delta\alpha = \frac{(b^2 - C_{DO}^2) - \sin^3\alpha\left[c_1\sin^2\alpha + c_2\sin\alpha + c_3\right]}{\sin^2\alpha\cos\alpha\left[6c_1\sin^2\alpha + 4c_2\sin\alpha + 3c_3\right]} \qquad (2.10.55)$$

A $\Delta\alpha$ is calculated from (2.10.55) using an assumed α (usually the last integration step value). This is added to α, and the new α is used to calculate a new $\Delta\alpha$. The process continues until $\Delta\alpha$ is negligible.

In the vicinity of $\alpha = 0, \frac{\pi}{2}$ or π, the denominator of (2.10.55) is likely to be quite small. To avoid this difficulty equation (2.10.53) was expanded as a series. The results are:

$$\text{If } \alpha \sim 0, \left(\sin\alpha = \alpha - \frac{\alpha^3}{3!}\right) \text{ then } \alpha = \left(\frac{b^2 - C_{DO}^2}{c_3}\right)^{1/3} \qquad (2.10.56)$$

$$\text{If } \alpha \sim \pi, \quad \left(\sin\alpha = (\pi-\alpha) - \frac{(\pi-\alpha)^3}{3!}\right) \tag{2.10.57}$$

$$\text{then } \alpha = \pi - \left(\frac{b^2 - C_{DO}^2}{c_3}\right)^{1/3}$$

$$\text{If } \alpha \sim \frac{\pi}{2}, \quad \left(\cos\alpha = \left(\frac{\pi}{2} - \alpha\right) - \frac{1}{3!}\left(\frac{\pi}{2} - \alpha\right)^3\right) \tag{2.10.58}$$

$$\text{then } \alpha = \frac{\pi}{2} \pm \left[\frac{(C_{DL} + C_{DO})^2 - b^2}{3C_{DL}(C_{DL}+C_{DO}) - C_{LO}^2}\right]^{1/2}$$

where: use + sign if $p_2 > 0$

use - sign if $p_2 < 0$.

If $p_2 = 0$, choose sign so that tan α is continuous. The choice of signs again comes from the minimum principle.

Finally, μ is calculated from equation (2.10.36) in the form:

$$\mu = \frac{1}{\left(\frac{\partial a_p}{\partial \alpha}\right)}\left[p_1 \frac{\partial f_1}{\partial \alpha} + p_2 \frac{\partial f_2}{\partial \alpha}\right]. \tag{2.10.59}$$

Now $\frac{\partial a_p}{\partial \alpha}$ may be written in the form:

$$\frac{\partial a_p}{\partial \alpha} = B \frac{\left[C_D \frac{\partial C_D}{\partial \alpha} + C_L \frac{\partial C_L}{\partial \alpha}\right]}{C_L^2 + C_D^2} \tag{2.10.60}$$

where:

$$C_D \frac{\partial C_D}{\partial \alpha} + C_L \frac{\partial C_L}{\partial \alpha} = \frac{1}{2}\sin^2\alpha \cos\alpha\left[6c_1\sin^3\alpha + 4c_2\sin\alpha + 3c_3\right]. \tag{2.10.61}$$

This is proportional to the denominator of equation (2.10.55).

It is also noted that the determinant R_2 of equation (2.3.31) is

$$R_2 = \left(\frac{\partial a_p}{\partial \alpha}\right)^2 \tag{2.10.62}$$

Since this must never be zero when $a_p = B$, it is seen that α must never be 0, $\frac{\pi}{2}$ or π.

The reduced differential equations of the extremals for the subarcs for which $a_p = B$ are equations (2.10.3)-(2.10.6),

(2.10.15) and (2.10.33)-(2.10.35) with $\mu \leq 0$. The control function is determined from the iterative equation (2.10.55) or the small angle equations (2.10.56)-(2.10.58) in the range (2.10.49), except $\alpha \neq 0, \frac{\pi}{2}$. The multiplier μ is determined from equation (2.10.59). It is easily seen that μ is continuous, and that the subarc terminal surface is, consequently, $\mu = 0$, provided $\dot{\mu} \neq 0$.

Potential corner points will now be examined. First, the points $\alpha = 0, \frac{\pi}{2}$ must be ruled out as candidates for these points since both R_1 and R_2, equations (2.3.21) and (2.3.31), must be non-singular at potential corner points. Then for all other values in the range (2.10.49) it turns out that α must be continuous at these points. This is seen most easily by looking at the Hamiltonian. It is a continuous function of $\bar{x}$, $\bar{p}$ and α. Since x and p are continuous at potential corner points, then α must be continuous at those points. Similarly, examination of equation (2.10.36) shows that μ must be continuous at those points. Thus, μ must start and end with the value zero on those subarcs for which $a_p = B$.

Finally, from sections 2.3 and 2.4, it is seen that this problem can be solved as an initial value problem.

2.11 A Special Form of the Newton-Raphson Method Several optimal trajectories have been obtained for the above problem using the Newton-Raphson method. The terminal surface for all of these has been described by equation (2.10.10). For this terminal surface only, the Newton-Raphson method reduces to an especially nice form. This will now be developed.

The function F for the equivalent minimization problem with the constraint equation (2.10.10) can be written

$$F = J(\hat{T}, \hat{p}) + e_1(a_1 - \hat{T}). \tag{2.11.1}$$

The values T, $\hat{T}$, $\bar{p}$, $\hat{p}$ are defined as in section 2.5. Then using equation (2.5.2) and the first necessary condition for the equivalent minimization problem, one finds

$$\frac{\partial F}{\partial \hat{T}} = 0 = \dot{J} + \nabla_p \dot{J}\,(\hat{p}-\bar{p}) - e_1 \tag{2.11.2}$$

$$\nabla_{\hat{p}} F = Z = \nabla_p J + \nabla_p^2 J\,(\hat{p}-\bar{p}) \tag{2.11.3}$$

$$0 = a_1 - \hat{T} \tag{2.11.4}$$

These are partially solved by

$$\hat{T} = a_1 \tag{2.11.5}$$

$$e_1 = \dot{J} + \nabla_p \dot{J}\,(\hat{p}-\bar{p}) \; . \tag{2.11.6}$$

The problem then reduces to solving (2.11.3) for $\hat{p}$ subject to the sufficiency condition (2.4.21).

The sufficiency condition will now be reduced to the equivalent from (2.4.33) by completing the set of constraint equations. The constraint equations need only be independent functions of $\hat{T}$ and $\hat{p}$. Consequently, the equations (2.4.23) may be chosen as

$$\bar{p} = \hat{p} + b\,\zeta \tag{2.11.7}$$

Then equation (2.4.30) becomes

$$\begin{bmatrix} \eta_o \\ \bar{\eta} \end{bmatrix} = I \begin{bmatrix} 0 \\ \zeta \end{bmatrix} \tag{2.11.8}$$

where I is the (n+1)x(n+1) identity matrix. Substitution of (2.11.8) into (2.4.31) then gives, for the form (2.4.33),

$$\zeta' \nabla_p^2 J\, \zeta \geq 0. \tag{2.11.9}$$

It is thus seen that the matrix $\nabla_p^2 J$ must be positive-definite. The special form of the Newton-Raphson method, then, consists of the iterative solution of equation (2.11.3) provided the matrix $\nabla_p^2 J$ of that equation is positive-definite.

2.12 Experiences in the Use of the Newton-Raphson Method The problem of section 2.10, as stated above, has been solved using the special form of the Newton-Raphson method. The overall performance of the method has been very satisfactory. Once the initial values, $\bar{p}$, are close enough to the minimizing values, $\hat{p}$, the method converges quite rapidly to the minimizing point.

There are certain problems involved in the use of the Newton-Raphson method, as there are with any other optimizing scheme. The first of these is the number of solutions of the reduced differential equations of the extremals required per Newton-Raphson step. As stated in section 2.7, $\frac{n(n+3)}{2}$ solutions are required. Gradient schemes, (6,8) on the other hand, require one forward and one backward solution per step. However the convergence properties of the Newton-Raphson method theoretically become better and better as the minimum point is approached, whereas gradient schemes tend to blow up. It is an open question, then, as to whether one method requires more total computer time than the other. Furthermore, one is never certain that a relative minimum has been obtained using gradient techniques.

The biggest problem involves the integration accuracy. Since the partial derivatives are approximated by differencing integrated solutions, the random error generated by the integration algorithm may result in bad predictions of the minimizing point. The value of the Hamiltonian, theoretically a constant, may be used as a measure of the goodness of the solution. It should be remembered that any integration algorithm will blow up after a sufficient number of integration steps.

Closely related to the above problem is that of picking the perturbations, $\delta\bar{p}$, of article 2.7. If these are too small,

the partial derivatives will reflect integration error alone. If they are too large, the partial derivatives will not reflect the nature of the surface in the vicinity of the point $\bar{T}$, $\bar{p}_o$. Intermediate values must be picked from experience.

Solution of the problem of article 2.10 was originally attempted using Bryson's method of steepest descent.[8] The terminal surface in this case was

$$\Psi_1 = v^2 - \frac{\mu}{r} = 0 \qquad (2.12.1)$$

This equation describes circular velocity at a given altitude. The problem turned out to be extremely sensitive for this terminal surface. Nevertheless, one trajectory, thought to be fairly close to optimum, was obtained. When the Newton-Raphson method was used for the same problem, the steepest descent value of J (equation (2.10.1)) was halved before the sensitivity problem became severe. Furthermore, the amount of human intervention required for the Newton-Raphson method was much less than that for the method of steepest descent. It is the writer's opinion, considering the problems involved with both methods, that the Newton-Raphson scheme is by far the more superior method.

REFERENCES

1. Berkovitz, L. D., "Variational Methods in Problems of Control and Programming", Journal of Mathematical Analysis and Applications, Vol. 3, No. 1, August 1961, pp. 145-169.
2. Bliss, G. A., Lectures on the Calculus of Variations, The University of Chicago Press, Chicago, 1946.
3. Breakwell, J. V., "The Optimization of Trajectories", J. Soc. Industr. Applied Mathematics 7, 215, 1959.
4. Chapman, D. R., "An Analysis of the Corridor and Guidance Requirements for Supercircular Entry into Planetary Atmospheres", NASA, TR R-55, 1959.
5. Kalman, R. E., "The Theory of Optimal Control and the Calculus of Variations", RIAS Technical Report 61-3.
6. Kelley, H. J., "Gradient Theory of Optimal Flight Paths", Jour. Amer. Rocket Soc., Vol. 30, No. 10, pp. 947-953, October, 1960.
7. Valentine, F. A., "The Problem of Lagrange with Differential Inequalities as Added Side Conditions", Contributions to the Theory of Calculus of Variations, pp. 403-447, 1933-1937, University of Chicago.
8. Bryson, A. E., Denham, W. F., Carroll, F. J., and Mikamic, K., "Determination of the Lift or Drag Program that Minimizes Re-Entry Heating with Acceleration or Range Constraints Using a Steepest Descent Computation Procedure" Paper given at the January, 1961, IAS meeting. This paper with abbreviated title will also be found in the Journ. of Aerospace Sci., Vol. 29, No. 4, April 1962, pp.420-430.

ASD-TDR-63-119

Application of Optimal Linear Control Theory to the Design of Aerospace Vehicle Control Systems

By

P. A. Reynolds
E. G. Rynaski

Cornell Aeronautical Laboratory, Inc.
Buffalo, New York

APPLICATION OF OPTIMAL LINEAR CONTROL THEORY TO THE DESIGN OF AEROSPACE VEHICLE CONTROL SYSTEMS

Philip A. Reynolds and Edmund G. Rynaski

Cornell Aeronautical Laboratory, Inc.
Buffalo 21, New York

INTRODUCTION

This paper describes some of the results of a current research project* to apply optimal control techniques to the design of flight control systems. The project will be completed in October 1962, and a final report containing the complete findings will be published soon thereafter.

The area of optimal control theory of primary interest in this study has been the optimal regulator problem for a constant linear plant with a performance index extended over an arbitrarily large time interval. The optimal control for this problem is a set of feedback loops with constant gains if the control input amplitude is not constrained. (Although the control is not restricted, the integral of control input squared is minimized.) This solution is of interest to flight control because constant feedback gains are easy to mechanize and because computer solution of the optimal regulator problem under these conditions is straightforward (non-iterative). It is also of interest as a possible technique for automatic digital computer synthesis of multi-input, multi-output linear systems of high dimension.

The paper briefly outlines the theory, discusses the flight control application in general and then presents some results on two specific problems: 1) the problem of obtaining adequate lateral handling qualities in an X-15-type aircraft, and 2) the attitude control problem in large, flexible boosters.

The optimal regulator problem for constant linear plants has been solved by S. S. L. Chang and others for the single-input, single-output case. R. E. Kalman, at the Martin Company's Research Institute for Advanced Studies, has solved the general linear case with a constant or time-varying plant. The authors are indebted to Dr. Kalman for his advice during this program and to RIAS for the use of their IBM 7090 computer program.

THEORY[1]

A constant, linear plant in state-vector form is represented as:

$$\frac{dx}{dt} = Fx + Gu \qquad (1)$$

* Sponsored by the Flight Control Laboratory, Aeronautical Systems Division, USAF - Lt. Edwin Stear, ASD Project Engineer.

where x is the n-dimensional state vector of the plant, u is the m-dimensional control vector, and F (dimension nxn) and G (dimension nxm) are constant matrices. There are no constraints on x or u.

Define the output of the plant as

$$y = Hx \tag{2}$$

where y is a p-dimensional vector and H is a constant matrix of dimension pxn. (The output can have more than one interpretation – vector of measurable quantities, error vector for the regulator problem, arbitrary linear function of the state, etc.).

Define a performance index as

$$2V = \lim_{T \to \infty} \left\{ x(T), Sx(T) + \int_t^T [y(\tau), Qy(\tau) + u(\tau), Ru(\tau)] d\tau \right\} \tag{3}$$

where t = present time
T = terminal time
S = non-negative definite symmetric matrix (nxn)
Q = positive definite symmetric matrix (pxp)
R = positive definite symmetric matrix (mxm)

and the notation a, Aa denotes the scalar product of the vectors a and and Aa. The forms $x(T)$, $Sx(T)$ and $y(\tau)$, $Qy(\tau)$ and $u(\tau)$, $Ru(\tau)$ are quadratic forms in the terminal state, the output and the control input respectively.

The optimal regulator problem is the problem of finding u as a function of x such that the plant moves from some initial x to the point x = 0 by the path that minimizes V. This problem was solved by Kalman[2] using the Hamiltonian calculus of variations and the work of Carathéodory. The result is briefly stated as follows:

Assuming that a minimum of V exists and that it can be written as:

$$2V_{min} = x(t), Px(t), \tag{4}$$

we have

$$u_{opt}(t) = -Kx(t) = -R^{-1}G'Px(t). \tag{5}$$

P is the positive-definite symmetric solution of the matrix equation

$$F'P + PF - PGR^{-1}G'P + H'QH = 0 \tag{6}$$

[V_{min} satisfies the Hamilton-Jacobi differential equation, and equation (6) is the steady state of the matrix Riccati equation,

$$\frac{-dP}{dt} = F'P + PF - PGR^{-1}G'P + H'QH,$$ that results when

V_{min} is substituted into the Hamilton-Jacobi equation.]

Kalman has shown that the solution exists whenever the plant is completely controllable, this quality being defined as the ability of finite control

inputs to return the plant from any initial condition to the origin in finite time. The solution is unique if the system is completely observable, this quality being defined as the ability to determine the state vector at present time by observing y and its past values. It should be emphasized that these are sufficient conditions for existence and uniqueness. Solutions have been computed even when the plant lacks controllability and observability. However, the uniqueness of these solutions has not been investigated.

Substituting u_{opt} for u in equation (1), we find that the closed-loop optimal system is described by the equation

$$\frac{dx}{dt} = (F-GK)x \equiv \hat{F}x \tag{7}$$

This system is asymptotically stable in the large, since V_{min} is a Lyapunov function.

Some examples of optimal regulators are illustrated in Figure 1. These examples are all single-input, single-output, second-order plants. Therefore the Q and R matrices in the performance index are scalars denoted as q and r. Since one of the scalars can be factored outside the integral sign in equation (3), there is only one parameter, q/r, that expresses the criterion of performance. By increasing q/r, we more heavily weight errors in output over control input action. Figure 1 shows the loci of the closed-loop poles as q/r is increased. These plots can be generated in two ways. First, the roots of $\hat{F}$ can be computed for each q/r by computing P (and thus K) from the Riccati equation. Secondly, the work of Chang[3] on single-input, single-output systems (root-square-locus technique) can be applied. The right-half plane images of the plant poles and zeros are added to the actual plant open-loop roots, and a standard root locus plot is constructed. The left-half plane poles of this root locus plot are then the closed-loop poles of the optimal system. The performance parameter q/r is related to the gain parameter of the plot. (The actual feedback parameters are represented by a matrix of loop gains.)

Using this second technique, it is relatively easy to examine the characteristics of optimal single-input, single-output systems. It is readily apparent how the optimal system will behave as a function of the performance index being minimized. At this point we can say that:

1. Not every stable linear system optimizes a quadratic index of the form of equation (3). The process of optimization is truly selective.
2. Not every optimal system is of high bandwidth and adequate damping (witness the pole positions for low q/r).
3. As $q/r \to \infty$ the optimal closed-loop poles not cancelled by zeros approach a Butterworth configuration (poles symmetrically arranged on a semicircle whose center is the origin).

GENERAL REMARKS ON FLIGHT CONTROL APPLICATION

The equations of motion of an aerospace vehicle are nonlinear. However, the motions of these vehicles can usually be adequately described by linearized equations if the perturbations from equilibrium are small. If the equilibrium conditions change due to changes in altitude, airspeed, gross weight, or other similar flight parameters, the coefficients in the linearized equations also change. If it can be assumed that the vehicle will not stray very far from some precalculated flight trajectory, then these coefficients can be specified by functions of time only. Frequently these coefficients change so slowly that they can be considered constant. It is therefore evident that the plant in this case fits the general linear optimal control theory and often the theory for a constant linear plant.

However, for optimal control to be meaningful, we must attach some physical meaning to the performance index. The error portion of the index is not difficult to justify. In the attitude control problem, we would like to minimize errors in attitude from some reference. In the handling qualities problem, we have a set of desirable closed-loop dynamic characteristics. We can represent these characteristics by a model. Thus using Kalman's model formulation of the performance index described below, we can effectively minimize the errors between model and system outputs. The second piece of the performance index requiring minimum control inputs is not as easily inte preted. In many applications, we don't really care to minimize control input as long as the control input and rate limits are not exceeded. Minimizing control deflections helps to accomplish this, but the requirement of minimization is basically too strong. We do not desire minimum control input – only input smaller than a given level.

The real weakness in the performance index as a meaningful criterion of performance is the fact that the choice of relative weighting between the output error and the control input is arbitrary. It has already been pointed out that the closed-loop poles are sensitive functions of the parameter q/r for a single-input, single-output system. The same is qualitatively true of the parameter $||Q||/||R||$ for multidimensional input and output ($||A|| \equiv$ norm of A).

This point is illustrated by the simple example of an aircraft pitch axis control system. The linearized equations assumed to represent the dynamic longitudinal motions of the aircraft for small deviations from an equilibrium flight path are:

$$\dot{q} = M_q q + M_\alpha \alpha + M_{\dot{\alpha}} \dot{\alpha} + M_{\delta_e} \delta_e \tag{7}$$

$$q - \dot{\alpha} = L_\alpha \alpha + L_{\delta_e} \delta_e \tag{8}$$

where q = pitching velocity
α = incremental angle of attack
δ_e = incremental elevator angle

and M_q, M_α, $M_{\dot{\alpha}}$, M_{δ_e}, L_α and L_{δ_e} are constants.

[$\dot{(\ )}$ denotes $d(\ )/dt$].

The elevator is driven by an actuator whose dynamics are represented by the equation

$$\tau\dot{\delta}_e + \delta_e = \delta_c \tag{9}$$

where δ_c = command input to the actuator

τ = elevator actuator time constant.

These equations written in state vector form are

$$\begin{bmatrix} \dot{\alpha} \\ \dot{q} \\ \dot{\delta}_e \end{bmatrix} = \begin{bmatrix} -L_\alpha & 1 & -L_{\delta_e} \\ M_\alpha - M_{\dot{\alpha}} L_\alpha & M_q + M_{\dot{\alpha}} & M_{\delta_e} - M_{\dot{\alpha}} L_{\delta_e} \\ 0 & 0 & -1/\tau \end{bmatrix} \begin{bmatrix} \alpha \\ q \\ \delta_e \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1/\tau \end{bmatrix} \begin{bmatrix} \delta_c \end{bmatrix} \tag{10}$$

Choosing α and q as system outputs, and choosing

$$Q = \begin{bmatrix} q_{11} & 0 \\ 0 & q_{22} \end{bmatrix}, \quad R = [\,1\,], \text{ and } S = 3 \times 3 \text{ null matrix}$$

(the matrix S does not influence the solution as long as it is non-negative definite and the plant is completely controllable and observable), we formulate the performance index

$$2V = \lim_{T \to \infty} \int_t^T (q_{11}\alpha^2 + q_{22}q^2 + \delta_c^2)\, d\tau \tag{11}$$

Taking numbers representative of the X-15 aircraft at a Mach number of 4.8 and an altitude of 77,000 feet (Table 1), we obtain the optimum responses to initial angle of attack shown in Figure 2 for several choices of Q with $q_{11} = q_{22}$.

Some of the responses are "more optimum" than others on the basis of what we know to be a desirable transient response, although all the closed-loop responses optimize their specified performance index. The choice of relative weighting between output error and input gives a wide range of closed-loop dynamics and there is no a priori basis at present for choosing the weighting factor.

However, this is not to say that optimal system synthesis is useless in flight control application. Quite to the contrary, it is evident from the example that, for the proper choice of weighting factor, a control system having an acceptable response has been synthesized. Figure 3, showing the control input necessary to produce the responses of Figure 2, indicates that quite reasonable magnitudes of elevator deflection are required.

Figure 4 – a flow diagram of the optimal system – shows that the system would not be difficult to mechanize. Examination shows that the feedback gain k_3 (where $K \equiv [k_1 k_2 k_3]$) serves to vary the equivalent time constant

of the elevator actuator. This equivalent time constant is given exactly by the relationship

$$\left(\frac{1}{\tau}\right)_{\text{closed-loop}} = 6.67\,(1 + k_3)$$

Since $K = R^{-1}G'P$ from equation (5), and P must be positive definite, then $k_3 = 1/\tau\ p_{33} > 0$. Therefore for this application the equivalent actuator closed-loop time constant is always smaller in the optimal system.

It is usually desirable to reduce the number of feedback parameters of a closed-loop system. This fact is particularly true of a system designed by optimal techniques because the optimal system nominally requires that each state variable be fed back to each input variable. The above example shows clearly that once a desired optimal response is achieved, it may be possible to replace an existing component by one having the characteristics required for optimal system response.

Another interesting point can be made from the example. Comparing closed-loop characteristic equations we have:

Q Matrix	Characteristic Equation
$q_{11} = q_{22} = 0$ (open loop)	$\lambda^3 + 7.13\lambda^2 + 20.09\lambda + 114.3 = 0$
$q_{11} = q_{22} = .2$	$\lambda^3 + 10.67\lambda^2 + 51.7\lambda + 120.5 = 0$
$q_{11} = q_{22} = .5$	$\lambda^3 + 12.5\lambda^2 + 72.9\lambda + 129.3 = 0$
$q_{11} = q_{22} = 1.0$	$\lambda^3 + 14.33\lambda^2 + 97.3\lambda + 142.1 = 0$

In each case, as Q is made larger, not only does the sum of the roots of the characteristic equation become larger, but the sum of the product of the roots taken two at a time and the product of all three roots increases. It is suspected that the characteristic equations for other optimal systems will also exhibit the properties shown above.

Throughout this study, choice of a performance index has been made using the two steps illustrated in the example. From physical intuition we would like to minimize errors in α and q. So we include α and q in the performance index. Then a weighting factor is selected that produces a desirable transient response. Thus the performance index is used as a performance index – that is, we choose elements of the H and Q matrices to minimize what we would like to minimize from physical considerations – and it is used as a system synthesis tool – that is, the weighting factor is used as a "cut-and try" parameter. The real criterion of performance is judgment applied during the "cut-and-try" procedure.

The beginnings of a straightforward synthesis procedure have been made by recognizing such generalizations as the image root interpretation of the single-input, single-output optimal problem, due to Chang. It is possible that more general results can be obtained by examining the so-called "reverse" optimal solution. That is, given a system and a feedback matrix, what performance index, if any, is minimized by this feedback matrix. The equations of the "reverse" solution are equations (5) and (6) with P, Q (or $H'QH$) and R as unknowns and F, G and K as knowns. These equations can be written as

$$F'P + P\hat{F} + HQ'H = 0 \tag{12}$$

and

$$RK - G'P = 0 \tag{13}$$

Equations (12) and (13) are now linear in the unknown matrices. By virtue of the linearity some facts are readily apparent. Depending on the number of inputs and outputs and the order of the system the "reverse" solution has more, less, or the same number of scalar unknowns as equations. When there are more unknowns than equations, a solution always exists for any assumed K matrix satisfying the inequalities that make P, Q and R positive definite. However, if the number of unknowns is less than or equal to the number of equations, a solution exists only if equalities are satisfied by the elements of the K matrix. Both the inequalities and the equalities become important when the designer is seeking optimal systems that are physically realizable, since the size of the numbers in the K matrix determines directly the feedback configuration by determining how many, if any, feedback gains can be neglected.

A brief example of the equalities that arise is the optimal feedback for the plant $y(s) / u(s) = 1/s^2$. In this case $k_1 = (k_2)^2/2$ is the equality and the "reverse" solution is $q/r = (k_1)^2$.

An example of the case where the number of unknowns exceeds the equations is the following:

Given $F = \begin{bmatrix} -4 & 1 \\ 2 & -3 \end{bmatrix}$, $G = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$, $H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $K = [-.828 \quad -.038]$,

then the "reverse" solution is any Q and R satisfying $Q = \begin{bmatrix} 4r & q_{12} \\ q_{12} & r + 4q_{12} \end{bmatrix}$

and $R = [\, r \,]$, $r > 0$, and $-.24\, r < q_{12} < 16.12\, r$.

SPECIFIC RESULTS

1. Lateral Handling Qualities Problem

The criteria for the design of the primary control system for a manned aircraft are generally expressed in terms of what are known as handling qualities. Handling qualities are a description of the match between the static and dynamic characteristics of the pilot and aircraft to accomplish the mission of the man-machine combination. Good handling qualities are achieved when the level of performance of the combination satisfies the mission objectives with the least amount of pilot concentration and compensation.

The Cornell Aeronautical Laboratory has done considerable research in the past fifteen years toward determining the optimum handling qualities of aircraft.[4] (Reference 4 is a representative recent publication.) The result of this research has led to a definition of combinations of dynamic and static characteristics judged by pilots to provide good handling qualities.

From these characteristics, and from general past experience, a mathematical model of the desired dynamics of an aircraft can be constructed. An optimum control system, in order to provide good handling qualities, must therefore satisfy two criteria:

1. The minimum of the quadratic performance index must be obtained.
2. The dynamic characteristics of the resulting system must be within the area defined by pilots as desirable.

There are at least three ways this problem can be approached.

The first is to obtain by trial and error a number of optimal solutions. From these solutions, it will probably be possible to choose a combination of H, Q, and R that results in a closed-loop optimal response acceptable from a handling qualities point of view.

The second approach is to obtain a direct relationship between the technique of design by performance index and design by time and frequency domain criteria. Preliminary steps have been taken along this path via the root-square locus and the "reverse" solution as described above.

The third approach is to include the mathematical expression for the model directly in the formulation of the system. This is equivalent to redefining the origin of the output space. The inclusion of the model redefines the origin to be a desired trajectory. This desired trajectory can be expressed explicitly by including in the system an actual physical model, or it can be implied by expressing the performance index in a manner that causes that part of the performance index weighing the state variables to go to zero as the optimal response approaches a desired response.

The use of a model simplifies the design procedure considerably. Its use is justified for a number of reasons:

1. A dual performance index is realized; one part expresses dynamic criteria on the output, the other as the optimality condition that minimizes a quadratic function of control action. The best use is made of the control to minimize those errors expressed in the performance index.

2. To approximate the response of the model more and more closely, it is necessary only to increase the $||Q||/||R||$ ratio. The $||Q||/||R||$ ratio will be limited by physical limitations on the feedback gains.

3. The time constants of higher-order plant dynamics can be made small by choosing a model that is lower-order than the plant.

The model is included in the performance index by Kalman* in the following way:

The model is expressed as the matrix L giving the autonomous dynamic system $\dot{\eta} = L\eta$ (η a ficticious state vector). Forming the difference

* Reference 1, Volume II

$\dot{y}$ - Ly and minimizing this difference, the plant output is forced to act like the system $\dot{\eta} = L\eta$. This leads to the performance index

$$2V = \lim_{T \to \infty} \int_t^T [\|\dot{y} - Ly\|_Q^2 + \|u\|_R^2] \, d\tau \tag{14}$$

where the notation $\|a\|_Q^2$ denotes the quadratic form a, Qa where Q is positive definite. Substituting equations (1) and (2) we have

$$2V = \lim_{T \to \infty} \int_t^T [\|(HF - LH)x + HGu\|_Q^2 + \|u\|_R^2] \, d\tau \tag{15}$$

If HG = 0 (which is true for single-input, single-output systems with the number of poles exceeding the number of zeros by at least two), then we define a new H matrix $\hat{H} = HF - LH$ and proceed as before. If $HG \neq 0$, new Lagrangian and Hamiltonian functions must be constructed, but the solution is similar in form to the previous one.

The optimal control system for achieving good lateral handling qualities is synthesized as follows, using the model formulation proposed by Kalman. The plant is assumed to be an X-15 – type aircraft on a re-entry flight path. The lateral-directional motions are described by

$$\left.\begin{aligned} \ddot{\beta} + \dot{r} &= Y_\beta \dot{\beta} + \frac{g}{V} p + Y_{\delta_r} \dot{\delta}_r \\ \dot{p} &= L_\beta \beta + L_p p + L_r r + L_{\delta_a} \delta_a + L_{\delta_r} \delta_r \\ \dot{r} &= N_\beta \beta + N_p p + N_r r + N_{\delta_a} \delta_a + N_{\delta_r} \delta_r \end{aligned}\right\} \tag{16}$$

where β = sideslip angle
p = rolling velocity
r = yawing velocity
δ_r = rudder deflection
δ_a = aileron deflection

and the remaining parameters are constants for a given point on the trajectory.

The aileron and rudder actuators are described by

$$\left.\begin{aligned} \tau_r \dot{\delta}_r + \delta_r &= \delta_{r_c} \\ \tau_a \dot{\delta}_a + \delta_a &= \delta_{a_c} \end{aligned}\right\} \tag{17}$$

The resulting six-dimensional system is written in state vector form as

$$\begin{bmatrix} \dot{\beta} \\ \ddot{\beta} \\ \dot{p} \\ \dot{r} \\ \dot{\delta}_r \\ \dot{\delta}_a \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -N_\beta & Y_\beta & \left(g/V - N_p\right) & -N_r & -\left(N_{\delta_r} + \frac{Y_{\delta_r}}{\tau_r}\right) & -N_{\delta_a} \\ L_\beta & 0 & L_p & L_r & L_{\delta_r} & L_{\delta_a} \\ N_\beta & 0 & N_p & N_r & N_{\delta_r} & N_{\delta_a} \\ 0 & 0 & 0 & 0 & -1/\tau_r & 0 \\ 0 & 0 & 0 & 0 & 0 & -1/\tau_a \end{bmatrix} \begin{bmatrix} \beta \\ \dot{\beta} \\ p \\ r \\ \delta_r \\ \delta_a \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \frac{Y_{\delta_r}}{\tau_r} & 0 \\ 0 & 0 \\ 0 & 0 \\ 1/\tau_r & 0 \\ 0 & 1/\tau_a \end{bmatrix} \begin{bmatrix} \delta_{r_c} \\ \delta_{a_c} \end{bmatrix} \tag{18}$$

Numbers for three flight conditions are listed in Table 2.

The model is exactly the same form as the plant with actuator dynamics omitted and numbers describing a near-optimum aircraft substituted for the re-entry aircraft numbers. In this case the model was chosen as the characteristics of a T-33 aircraft at 25, 000 feet and 250 knots with some artificial damping added to the Dutch roll mode. The resulting L matrix is

$$L \equiv \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -3.72 & -2.156 & .0526 & .317 \\ -8.54 & 0 & -2.56 & 2.50 \\ 3.72 & 2.00 & 0 & -.317 \end{bmatrix}$$

$$H = \begin{bmatrix} I_{4 \times 4} & \begin{matrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{matrix} \end{bmatrix}$$

From previous experience we know that matching the β and p time histories of the model results in adequate overall matching. Therefore we choose Q and R to be

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & q_{22} & 0 & 0 \\ 0 & 0 & q_{33} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

(Although Q is non-negative definite here for convenience, the results show no detrimental effects of this choice.) The choice of Q and R gives the performance index

$$2V = \lim_{T \to \infty} \int_t^T [q_{22}(\dot{\beta} - L_2 y)^2 + q_{33}(\dot{p} - L_3 y)^2 + \delta_{r_c}^2 + \delta_{a_c}^2]\, d\tau \qquad (19)$$

which expresses the desire to make the aircraft side force and yawing equations (combined in the canonical formulation) and rolling equations match those of the model.

Figure 5 shows the response of the open-loop system and the closed-loop optimal system at the first flight condition listed in Table 2. The choice of q_{22} and q_{33} is shown in the figure. The aircraft is nearly out of the atmosphere, where the natural frequency of the Dutch roll mode is very low, with practically no damping. The optimal response does not follow the model closely, but achieves good Dutch roll damping with feedback gain values that are probably physically realizable (with the possible exception of $\dot{\beta}$ feedback to the rudder actuator).

Figure 6 shows the same series of responses except at a later time during the re-entry. Again the optimal system does not follow the model well but the improvement over the open-loop is considerable.

At a still later time, t = 40 sec, the aircraft follows the model more closely and, from a handling qualities point of view, would probably be regarded as a good aircraft. The transient responses are shown in Figure 7.

By increasing the ratio of $\|Q\|$ to $\|R\|$, it is presumed possible to make the aircraft follow the model as closely as desired. Figure 8 shows the response of an optimal system where the $\|Q\|$ to $\|R\|$ ratio has been increased by a factor of 5. The open-loop aircraft is that of Figure 7. The closed-loop aircraft follows the model closely and it may be presumed that this fidelity can be improved even more by increasing $\|Q\|/\|R\|$. Figure 8 shows, however, that as this ratio is increased, a corresponding increase in the feedback gains is required.

If one desired to build these systems without simplification, two obvious disadvantages of the optimal systems from the hardware viewpoint can be seen from these results. Assuming that the feedbacks of actuator position can be incorporated in the original actuator design, one must still provide sideslip, roll rate and yaw rate sensors. Adding a differentiating network for the sideslip signal brings the total number of channels up to eight. This

is a lot of complexity. Also, the feedback gains must be programmed as a function of time or changed in some other manner during the re-entry. This also adds complexity. The advantages of designing a control system of this complexity are not apparent in the results.

It should be emphasized, however, that these results do not preclude the possibility of deriving optimal systems that are less complex or closely approximating optimal systems with less complex systems.

2. Booster Pitch Attitude Control Problem

The first-stage attitude control system of a large booster must be designed to satisfy several criteria. It must stabilize a usually unstable airframe; it must maintain deviation from commanded attitude at a minimum; it must respond to wind disturbances so that angle of attack does not exceed the design value, at the same time limiting drift off the nominal trajectory; and it must accomplish the control task without exceeding engine deflection limits and rate limits. The quadratic performance index by itself cannot express all these desires. However, for present purposes the problem is restricted to that of maintaining minimum attitude errors – a simplification that will yield results and yet that is not far removed from the real problem.

The motions of the flexible booster in the pitch plane and the engine-deflection actuator are described by:

$$\text{Rigid Body} \quad \begin{cases} \ddot{\theta}_R = -M_1 \alpha_R - M_2 \delta - M_3 \ddot{\delta} \\ \dot{\theta}_R - \dot{\alpha}_R = N_1 \theta_R + N_2 \delta + N_3 \alpha_R + N_4 \ddot{\delta}_R \end{cases}$$

$$\text{Structural Modes} \quad \begin{cases} \ddot{\theta}_{s_1} + 2\zeta\omega_{n_1} \dot{\theta}_{s_1} + \omega_{n_1}^2 \theta_{s_1} = k_1 \delta + k_1' \ddot{\delta} \\ \ddot{\theta}_{s_2} + 2\zeta\omega_{n_2} \dot{\theta}_{s_2} + \omega_{n_2}^2 \theta_{s_2} = k_2 \delta + k_2'' \ddot{\delta} \end{cases}$$

$$\text{Actuator} \quad a_3 \dddot{\delta} + a_2 \ddot{\delta} + a_1 \dot{\delta} + \delta = \delta_c \qquad (20)$$

where θ_R = pitch attitude error from nominal trajectory

α_R = rigid body angle of attack

δ = engine deflection

θ_{s_i} = local attitude of structure relative to rigid body mode (i^{th} mode)

δ_c = commanded engine deflection,

and M, N, ζ, ω_n, a_1, etc. are constants.

Written in state-vector form, these equations are as follows

$$\frac{d}{dt}\begin{bmatrix} \theta_R \\ \dot{\theta}_R \\ \alpha_R \\ \delta \\ \dot{\delta} \\ \ddot{\delta} \\ \theta_{s_1} \\ \dot{\theta}_{s_1} \\ \theta_{s_2} \\ \dot{\theta}_{s_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -M_1 & -M_2 & 0 & -M_3 & 0 & 0 & 0 & 0 \\ -N_1 & 1 & -N_3 & -N_2 & 0 & -N_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{a_3} & -\frac{a_1}{a_3} & -\frac{a_2}{a_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & k_1 & 0 & k_1'' & -\omega_{n_1}^2 & -2\zeta\omega_{n_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & k_2 & 0 & k_2'' & 0 & 0 & -\omega_{n_2}^2 & -2\zeta\omega_{n_2} \end{bmatrix} \begin{bmatrix} \theta_R \\ \dot{\theta}_R \\ \alpha_R \\ \delta \\ \dot{\delta} \\ \ddot{\delta} \\ \theta_{s_1} \\ \dot{\theta}_{s_1} \\ \theta_{s_2} \\ \dot{\theta}_{s_2} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{a_3} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} [\delta_c] \tag{21}$$

One flight condition is examined here – Mach Number = 5 at 160,000 feet. This condition is typical of first-stage burnout (numbers given in Table 3).

Since we desire to minimize rigid body attitude errors but have no interest in adding to the structural damping or changing the actuator in any way, the H matrix is chosen as:

$$H = [\,1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\,].$$

This makes the Q and R matrices both scalars and leads to the performance index

$$2V = \lim_{T\to\infty} \int_t^T (q/r\ \theta_R^2 + \delta_c^2)\,d\tau \tag{22}$$

(The choice of H in this case gives a plant that is not completely observable. However the optimal regulator solution may, nevertheless, be unique, due to the stability of the unobservable part of the plant – the structural modes.)

After running several solutions, the ratio $q/r = 10$ was chosen as giving a good compromise between speed of response and peak engine deflections. This choice, of course, had to be based on previous experience with the booster problem. The transient response to the initial condition $\theta_R = 1.0$ deg, $\alpha_R = 1.0$ deg, is shown in Figure 9. The closed-loop structural damping is no greater than the open-loop damping ($\zeta = .01$ from Table 3), but this is to be expected since the structural mode state variables are not included in the performance index.

The feedback gains for Figure 9 are

$$\delta_c = -Kx$$

$$\delta_c = 3.16\,\theta_R + 1.62\,\dot{\theta}_R + .018\,\alpha_R - .246\,\delta - .0007\,\dot{\delta} - .194 \times 10^{-4}\,\ddot{\delta} - (0)\,\theta_{s_1} - (0)\,\dot{\theta}_{s_1} - (0)\,\theta_{s_2} - (0)\,\dot{\theta}_{s_2}. \qquad (23)$$

This result is both encouraging and discouraging, for although α_R, $\dot{\delta}$, $\ddot{\delta}$, θ_{s_1}, $\dot{\theta}_{s_1}$, θ_{s_2}, and $\dot{\theta}_{s_2}$ can be omitted from the feedback equation without deviating appreciably from the optimum system, the remaining rigid-body variables cannot be measured without sensing some structural motion. For some sensor positions, this structural feedback is destabilizing.

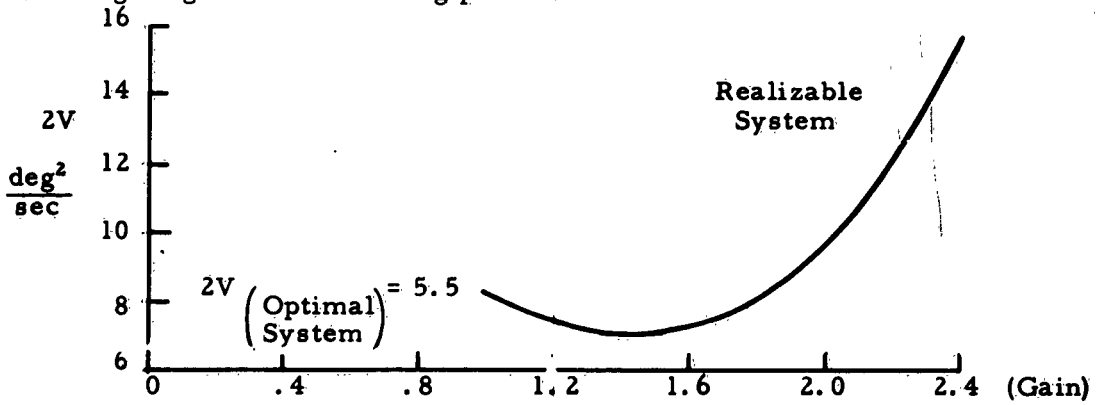
The booster representation was next simplified to shed some light on the question of ignorable higher-order dynamics. The designer would like to know whether it is legitimate to ignore higher-order dynamics in constructing an optimal solution when physical intuition indicates that they can be ignored. For the flight condition examined here, one would suspect that the structural modes might be ignorable (that is, ignorable when determining the feedback equation, not when mechanizing it). This was shown to be the case, because the feedback equation for the booster without flexibility is the same as equation (23) with θ_{s_1}, $\dot{\theta}_{s_1}$, etc. omitted. Dropping the actuator and the heaving degree of freedom from the booster formulation, we find the feedback equation is

$$\delta_c = 3.17\,\theta_R + 1.45\,\dot{\theta}_R \qquad (24)$$

Comparing equations (23) and (24), we suspect that the omitted dynamic effects are almost but not quite ignorable. The feedback of actuator position is the only significant term omitted in equation (24).

These results indicate that dynamic effects far outside the bandwidth of interest can be ignored in determining the optimum feedback configuration much the same as they are ignored in conventional design practice.

The final result in the booster application is the comparison of the optimal system with a realizable system. Figure 10 is the block diagram of a realizable system. This loop, taken from a previous study, was not designed to approximate the feedback of equation (23) except in the sense that low-pass filters are included to greatly attenuate the feedback of θ_{s_1} and θ_{s_2}. The compensation is lead-lag. However, plotting 2V from equation (22) versus gain gives the following picture.



The realizable system, in this flight condition, approaches the optimal system quite closely although the transient response is slightly oscillatory. As gain is increased from the minimum point, system oscillations increase (so $\int_0^\infty \delta_c^2 d\tau$ increases), but the bandwidth also increases so that $\int_0^\infty \Theta_R^2 d\tau$ continues to decrease. Figure 11 shows the transient response at a gain of 2.2. As is shown on the plot, the error portion of the performance index is almost equal to that of the optimum, but the control input integral is considerably larger.

CONCLUSIONS

Only some of the results of the study mentioned in the introduction have been included in this paper, but they illustrate the general experience gained. Both success and failure have been encountered in choosing performance indices giving satisfactory optimum systems. However, the following conclusions can be made:

1. State variables characterizing higher-order system dynamics need not, in some cases, be fed back to obtain an optimal system if they do not appear explicitly in the performance index.
2. Choosing the performance index from physical intuition, except for the relative weights given to the error and the control, works fairly well in obtaining satisfactory transient response, although this procedure will not usually lead to a feedback configuration that is easily realized physically.
3. Further work is needed to establish general relations between the performance index and the resulting feedback configuration if the optimal control theory is to be useful as a practical system synthesis tool.

REFERENCES

1. Kalman, R. E., Englar, T. S., and Bucy, R. S., Fundamental Study of Adaptive Control Systems, Vol. I and II. ASD TR 61-27, March 1961 and March 1962.

2. Kalman, R. E., The Theory of Optimal Control and the Calculus of Variations. AFOSR 1121; RIAS TR 61-3, October 1960.

3. Chang, S. S. L., Synthesis of Optimum Control Systems. McGraw-Hill Book Company, New York, pp. 11 - 35.

4. Harper, R. P., Jr., In-Flight Simulation of the Lateral-Directional Handling Qualities of Entry Vehicles. WADD TR 61-147, November 1961.

TABLE 1

X-15 - AIRCRAFT LONGITUDINAL DERIVATIVES

M = 4.8, h = 77,000 feet

M_q	$= -.132\ sec^{-1}$	L_α	$= .277\ sec^{-1}$
$M_{\dot\alpha}$	$= -.046\ sec^{-1}$	L_{δ_e}	$= .037\ sec^{-1}$
M_α	$= -17.1\ sec^{-2}$	τ	$= .15$ sec
M_{δ_e}	$= -12.2\ sec^{-2}$		

TABLE 2

RE-ENTRY AIRCRAFT DERIVATIVES

time after re-entry initiation (sec)	Mach No.	h (1000 ft)	Y_β sec^{-1}	L_β sec^{-2}	L_p sec^{-1}	L_r sec^{-1}	N_β sec^{-2}
0	5.3	226	-.0018	.0208	-.0062	-.0002	.492
20	5.6	211	-.0036	.0637	-.0123	-.0004	.932
40	5.5	147	-.0019	.5140	-.0625	+.0376	2.95

N_r sec^{-1}	Y_{δ_r} sec^{-2}	L_{δ_a} sec^{-2}	L_{δ_r} sec^{-2}	N_{δ_a} sec^{-2}	N_{δ_r} sec^{-2}	g/V sec^{-1}
-.0015	.0003	-.335	.150	-.0161	-.178	.00552
-.0028	.0007	-.647	.353	-.0311	-.349	.00550
-.0103	.0025	-5.11	-2.32	-.249	-1.43	.00535

TABLE 3

BOOSTER DERIVATIVES

Mach No. = 5.0 h = 160,000 feet

$M_1 = -.036\ (sec^{-2})$

$M_2 = 3.02\ (sec^{-2})$

$M_3 = .00175\ (---)$

$N_1 = .0292\ (sec^{-1})$

$N_2 = .0148\ (sec^{-1})$

$N_3 = .0012\ (sec^{-1})$

$N_4 = .936 \times 10^{-5}\ (---)$

$a_1 = 56.3 \times 10^{-3}\ (sec)$

$a_2 = .809 \times 10^{-3}\ (sec)^2$

$a_3 = .00447 \times 10^{-3}\ (sec)^3$

$\omega_{n_1} = 15.1\ (rad/sec)$

$\omega_{n_2} = 86.1\ (rad/sec)$

$\zeta = .01\ (---)$

$k_1 = 6.21\ (sec^{-2})$

$k_1'' = .00515\ (---)$

$k_2 = 59.6\ (sec^{-2})$

$k_2'' = .0843\ (---)$

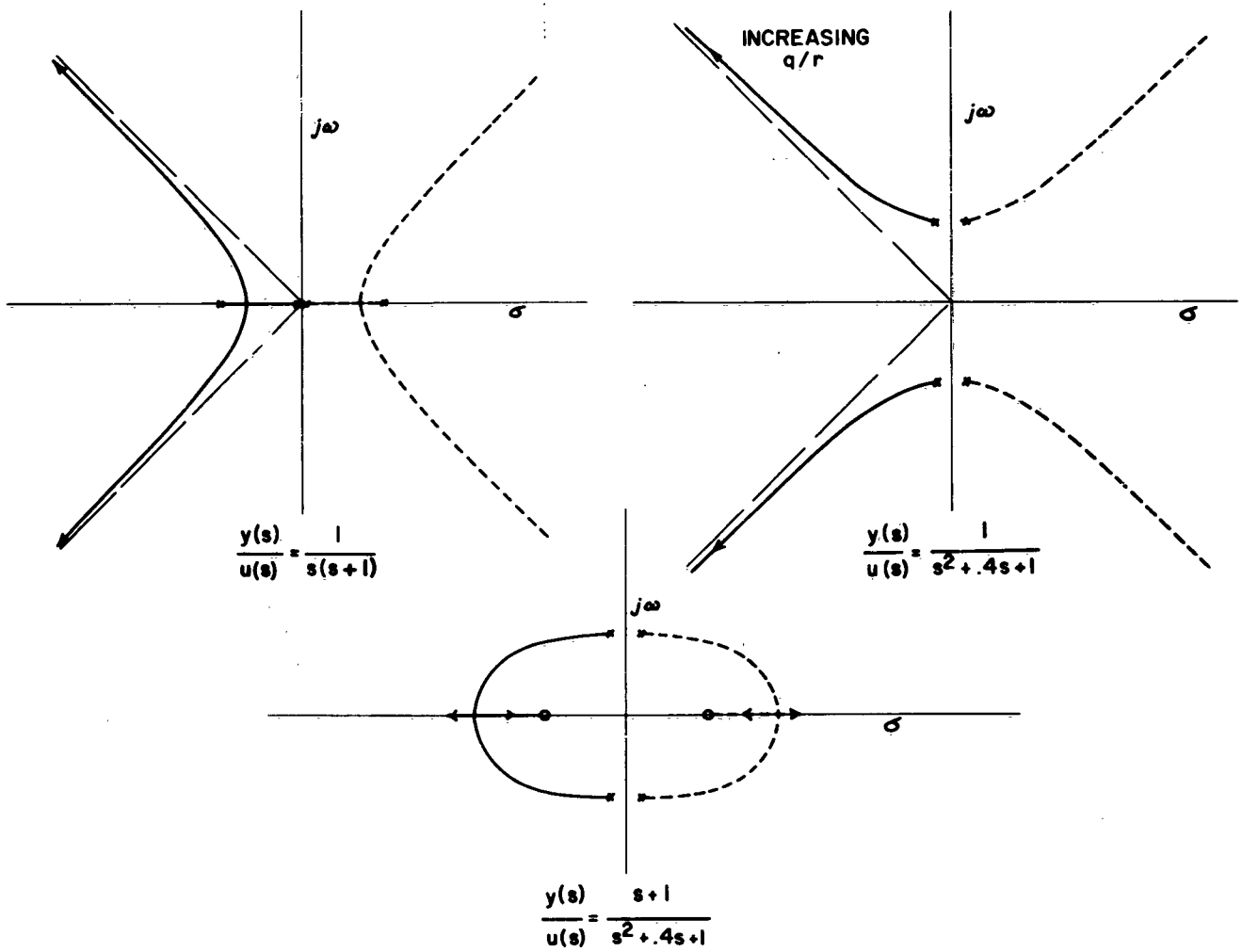


Figure 1. Optimal Second-Order Systems
(Single Input, Single Output)

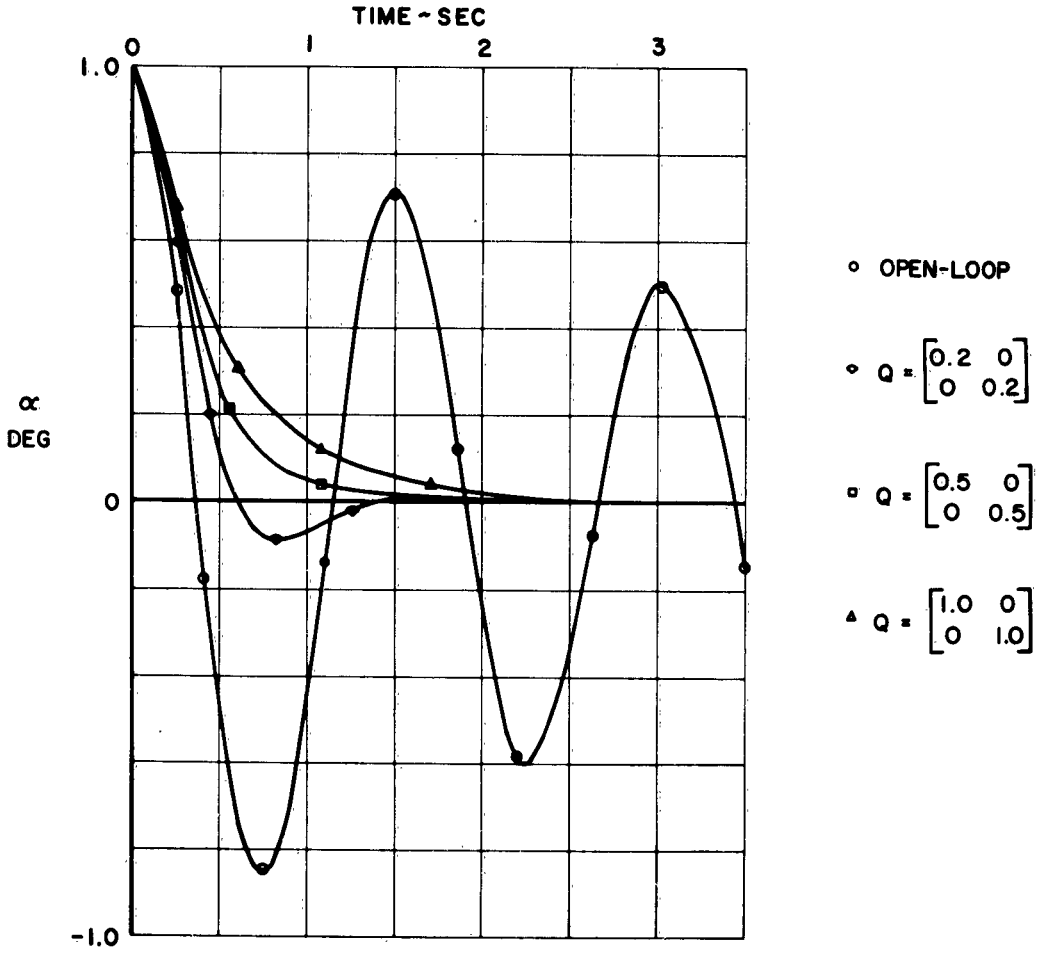


Figure 2. Response of Optimal System to Initial Angle of Attack

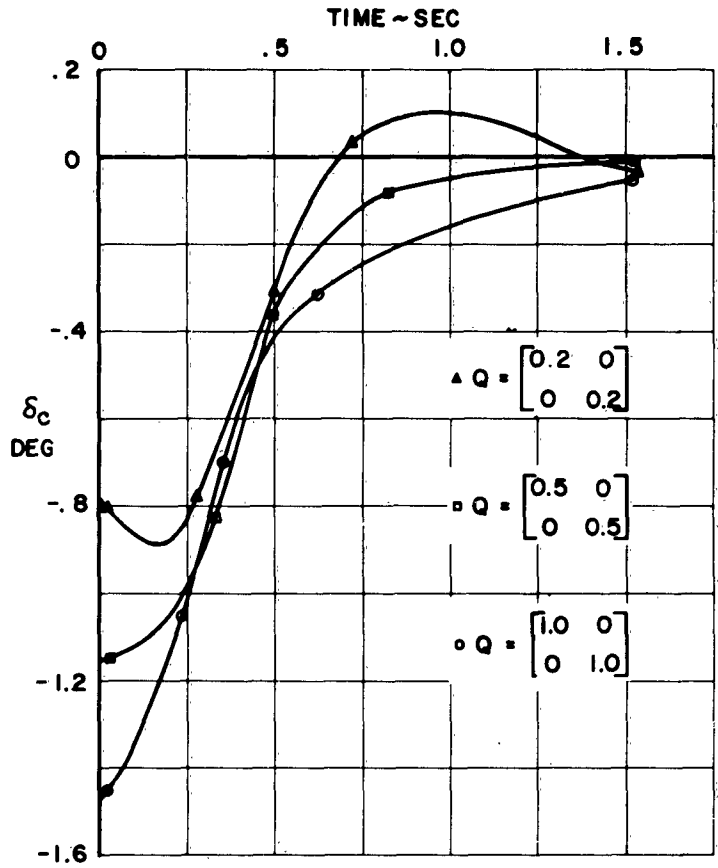


Figure 3. Control Input (δ_c) Response to Initial Angle of Attack

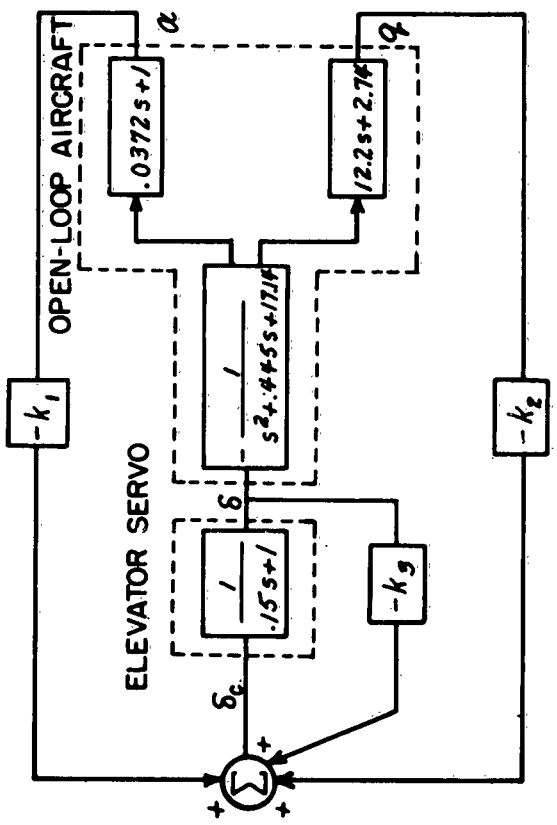
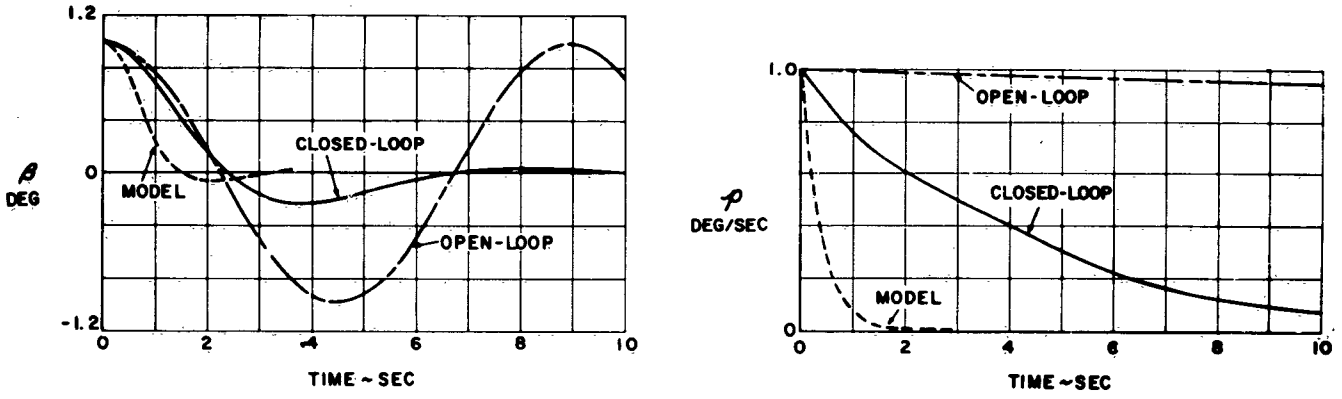


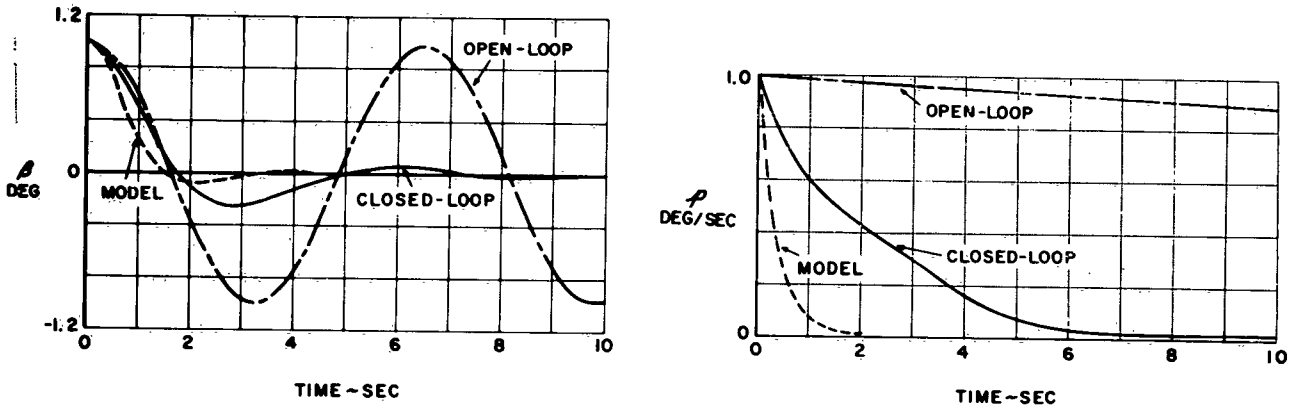
Figure 4. Optimal Regulator for Aircraft Pitch Axis Control



$$K = \begin{matrix} \\ \delta_{r_c} \\ \delta_{a_c} \end{matrix} \begin{matrix} \beta & \dot{\beta} & p & r & \delta_r & \delta_a \end{matrix} \begin{bmatrix} 2.3609 & 3.7336 & 0.8448 & -.8679 & .0410 & -.0405 \\ -1.2390 & -.0538 & -0.7300 & .6626 & -.0108 & .0444 \end{bmatrix}$$

$$V = \int_{t=0}^{T\to\infty} \left[0.2(\ddot{\beta} - L_2 Y)^2 + 0.2(\dot{p} - L_3 Y)^2 + \delta_{r_c}^2 + \delta_{a_c}^2 \right] d\tau$$

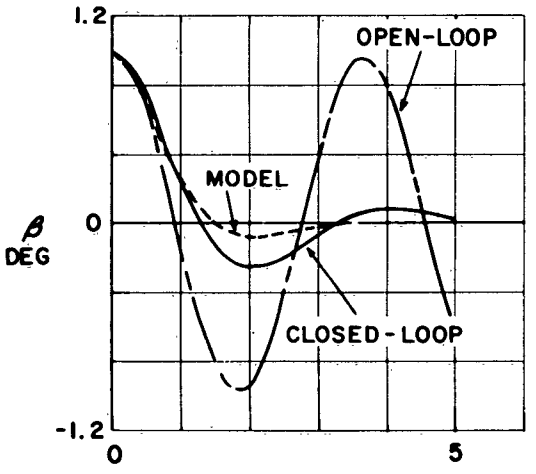
Figure 5. Optimal Regulator Response Simplified X-15 Lateral Dynamics Mach = 5.3 h = 226,000 Ft

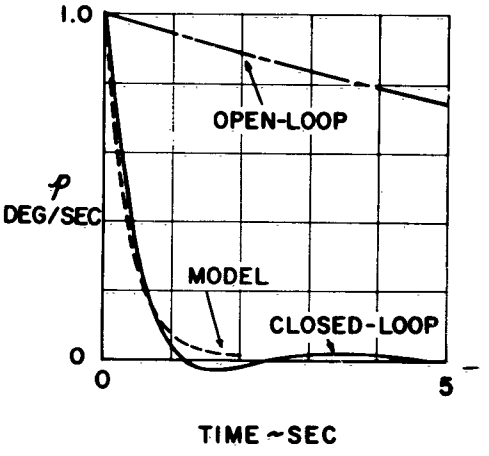


$$K = \begin{matrix} \\ \delta_{r_c} \\ \delta_{a_c} \end{matrix} \begin{matrix} \begin{matrix} \beta & \dot{\beta} & \varphi & r & \delta_r & \delta_a \end{matrix} \\ \begin{bmatrix} 2.4966 & 2.2788 & 0.8273 & -0.8598 & 0.0743 & -0.0998 \\ -1.6652 & 0.0353 & -0.7657 & 0.7020 & -0.0266 & 0.1030 \end{bmatrix} \end{matrix}$$

$$V = \int_{t=0}^{T \to \infty} \left[0.2(\ddot{\beta} - L_2 Y)^2 + 0.2(\dot{p} - L_3 Y)^2 + \delta_{r_c}^2 + \delta_{a_c}^2 \right] d\tau$$

Figure 6. Optimal Regulator Response Simplified X-15 Lateral Dynamics Mach = 5.6 h = 211,000 Ft

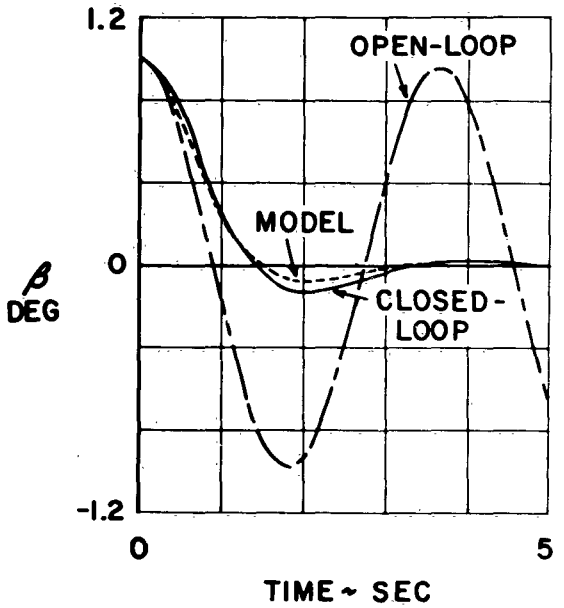


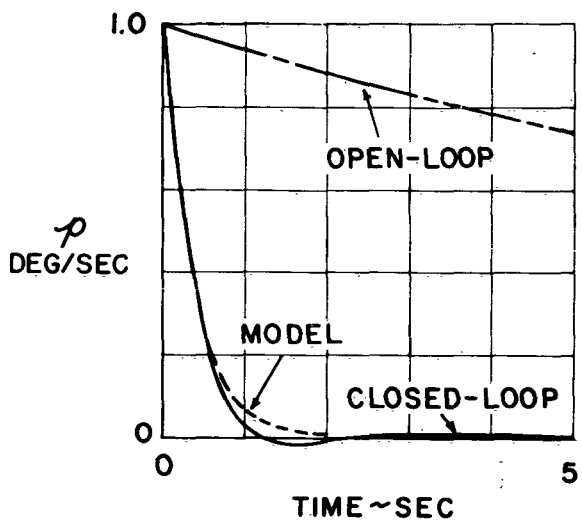


$$K = \begin{matrix} \\ \delta_{r_c} \\ \delta_{a_c} \end{matrix} \begin{matrix} \begin{matrix} \beta & \dot{\beta} & p & r & \delta_r & \delta_a \end{matrix} \\ \begin{bmatrix} -1.6150 & 1.1565 & -.4953 & .4136 & .5855 & 1.0995 \\ -3.0560 & -.8129 & -.9953 & 1.0146 & .2955 & 1.5340 \end{bmatrix} \end{matrix}$$

$$V = \int_{t=0}^{T \to \infty} \left[0.2(\dot{\beta} - L_2 Y)^2 + 0.2(\dot{p} - L_3 Y)^2 + \delta_{r_c}^2 + \delta_{a_c}^2 \right] d\tau$$

Figure 7. Optimal Regulator Response Simplified X-15
Lateral Dynamics Mach = 5.5 h = 147,000 Ft





$$K = \begin{matrix} \\ \delta_{r_c} \\ \delta_{a_c} \end{matrix} \begin{matrix} \begin{matrix} \beta & \dot{\beta} & p & r & \delta_r & \delta_a \end{matrix} \\ \begin{bmatrix} -4.6945 & 1.8520 & -1.4191 & 1.1768 & 1.7547 & 3.0459 \\ -7.2242 & -1.6167 & -2.0470 & 2.1792 & 0.8195 & 3.5480 \end{bmatrix} \end{matrix}$$

$$V = \int_{t=0}^{T\to\infty} \left[(\ddot{\beta} - L_2 Y)^2 + (\dot{p} - L_3 Y)^2 + \delta_{r_c}^{\,2} + \delta_{a_c}^{\,2} \right] d\tau$$

Figure 8. Optimal Regulator Response Simplified X-15

Lateral Dynamics Mach = 5.5 h = 147,000 Ft

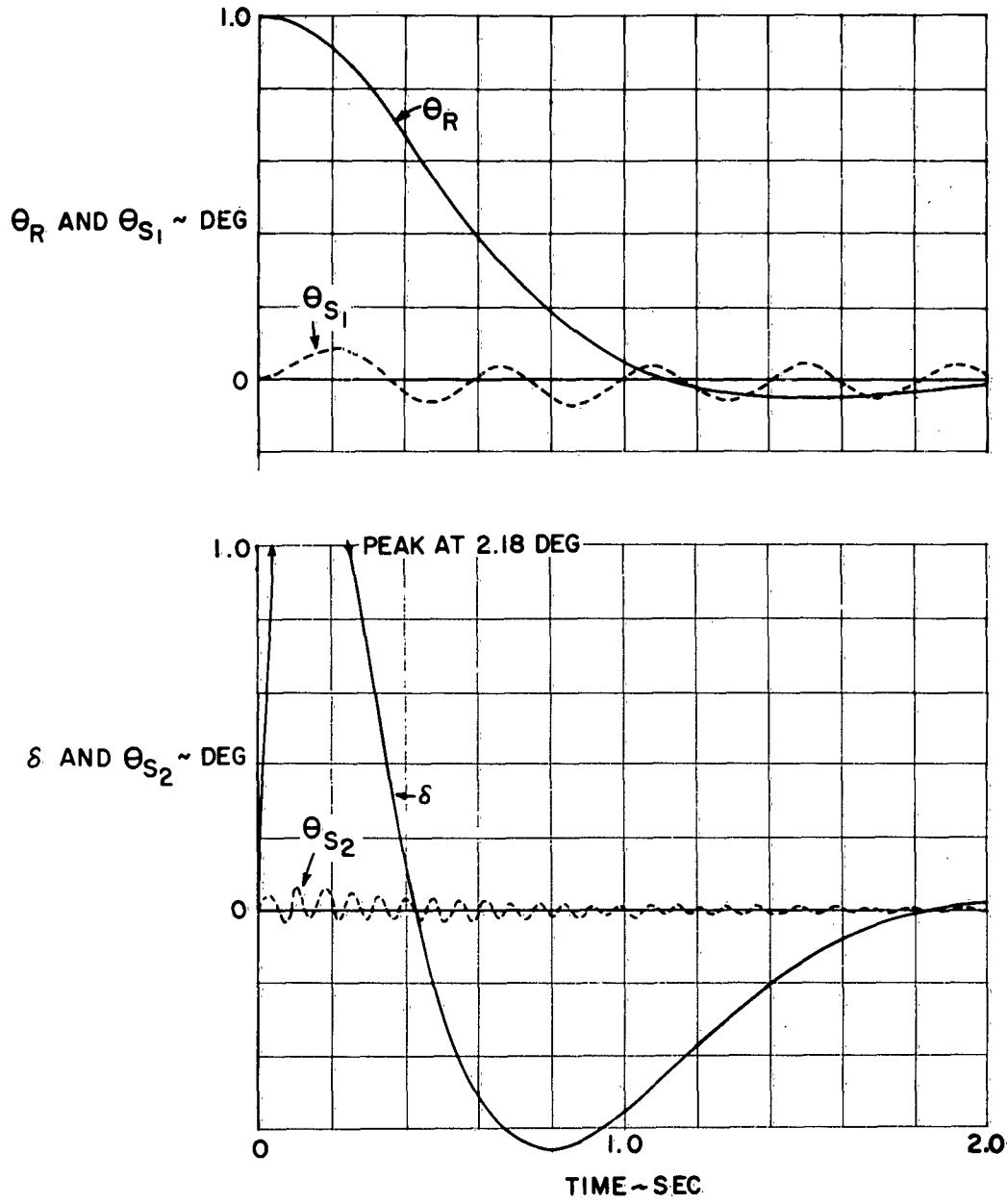


Figure 9. Optimum Regulator Response to Initial θ_R and α_R Booster at Mach 5, 160,000 Ft

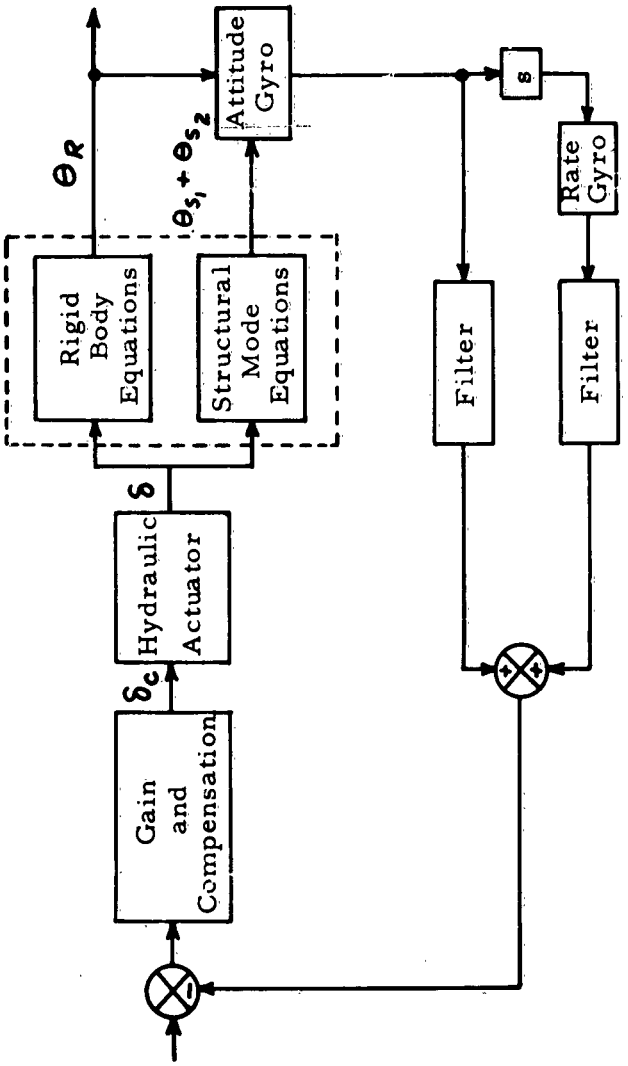
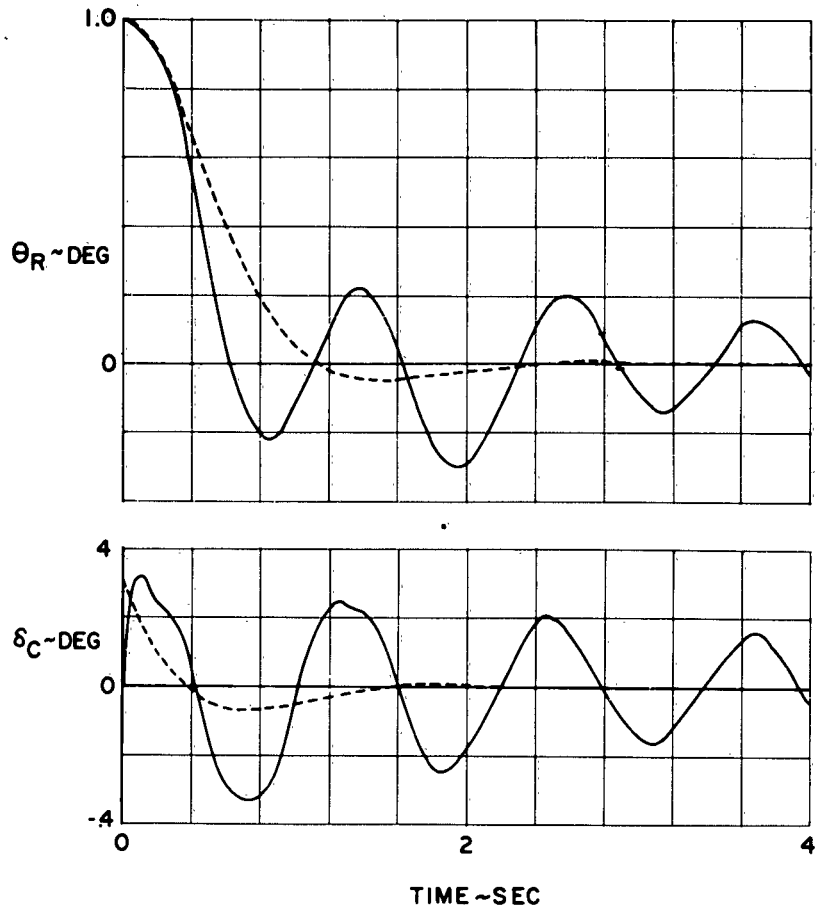


Figure 10. Realizable Booster Attitude Control System



—— HIGH-GAIN SYSTEM

$$2V = \int_0^\infty 10\,\theta_R^2\,dt + \int_0^\infty \delta_C^2\,dt = 4.12 + 12.28 = 16.40\ \text{DEG}^2\text{-SEC}$$

----- OPTIMAL SYSTEM

$$2V_{min} = \int_0^\infty 10\,\theta_R^2\,dt + \int_0^\infty \delta_C^2\,dt = 3.97 + 1.54 = 5.51\ \text{DEG}^2\text{-SEC}$$

Figure 11. Comparison of Optimal Regulator With a Realizable Booster Control System (Gain = 2.2)

Sortie Boost and Glide Trajectories Determined for Specific Mission Requirements by the Steepest Ascent Technique

By

K. Mikami
C. T. Battle
R. S. Goodell

A. E. Bryson, Consultant

Raytheon Company
Missile and Space Division

ABSTRACT

A three-dimensional trajectory optimization computer program is described for determining vehicle performance boundaries. Simultaneous programs for angle-of-attack and bank-angle, with specified initial and terminal conditions, are determined by this program using steepest-ascent computational procedures. Terminal constraints in the form of upper bounds on total stagnation point heat absorption and total acceleration dose and in-flight constraints in the form of upper bounds on altitude and stagnation point heating rate are also included. The vehicle is treated as a mass particle and the earth is approximated as on oblate-spheroid with an ARDC atmosphere rotating with the earth. Only coordinated turn maneuvers are considered. The terminal functions which can be optimized include all the state variables individually and a number of functions of the state variables. These computer programs are specifically oriented for computations on the IBM 7090.

A numerical example is given for a SORTIE configuration consisting of a three-stage booster system and a payload, the re-entry vehicle, SORTIE. Three separate but related performance boundaries were computed for this configuration: 1) the boost boundary for the three-stage booster system, 2) SORTIE re-entry and glide, 3) SORTIE lateral maneuvers in the recovery area.

SORTIE BOOST AND GLIDE TRAJECTORIES DETERMINED FOR SPECIFIED MISSION REQUIREMENTS BY THE STEEPEST-ASCENT TECHNIQUE*

Kinya Mikami
C. Tucker Battle
Ralph S. Goodell
and
Arthur E. Bryson, Consultant
Raytheon Company
Missile and Space Division

1. INTRODUCTION

The determination of the performance capabilities of a hypersonic vehicle requires a three-dimensional analysis because of the rotation of the atmosphere with the earth, the oblateness of the earth, and the fact that the vehicle is capable of making lateral turns. Constraints on acceleration dosage, heating, heating rate, and altitude may also be placed upon the vehicle.

To map out the performance boundaries of a given vehicle by a cut-and-try approach requires many individual tries. The technique presented in this report allows one to proceed in a systematic and efficient manner to map out the performance boundaries. It requires the use of a high-speed computer to determine each trajectory and employs a steepest-ascent or successive-improvement method.

Two major computer programs are presented: 1) for a gliding vehicle and 2) for a vehicle which has a fixed direction of thrust relative to the vehicle. Only coordinated turns are considered for both gliding and thrusting vehicles; thus angle-of-attack and bank angle are the control variables. The vehicle is approximated by a mass particle acted upon by aerodynamic, propulsive, and gravitational forces. This represents a considerable simplification over treating the vehicle as a rigid body, which would involve, in addition, the pitch,

*This study program was supported in part by funds under USAF Contract No. AF 33(616)-8300, Task No. 143107 of Project 1431. This paper consists of a condensation of References (10-12).

yaw and roll dynamics. It is believed that this approximation is reasonable when determining performance boundaries.

In-flight constraints are satisfied by the use of penalty functions which effectively converts them into terminal constraints. Two in-flight constraints are considered, either one of which may be applied in a given problem: 1) An upper bound on altitude during the flight and 2) an upper bound on the heating rate at the stagnation point during the flight.

2. COORDINATE SYSTEMS

2.1 Spherical Coordinate System

The choice of the coordinate systems is an important part of the optimization analysis. Since the atmosphere and the desired landing point both rotate with the earth, it seems that a simple and desirable coordinate system to use is the conventional geographic coordinate system. This coordinate system is a spherical polar coordinate system described by a radius from the center of the earth, r, a co-latitude angle measured from the North Pole, θ, and a longitude angle measured from the Greenwich or prime meridian, ϕ. (See Figure 1). The earth is assumed to be an oblate spheroid with an equatorial radius, R_e, of 20.925631×10^6 feet and a polar radius, R_p, of 20.855965×10^6 feet. (See Reference 1.) Thus the radius of the earth becomes a function of the co-latitude and can be given as

$$R(\theta) = \frac{R_p}{\sqrt{1 - \frac{R_e^2 - R_p^2}{R_e^2} \sin^2 \theta}} \qquad (2\text{-}1)$$

Since the second term in the radical is much less than unity, the radical can be expanded into a series. Retaining only the first two terms, the radius of the earth is given approximately by:

$$R(\theta) = R_e \left(1 - \frac{R_e - R_p}{R_e} \cos^2 \theta\right) \qquad (2\text{-}2)$$

Because the earth is an oblate spheroid of eccentricity near unity, the radius vector is not quite perpendicular to the local earth mean-surface. However, the maximum angle between the radius vector and the normal to the spheroid is 11.5 minutes of arc and occurs at a co-latitude of 45°. Thus the error in assuming that the altitude is given by:

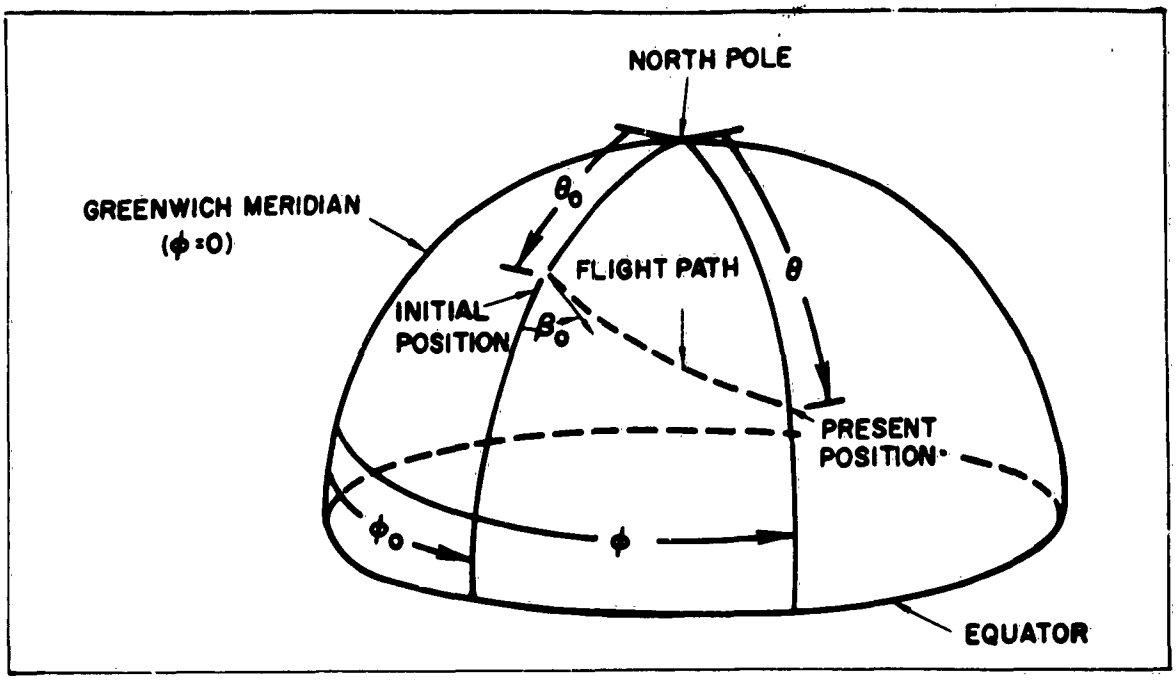


Figure 1 - Geographic Position Coordinates r, θ, ϕ

$$h = r - R(\theta) \tag{2-3}$$

is negligibly small. The atmospheric density is assumed to be a function of altitude h above the earth's surface and is taken to be the standard 1959 ARDC atmosphere model.

2.2 Auxiliary Coordinate Systems θ^*, ϕ^* and μ, ν

Two auxiliary coordinate systems are used for convenience in determining vehicle performance. One coordinate system is described by the angles θ^*, ϕ^*. This system places a "north pole" at the initial point of the problem located by the geographic angles θ_o and ϕ_o. The "prime meridian" is oriented to lie in the great circle plane of the initial flight path direction. Thus, θ^* is the great circle distance from the initial point and ϕ^* is the lateral longitude from the initial great circle. (See Figure 2).

The "equator" of the second coordinate system is located on the initial great circle and the "prime" meridian lies through the initial point at right angles to the "equator". μ is the longitude measured along the "equator" and

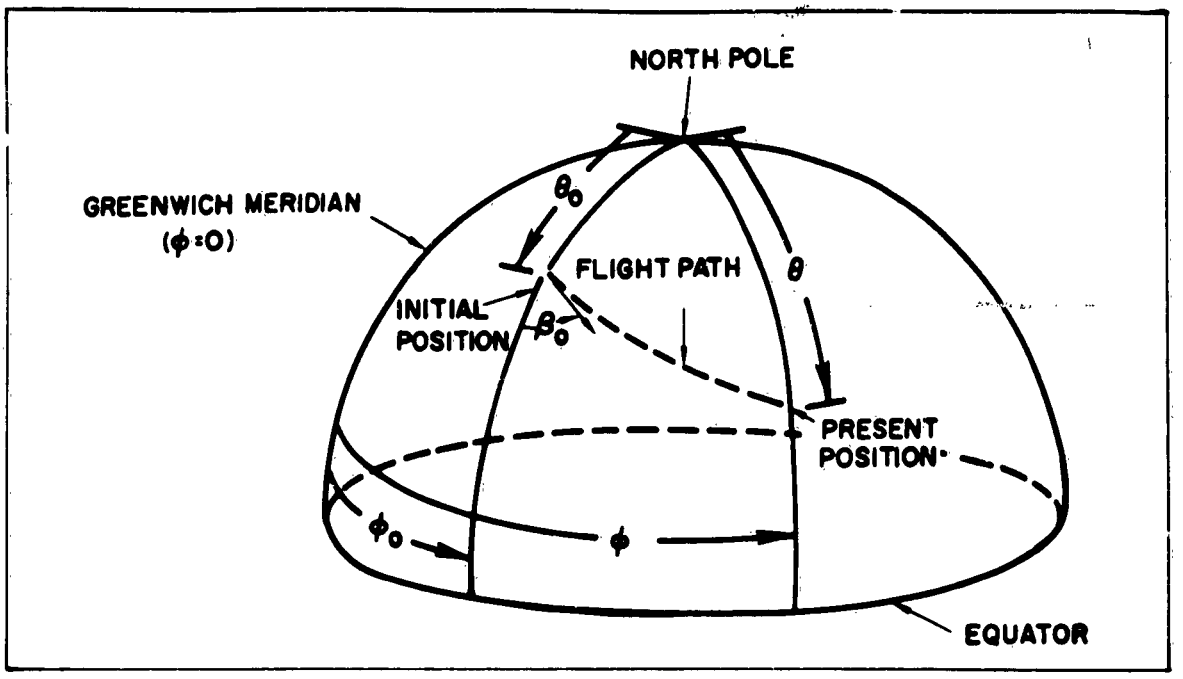


Figure 1 - Geographic Position Coordinates r, θ, ϕ

$$h = r - R(\theta) \tag{2-3}$$

is negligibly small. The atmospheric density is assumed to be a function of altitude h above the earth's surface and is taken to be the standard 1959 ARDC atmosphere model.

2.2 Auxiliary Coordinate Systems θ^*, ϕ^* and μ, ν

Two auxiliary coordinate systems are used for convenience in determining vehicle performance. One coordinate system is described by the angles θ^*, ϕ^*. This system places a "north pole" at the initial point of the problem located by the geographic angles θ_o and ϕ_o. The "prime meridian" is oriented to lie in the great circle plane of the initial flight path direction. Thus, θ^* is the great circle distance from the initial point and ϕ^* is the lateral longitude from the initial great circle. (See Figure 2).

The "equator" of the second coordinate system is located on the initial great circle and the "prime" meridian lies through the initial point at right angles to the "equator". μ is the longitude measured along the "equator" and

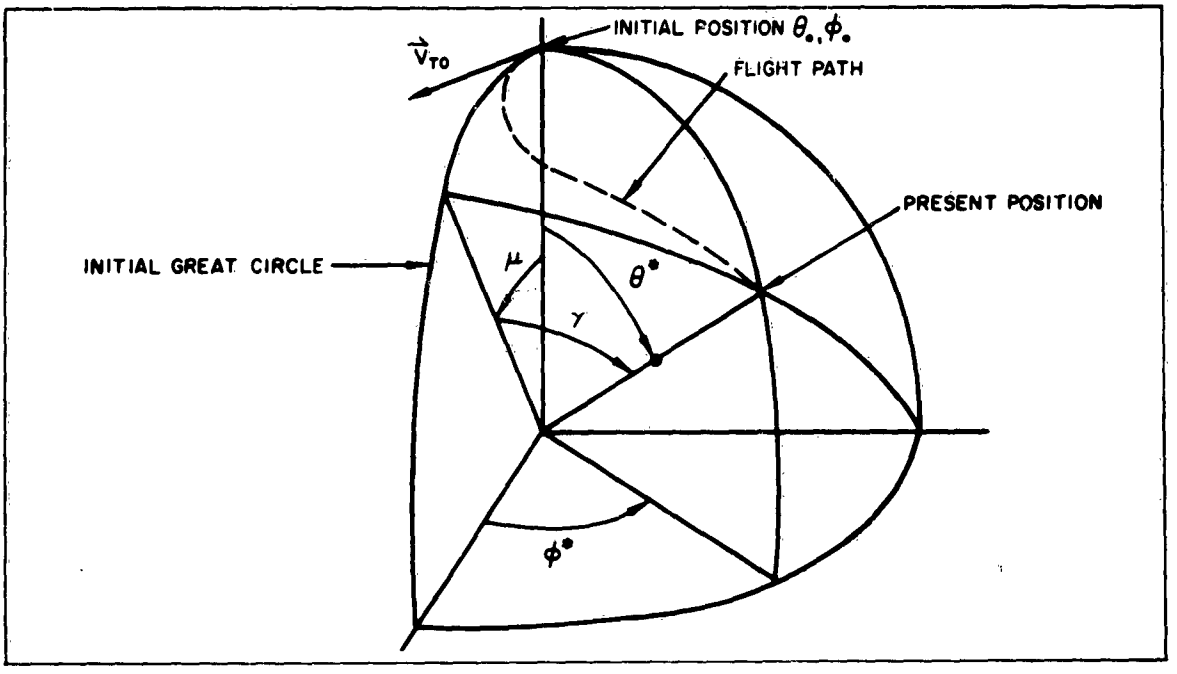


Figure 2 - Auxiliary Position Coordinates θ^*, ϕ^* or μ, ν

ν is the great circle distance from the initial great circle plane, i.e. latitude from the "equator". (See Figure 2).

The radius vector, $\vec{r}$, is the same in all three coordinate systems. Knowing the initial co-latitude, θ_o, the initial longitude, ϕ_o, and the initial flight path heading angle, β_o, along with the present co-latitude, θ, and longitude, ϕ, the coordinates θ^*, ϕ^*, μ and ν can be found from:

$$\cos\nu\cos\mu = \cos\theta^* = \sin\theta_o \sin\theta \cos(\phi_o - \phi) + \cos\theta_o \cos\theta$$

$$\sin\nu = \sin\theta^* \sin\phi^* = \sin\theta_o \cos\theta \sin\beta_o$$

$$- \sin\theta \left[\sin\beta_o \cos\theta_o \cos(\phi_o - \phi) + \cos\beta_o \sin(\phi_o - \phi) \right] \quad (2\text{-}4)$$

$$\cos\nu\sin\mu = \sin\theta^* \cos\phi^* = \sin\theta \left[\cos\beta_o \cos\theta_o \cos(\phi_o - \phi) \right.$$

$$\left. - \sin\beta_o \sin(\phi_o - \phi) \right] - \cos\beta_o \sin\theta_o \cos\theta$$

or

$$\sin\theta^* = \sqrt{\sin^2\nu + \sin^2\mu\cos^2\nu}$$
$$\tan\phi^* = \tan\nu\csc\mu \qquad (2\text{-}5)$$

and

$$\tan\mu = \tan\theta^*\cos\phi^*$$
$$\sin\nu = \sin\theta^*\sin\phi^* \qquad (2\text{-}6)$$

3. STATE EQUATIONS

3 1 Equations of Motion

Vehicle position is described by the spherical coordinates r, θ, and ϕ. (See Figure 3). Let $\hat{r}$, $\hat{\theta}$, and $\hat{\phi}$ be unit vectors in the direction of increasing r, θ and ϕ, respectively; then the velocity of the vehicle relative to the rotating atmosphere is given by

$$\vec{V}_T = u\,\hat{r} + v\,\hat{\theta} + w\,\hat{\phi} \tag{3-1}$$

It is convenient at times to describe the velocity vector by its magnitude V_T, its elevation angle above horizontal, γ, (actually above an r = constant surface), and its azimuth angle (or heading) counter clockwise from south, β. Thus, from Figure 4.

$$\begin{aligned} V_T &= \sqrt{u^2 + v^2 + w^2} \\ \tan\beta &= \frac{w}{v} \\ \tan\gamma &= \frac{u}{\sqrt{v^2 + w^2}} = \frac{u}{V_H} \\ V_H &= \sqrt{v^2 + w^2} \end{aligned} \tag{3-2}$$

The relationship between position and velocity components is

$$\begin{aligned} \dot{r} &= u \\ \dot{\theta} &= \frac{v}{r} \\ \dot{\phi} &= \frac{w}{r \sin\theta} \end{aligned} \tag{3-3}$$

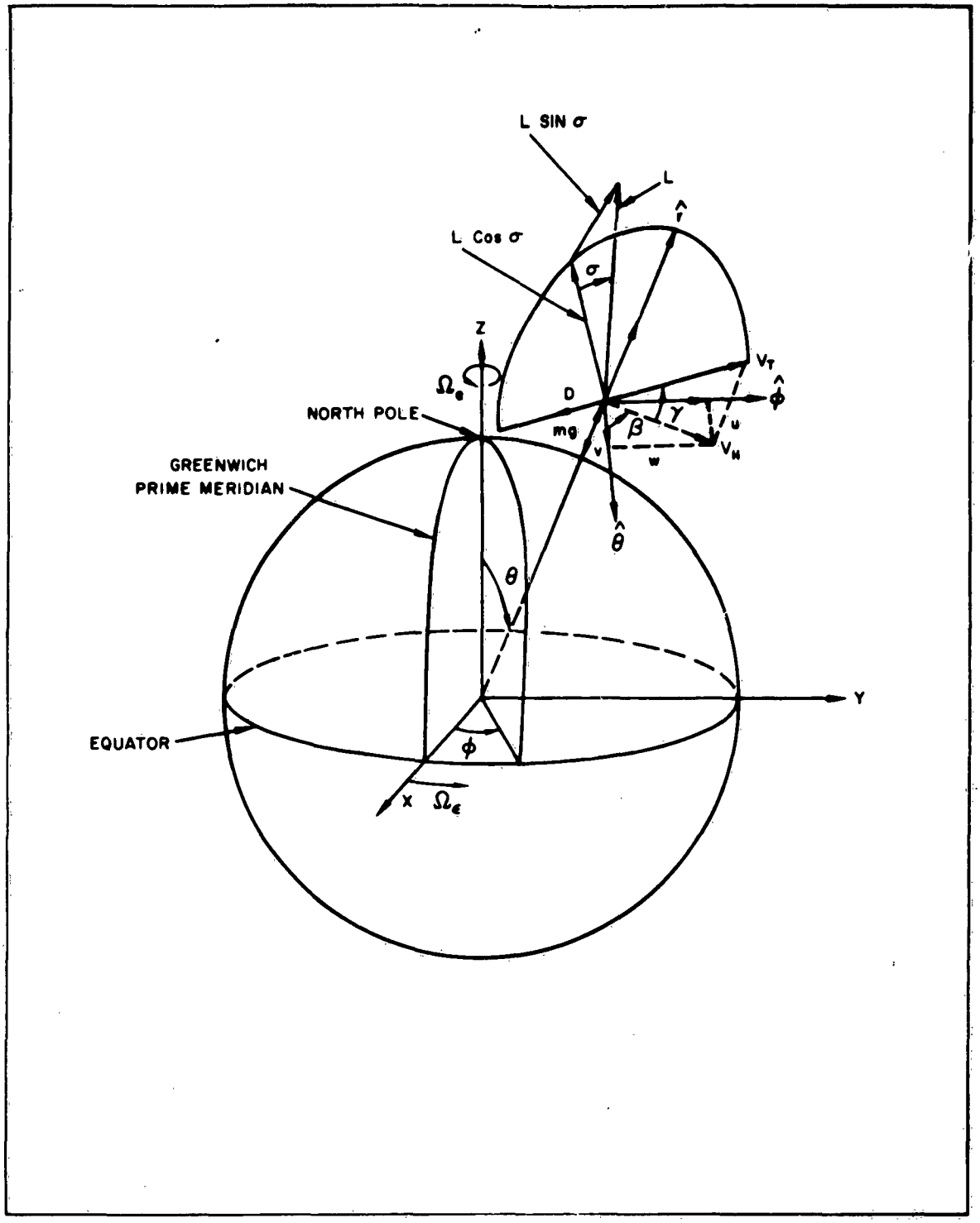


Figure 3 - The Coordinate System and Forces of the Problem

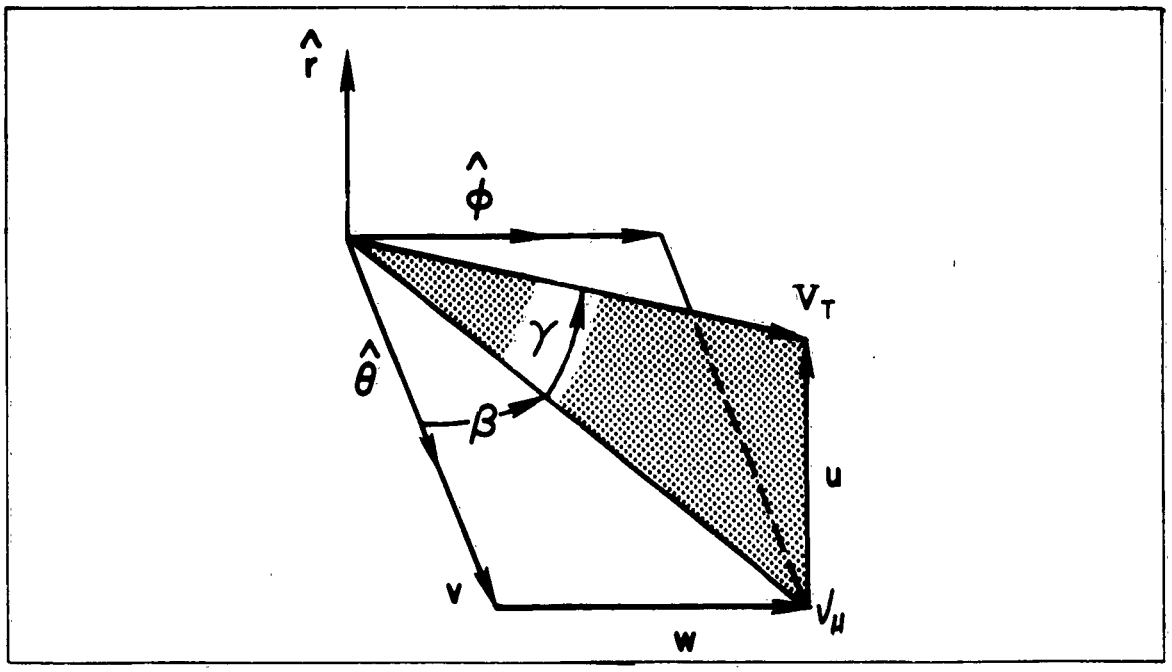


Figure 4 - Velocity Components u, v, w or V_T, γ, β

The dynamic equations for the mass center of the vehicle are given as:

$$\dot{u} = \frac{1}{r}(v^2 + w^2) + r\,\Omega_e{}^2 \sin^2\theta + 2\,\Omega_e\, w \sin\theta + \frac{1}{m} F_r$$

$$\dot{v} = \frac{1}{r}(w^2 \cot\theta - uv) + r\Omega_e{}^2 \sin\theta\cos\theta + 2\,\Omega_e\, w\cos\theta + \frac{1}{m} F_\theta \qquad (3\text{-}4)$$

$$\dot{w} = -\frac{1}{r}(vw\cot\theta + uw) - 2\Omega_e\,(u\sin\theta + v\cos\theta) + \frac{1}{m} F_\phi$$

where the components of the force vector

$$\vec{F} = F_r\,\hat{r} + F_\theta\,\hat{\theta} + F_\phi\,\hat{\phi} \qquad (3\text{-}5)$$

are given by

$$F_r = -m\left[\frac{\mu_e}{r^2} + \frac{\mu_e JR_f^2}{r^4}(1 - 3\cos^2\theta)\right] + A_r + T_r$$

$$F_\theta = \frac{m2\mu_e JR_f^2}{r^4}\cos\theta\sin\theta + A_\theta + T_\theta \qquad (3\text{-}6)$$

$$F_\phi = A_\phi + T_\phi$$

The thrust components T_r, T_θ and T_ϕ are given in terms of the components of thrust along the vehicle axes by the transformation:

$$\begin{bmatrix} T_r \\ T_\theta \\ T_\phi \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} \qquad (3\text{-}7)$$

The a_{ij}'s are functions of β, γ, σ, and α, given in the Appendix. Two constant offset angles (i_T, δ_T) orient the thrust vector relative to the vehicle axis system. The thrust components in the vehicle axis system are given as

$$T_x = T\cos i_T \cos\delta_T$$

$$T_y = T\cos i_T \sin\delta_T \qquad (3\text{-}8)$$

$$T_z = T\sin i_T$$

where T is a known function of time and stage number.

The aerodynamic force components A_r, A_θ, and A_ϕ are given in terms of force components perpendicular and parallel to the velocity vector (lift and drag) by the transformation:

$$\begin{bmatrix} A_r \\ A_\theta \\ A_\phi \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \begin{bmatrix} L \\ -D \end{bmatrix} \tag{3-9}$$

The b_{ij}'s are functions of β, γ, and σ, given in the Appendix.

Derivations of the thrust transformation matrix, and the aerodynamic force transformation matrix are given in the Appendix.

3.2 In-Flight Constraints on Altitude and Heating Rate

Two in-flight constraints are considered, either one of which may be used in any given problem:

1) An upper bound on the altitude, h, during the flight.
2) An upper bound on the heating rate at the stagnation point, $\dot{q}$, during the flight.

These constraints are handled by the use of penalty functions:

$$\dot{P} = \begin{cases} F_p = A_{p1} \left\{ \dfrac{h - A_{p2}}{A_{p2}} \right\}^2 ; & h > A_{p2} \\ 0 ; & h < A_{p2} \end{cases} \tag{3-10}$$

$$\dot{P} = \begin{cases} F_P = A_{P1} \left\{ \dfrac{\dot{q} - A_{P2}}{A_{P2}} \right\}^2 ; & \dot{q} > A_{P2} \\ 0 ; & \dot{q} < A_{P2} \end{cases} \tag{3-11}$$

where

$$\dot{q} = \rho^{1/2} V_T^3$$

See Reference(2) for details about penalty functions.

If the nominal trajectory exceeds the desired upper bound on altitude (say) then the terminal value of p will be positive. The successive-improvement scheme will reduce this terminal value of p to a small positive quantity which will cause the final trajectory to exceed the desired upper bound on altitude by only a negligibly small amount.

3.3 Terminal Constraints on Pilot Acceleration Dose and Total Heat Absorbed

A measure of the acceleration dose absorbed by a human pilot subjected to an acceleration history a(t) is given by

$$\dot{e} = F_e = \frac{1}{\tau(a)} \tag{3-12}$$

where $\tau(a)$ is the time (empirically measured) that a human pilot can stay usefully conscious under a constant acceleration level a. Whenever e exceeds a value of unity, it is assumed that the pilot has had an excessive dose of acceleration and cannot any longer perform a useful piloting operation. Figure 5 shows a plot of τ vs a using data from Reference (3).

An approximate measure of the rate of heat absorption at the stagnation point of an ablative nose is given by

$$\dot{q} = F_q = \frac{\chi}{\sqrt{R_s}} \rho^{1/2} V_T^3 = A_{\dot{q}} \rho^{1/2} V_T^3 \tag{3-13}$$

where $A_{\dot{q}}$ is a constant. This relation does not include the heat due to radiation from the hot gases surrounding the nose. It should be noted that $A_{\dot{q}}$ can be specified so that $\dot{q}$ will be a normalized value. For further details on the heat absorption equation see, for example, Reference (4).

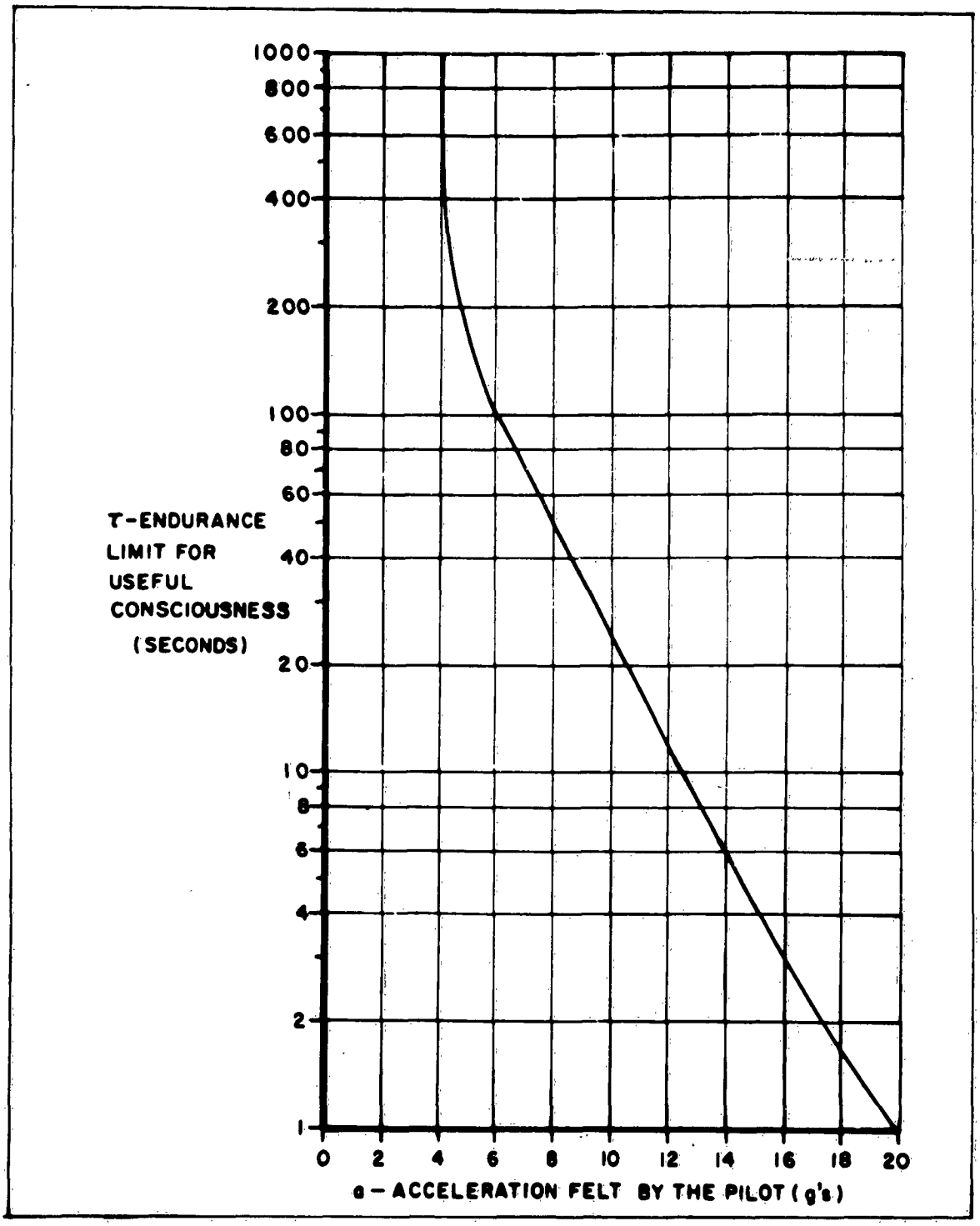


Figure 5 - Limiting Acceleration Dose for Useful Consciousness, τ vs a

3.4 Summary of State Equations

The eight first-order non-linear differential equations of state are summarized below:

$$\dot{u} = \frac{V_H^2}{r} + r\,\Omega_e^2 \sin^2\theta + 2\,\Omega_e\, w \sin\theta + \frac{\rho S}{2m}\left[-C_D\, u\, V_T + \cos\sigma\, C_L\, V_H\, V_T\right] - \frac{\mu_e}{r^2} - \frac{\mu_e J R_f^2}{r^4}\left(1 - 3\cos^2\theta\right) + \frac{T_r}{m} \tag{3-14}$$

$$\dot{v} = -\frac{uv}{r} + \frac{w^2 \operatorname{ctn}\theta}{r} + r\,\Omega_e^2 \sin\theta\cos\theta + 2\,\Omega_e\, w\cos\theta - \frac{\rho S}{2m}\left[C_D\, v\, V_T + \cos\sigma\, C_L \frac{u\, v\, V_T}{V_H} + \sin\sigma\, C_L \frac{w\, V_T^2}{V_H}\right] + 2\,\frac{\mu_e J R_f^2}{r^4}\cos\theta\sin\theta + \frac{T_\theta}{m} \tag{3-15}$$

$$\dot{w} = -\frac{uw}{r} - \frac{v\, w \operatorname{ctn}\theta}{r} - 2\,\Omega_e\left(u\sin\theta + v\cos\theta\right) - \frac{\rho S}{2m}\left[C_D\, w\, V_T + \cos\sigma\, C_L \frac{u\, w\, V_T}{V_H} - \sin\sigma\, C_L \frac{v\, V_T^2}{V_H}\right] + \frac{T_\phi}{m} \tag{3-16}$$

$$\dot{r} = u \tag{3-17}$$

$$\dot{\theta} = \frac{v}{r} \tag{3-18}$$

$$\dot{\phi} = \frac{w}{r\sin\theta} \tag{3-19}$$

$$\dot{e} = F_e \tag{3-20}$$

$$\dot{q} = F_q \tag{3-21}$$

where

m = vehicle mass

$\mu_e = g_o R_e^2 = 1.407698 \times 10^{16}\ ft^3\ sec^{-2}$

$R_f = 20.925631 \times 10^6\ ft$

$J = 1623.41 \times 10^{-6}$

$\Omega_e = 7.29211508 \times 10^{-5}\ rad/sec$

$L = C_L(\alpha, M) \frac{\rho V_T^2}{2} S$ = lift

$D = C_D(\alpha, M) \frac{\rho V_T^2}{2} S$ = drag

$\rho = \rho(h)$ = air density ARDC tables

S = reference area

$M = \frac{V_T}{c}$

4. OPTIMUM PROGRAMMING PROBLEMS AND THEIR SOLUTIONS

4.1 Optimum Programming Problems

A process which develops in space and/or time in which one or more control variables must be programmed to optimize a performance criterion is an optimum programming problem. The problem is to determine, out of all possible programs for the control variables, the one program which maximizes (or minimizes) the performance criterion (often one of the thermal quantities) while simultaneously yielding specified values of certain terminal quantities and satisfying the in-flight constraints of the system.

A formulation of a general class of optimum programming problems is given below:

Let the system or process be described by n first-order non-linear (or linear) ordinary differential equations of the form

$$\frac{dx_i}{dt} = \dot{x}_i = f_i\,(x_1, x_2, \ldots, x_n, \alpha_1, \alpha_2, \ldots, \alpha_m, t) \qquad (4\text{-}1)$$

where $i = 1, 2, 3, \ldots, n$, and the control variables are given by α_j, $j = 1, 2, 3, \ldots, m$, which are free to be selected. The problem is to determine the α_j's in the interval $t_o \leq t \leq T$, such that some quantity

$$\Phi = \Phi\,\left[\, x_i(T),\ T \,\right] \qquad (4\text{-}2)$$

is a maximum (or a minimum), subject to satisfying the physical equations given by Equation (4.1) and certain other specified terminal constraints

$$\psi = \begin{bmatrix} \psi_1 \ (x_i(T),\ T) \\ \psi_2 \ (x_i(T),\ T) \\ \vdots \\ \psi_p \ (x_i(T),\ T) \end{bmatrix} \tag{4-3}$$

$$\Omega = \Omega \ (x_i \ (T),\ T \ ) = 0 \tag{4-4}$$

All the terminal constraint functions are known functions of the state variables and time. The stopping condition, Ω, determines the final time T and is a known function of x_i (T) and T.

4.2 Classical Approach to Solving Optimum Programming Problems

The classical approach to solving optimum programming problems is by the use of the calculus of variations. In the classical formulation, these problems are two-point boundary value problems for a set of non-linear ordinary differential equations. The boundary conditions for the physical equations are usually specified at the initial point, and the boundary conditions for the Euler-Lagrange equations are usually specified at the terminal point. Thus, solving the set of equations numerically requires that a "guess" be made for the initial values of the Lagrange multipliers in the Euler-Lagrange equations and the two sets of equations must then be integrated to the terminal point. A check of the terminal boundary conditions is then made to see how badly they were missed. A correction must then be made to the "guessed" initial values to improve agreement with the desired terminal boundary conditions. This process must be repeated until all of the terminal boundary conditions are satisfied. This process is not only tedious, expensive and frustrating, but great patience must be exercised before any worthwhile results can be expected.

4.3 Steepest-Ascent Technique for Solving Optimum Programming Problems

A systematic and efficient method of solving an optimum programming problem on a high-speed digital computer using the steepest-ascent technique will be described.

Recently, a little-known procedure for determining numerical solutions to complicated optimum programming problems has been revived by Kelley (Reference 2) and Bryson et al (References 6, 7 and 8). These references present a number of examples using the steepest-ascent technique.

In this method the control functions $\alpha_j(t)$ are guessed and a trajectory is computed. The equations adjoint to the equations describing small perturbations about this trajectory are integrated backwards over the trajectory with suitable boundary conditions. The results of this integration yield the impulse-response functions of the final conditions to small changes in the $\alpha_j(t)$ programs. By suitably combining these impulse-response functions, changes in the $\alpha_j(t)$ programs can be made which will simultaneously increase the final value of the optimizing function and approach more closely the desired values of the terminal constraints. This process is repeated until it becomes apparent that only small changes are occurring in the optimizing function and for all practical engineering purposes this last trajectory is the optimum trajectory.

4.4 Adjoint Equations for Steepest-Ascent Optimization

The influence (adjoint) functions are devoted by

$$\lambda_u,\ \lambda_v,\ \lambda_w,\ \lambda_r,\ \lambda_\theta,\ \lambda_\phi,\ \lambda_e,\ \lambda_q \tag{4-5}$$

where the subscript refer to the physical variable associated with that particular influence function. The eight first-order non-linear differential equations of state given in Section 3.4 can be written in the shorthand form of Equation (4-1) where

$$(x_1,\ x_2,\ x_3,\ x_n,\ \alpha_1,\ \alpha_2, \ldots, \alpha_m) = (u,\ v,\ w,\ r,\ \theta,\ \phi,\ e,\ q,\ \alpha,\ \sigma) \tag{4-6}$$

The adjoint differential equations can be written as:

$$\dot{\lambda}_i + \sum_{j=1}^{8} \frac{\partial f_j}{\partial x_i} \lambda_j = 0, \quad i = 1,\ 2,\ 3, \ldots,\ 8 \tag{4-7}$$

where $(\lambda_1, \lambda_2, \ldots \lambda_8) = (\lambda_u, \lambda_w, \lambda_r, \lambda_\theta, \lambda_\phi, \lambda_e, \lambda_q)$

And the impulse response functions for the angle of attack and the bank angle, α and σ, respectively, are

$$\sum_{j=1}^{8} \frac{\partial f_j}{\partial \alpha} \lambda_j \text{ and } \sum_{j=1}^{8} \frac{\partial f_j}{\partial \sigma} \lambda_j \tag{4-8}$$

The ten relationships given by Equations (4-7) and (4-8) are quite lengthy when written out but can be obtained by straight forward differentiation. Hence, the ten relationships are not given here, however, they are written out in full Reference (10), which also gives details on the boundary conditions and the computational procedures used, as well as other details on the computer programs.

5. EXAMPLE PROBLEM-THE PERFORMANCE BOUNDARIES FOR A SORTIE CONFIGURATION

5.1 The Example Problem

As an example three separate but related performance boundaries for a typical three-stage booster system with a re-entry vehicle as its payload is presented. This example demonstrates the main features of computer programs developed to solve trajectory optimization problems. The first part demonstrates the practical way of determining control programs for a booster system with specified terminal conditions. The second part of the example presents an example problem with an in-flight constraint during the re-entry phase for near escape speeds. Finally, the third part presents the computer program capability to determines lateral maneuvers for a gliding vehicle. Each of these examples are presented in turn.

6. THE VEHICLE STAGE AND AERODYNAMIC DATA

The booster used in calculating the SORTIE performance was the Atlas-Agena. Figure 6 shows the lift and drag characteristics for the vehicle during first stage and second stage boost. Figure 7 shows the lift and drag characteristics for the third stage boost and for the re-entry SORTIE vehicle. Figure 8 shows a sketch of the SORTIE vehicle and Figure 9 presents the L/D ratio vs α for this vehicle. $(L/D)_{max}$ is equal to 1.59 at $\alpha = 17.5°$.

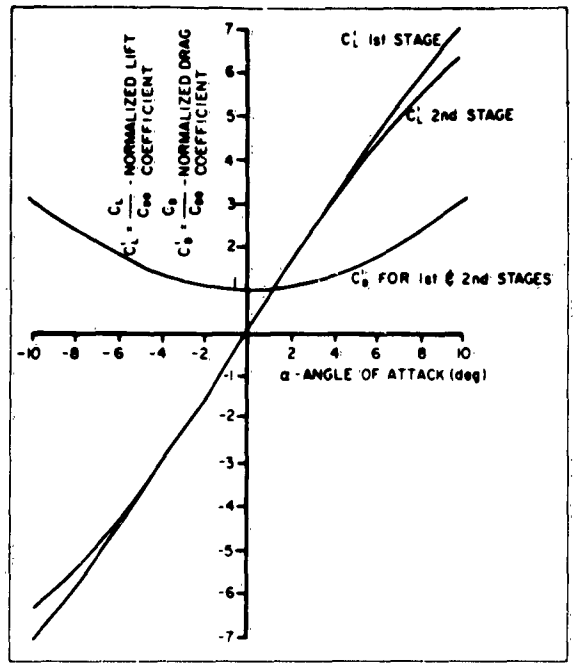


Figure 6 - C'_L and C'_D vs α for the SORTIE Configuration

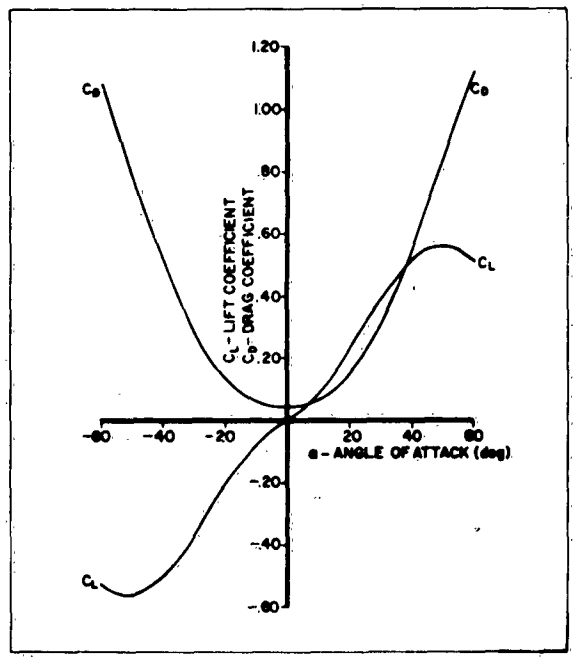


Figure 7 - C_L and C_D vs α for the Third Stage Booster and the SORTIE Vehicle

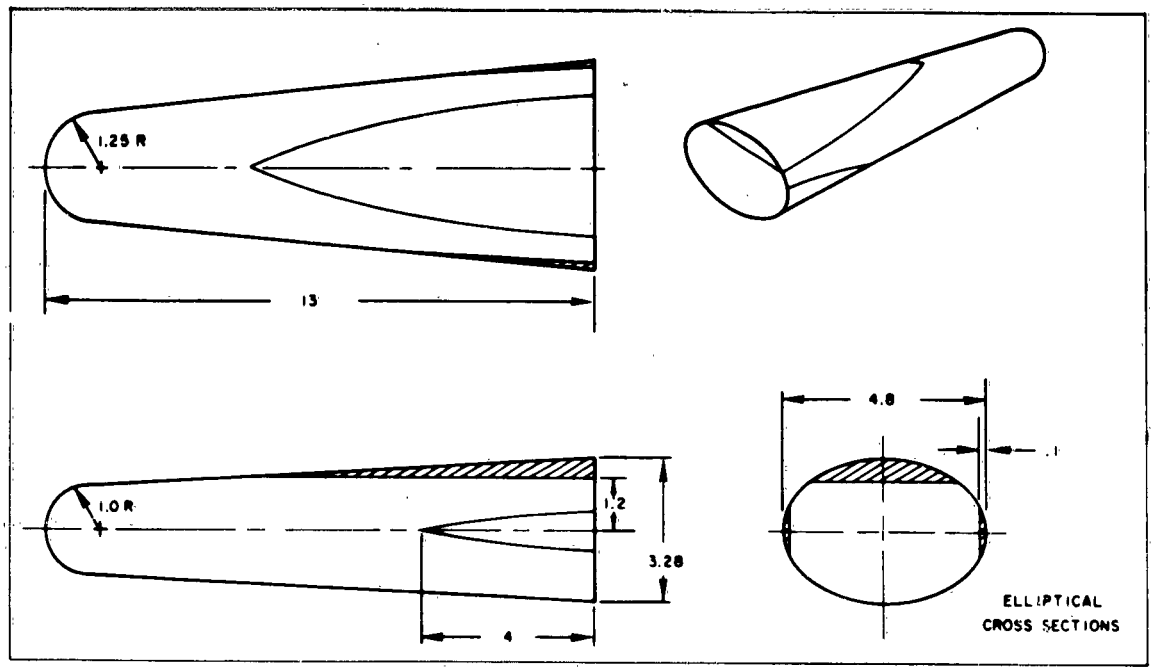


Figure 8 - SORTIE Vehicle

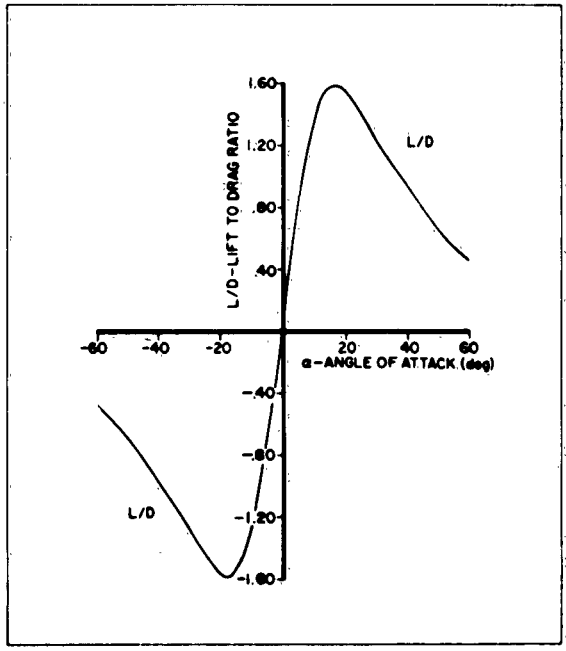


Figure 9 - L/D vs α for SORTIE Vehicle

7. BOOST TRAJECTORY

The trajectory optimization computer program for vehicles with thrust has been demonstrated by determining the maximum velocity at burnout with specified terminal values for the altitude and flight-path angle for the three-stage booster system. The specified terminal conditions were h_f = 400,000 ft and γ_f = -5 degrees. The initial conditions for the optimization procedure were determined by integrating the equations of motion, neglecting the effects of drag, for a period of 15 seconds for a vertically launched vehicle. The following conditions were used as the initial conditions for the optimization of the launch trajectory:

V_o = 190 ft/sec

γ_o = 89°

h_o = 1300 ft

β_o = 70°

θ_o = 61.5°

ϕ_o = -80.5°

The angle-of-attack program for the nominal trajectory was taken from a preliminary series of boost trajectories which had been run to check out the computer program. It should be noted that the terminal values of h and γ for the nominal case were at least one order of magnitude too large. The nominal trajectory had the following terminal conditions:

V_f = 28,530 ft/sec

h_f = 4,603,350 ft

γ_f = 32.39°

The computer program in 34 iterations, very effectively, reduced the terminal values of h and γ to within 90% of the specified values. The terminal conditions at the end of 34 iterations were

$$V_f = 35,620 \text{ ft/sec}$$

$$h_f = 430,579 \text{ ft}$$

$$\gamma_f = -4.96°$$

with an increase of 7090 ft/sec in the terminal velocity, nearly a 25% improvement over the nominal case. The altitude differential has been reduced from 4,203,350 ft to 430,579 ft and the flight-path angle is within .04 degree of the specified value. Two additional series of iterations were made and the results of these runs are presented in Table 1.

TABLE 1
SUMMARY OF SUCCESSIVE IMPROVEMENTS FOR THREE-STAGE BOOST TRAJECTORY

Case	V_{Tf}-ft/sec^2	γ_f-deg	h_f-ft	$\frac{(dP)^2}{T-t_o}$ - deg^2
Nominal	28,530	32.391	4,603,350	1.0×10^{-1}
34	35,620	-4.963	430,579	1.0×10^{-2}
36	35,665	-4.974	404,696	1.0×10^{-3}
38	35,700	-4.991	401,559	1.0×10^{-4}

For all practical purposes, the 38th iteration is an optimum trajectory having maximum terminal velocity for the specified terminal values on altitude and flight-path angle.

Figure 10 shows h, γ, V_T, and α vs t for the nominal and optimum trajectories. The angle-of-attack program for the optimum trajectory is a modest program, starting from 0 decreasing slowly to -7 degrees and increasing to 2 degrees before falling off slowly to -17.5 degrees at the end of burnout of the three-stage. A maximum altitude of 630,000 ft occurred 370 seconds after launch for the optimum trajectory.

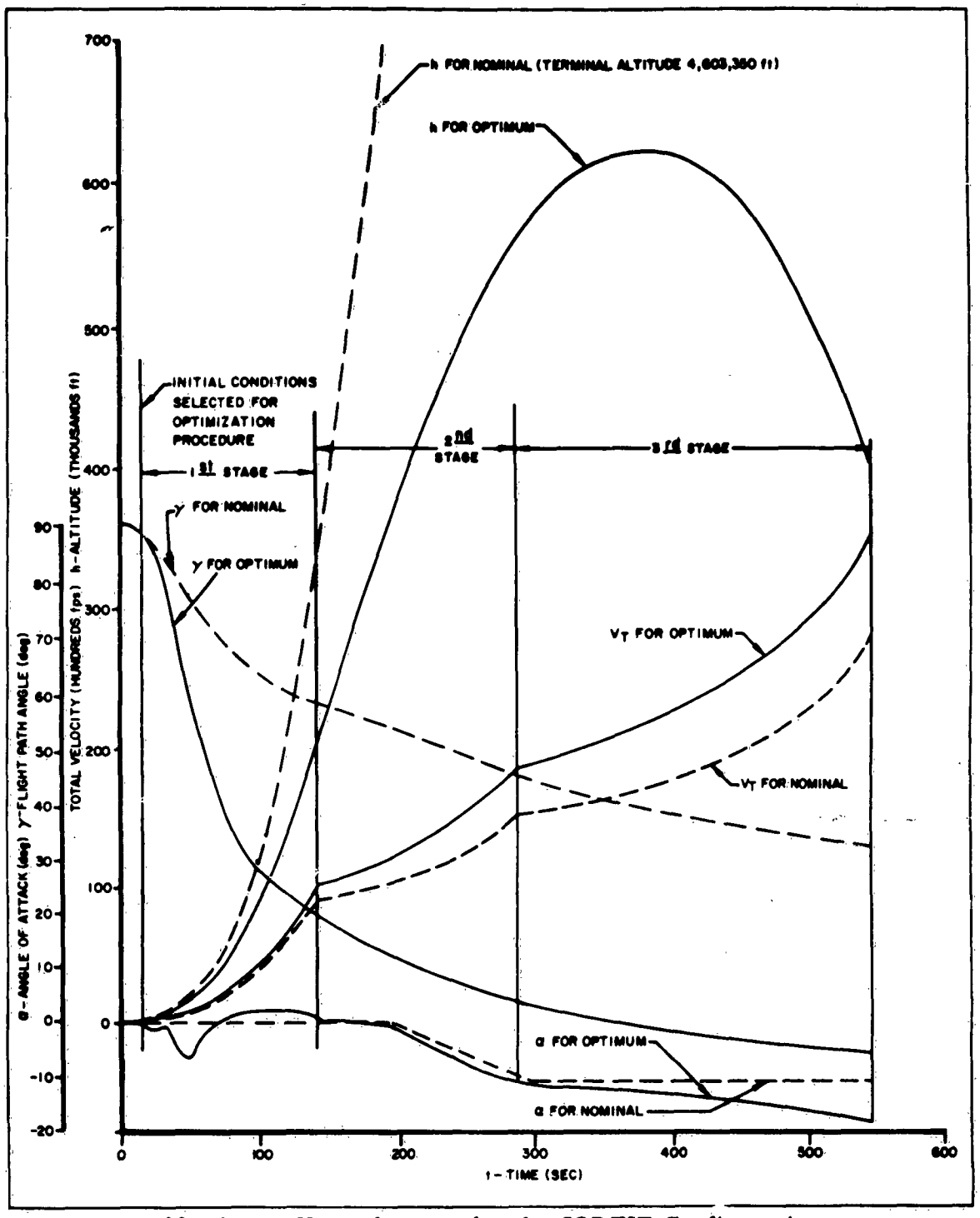


Figure 10 - h, γ, V_T and α vs t for the SORTIE Configuration for Maximum Re-Entry Velocity for Terminals Conditions of Altitude of 400,000 ft and Flight-Path Angle of -5°

This example has demonstrated how efficiently and effectively the trajectory optimization procedure can determine the optimum trajectory, maximizing the terminal velocity and meeting the specified terminal conditions, even though the terminal constraints on the nominal trajectory differed by an order of magnitude from the specified values.

8. SORTIE GLIDING RE-ENTRY WITH A STAGNATION-POINT HEATING RATE IN-FLIGHT CONSTRAINT

The second part of the demonstration example considered the problem of determining the re-entry glide trajectory with minimum stagnation-point heating rate penalty function, P, for the SORTIE vehicle. (See Section 3.2 for details on in-flight constraints.) This example has demonstrated the trajectory optimization computer program for gliding vehicles with an in-flight constraint for the following initial conditions:

V_o = 34,000 ft/sec

γ_o = -5°

h_o = 400,000 ft

θ_o = 70°

ϕ_o = -60°

β_o = 70°

One of the conditions of the re-entry trajectory was that the vehicle reenter within the Atlantic Missile Range; thus, trajectories which require more than one pass over the range were not considered. The re-entry angle of -5 degrees was determined from a series of trajectories in which the re-entry angle was varied for a fixed re-entry velocity of 34,000 ft/sec. The re-entry velocity of 34,000 ft/sec has been specified by ASD as a possible lower limit for the SORTIE test vehicle.

The re-entry angle (for fixed re-entry velocity and altitude) determines a peak value for the stagnation-point heating rate. The flight-path angle, γ, is principally dependent upon the centrifugal and gravity forces acting upon the vehicle during the initial phase of re-entry. Aerodynamic forces of drag and lift do not affect the vehicle motion until the vehicle enters appreciably into the atmosphere. Thus, during the initial phase of re-entry (approximately the first 90 seconds in this problem), the flight-path-angle program is solely dependent upon the initial velocity, flight-path angle, and altitude. Therefore, there is an

initial flight angle (for every initial velocity and and altitude) for which the vehicle can just stay in with the application of maximum negative (negative is defined in the $-\hat{r}$ direction) lift. For larger (less negative) initial flight-path angles the vehicle will skip out and a single-pass re-entry within the Atlantic Missile Range is not possible. For these initial re-entry conditions, the peak value of the stagnation-point heating rate, $\dot{q}$, will be determined and will occur at the point where the effect of lift enters. Thus, associated with every initial velocity and re-entry angle, there will be a fixed value for $\dot{q}$.

Once lift can be used to alter the flight-path angle, a modulated lift reentry is possible. The minimum flight-path angle for reentry in a single-pass trajectory was found to be in the neighborhood of -5° for the specified initial conditions. This value was selected as being well within the re-entry capability of the vehicle, but not so sensitive as to allow the vehicle to skip out with small changes in the angle-of-attack program. The angle-of-attack program for the nominal trajectory was a positive constant angle-of-attack (maximum lift) until the flight-path angle reached a value close to zero, at which point the vehicle's maximum negative lift capability is utilized to keep it from skipping out. This negative lift is used for the remainder of the flight, and the bank angle is used to

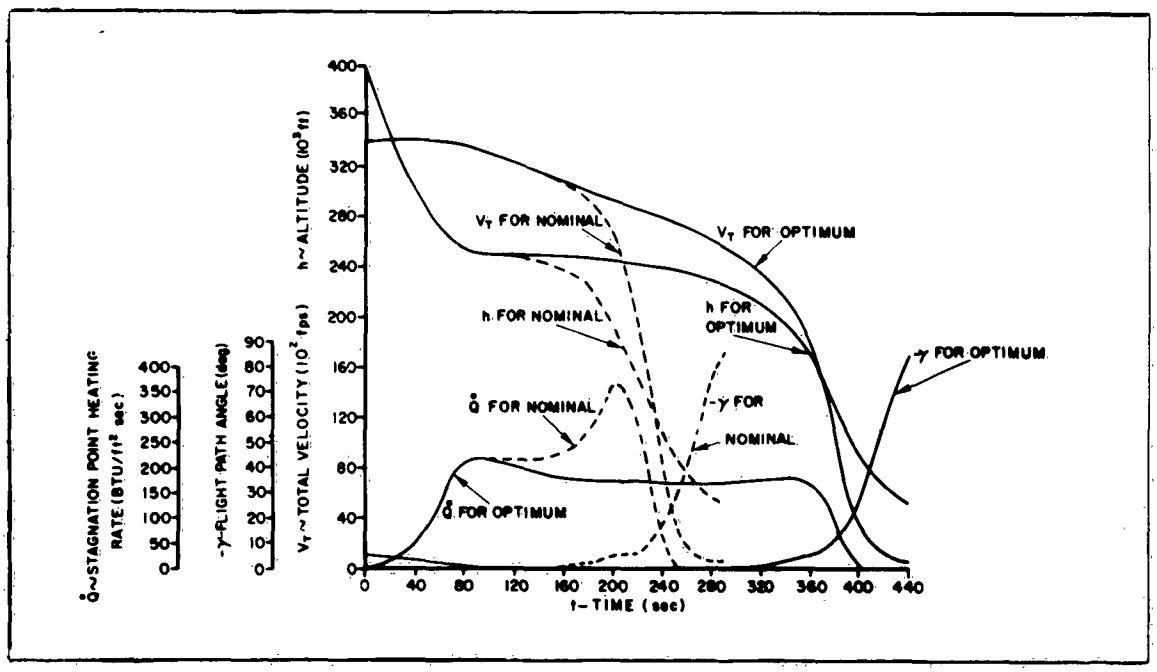


Figure 11 - h V_T, γ and $\dot{q}$ vs t for SORTIE Re-Entry which Minimizes Stagnation-Point Heating Rate Penalty Function

modulate the negative lift component. A bank-angle program to keep the vehicle as near as possible on the original heading angle was used; however, the vehicle did wander off course.

Figure 11 shows the velocity, altitude, flight-path angle and the stagnation-point heating rate as a function of time. The stagnation heating rate for the nominal case peaked at 371 BTU/ft^2 sec 205 seconds after reentry. This shows that the lift program could be improved upon as well as reduce the value of the penalty function for stagnation-point heating rate from a value of 29. 2. The penalty function level, $\dot{q}_o$, was set at 186 BTU/ft^2 sec, a value considerably below the 220 BTU/ft^2 sec peak reached at the bottom of the first skip-out point. After 21 iterations, the penalty function P was reduced to 18. 3 and the maximum heating rate has dropped to 302 BTU/ft^2 sec. In 7 additional iterations, $\dot{q}_{max}$ was reduced to 218 BTU/ft^2 sec and P to a value of 1. $\dot{q}$ drops from the peak value of 218 to approximately 175 and remains close to this value until the velocity drops to about 18,000 ft/sec and the altitude at this point is approximately 180,000 ft. From this point on, the stagnation-point heating rate drops rapidly as the velocity drops. A summary of these computations are given in Table 2.

TABLE 2

SUMMARY OF SUCCESSIVE IMPROVEMENTS FOR SORTIE RE-ENTRY GLIDE

Case	$\dot{q}_{max}$ - BTU/ft^{2}sec	P($\dot{q}_o$ = 186 BTU/ft^{2}sec)	$(dP)^2/T$ - deg^2
nominal	371	29. 2	1×10^{-1}
21	302	18. 3	5×10^{-2}
28	218	. 99	5×10^{-2}

Figure 12 shows the angle-of-attack programs for these re-entry trajectories. Note that the angle-of-attack program has had only minor changes for the first 90 seconds of flight; however, these small changes are important for they determine the peak value for the stagnation-point heating rate. The "spike" appearing in the angle-of-attack program at 200 seconds appears to be due to an error in input for the angle of attack in the nominal program. The value used in the nominal program was +50 degrees and the final trajectory has smoothed out this value to about -8 degrees. The angle-of-attack changes are again modest, though important. The bank-angle program is presented in Figure 13.

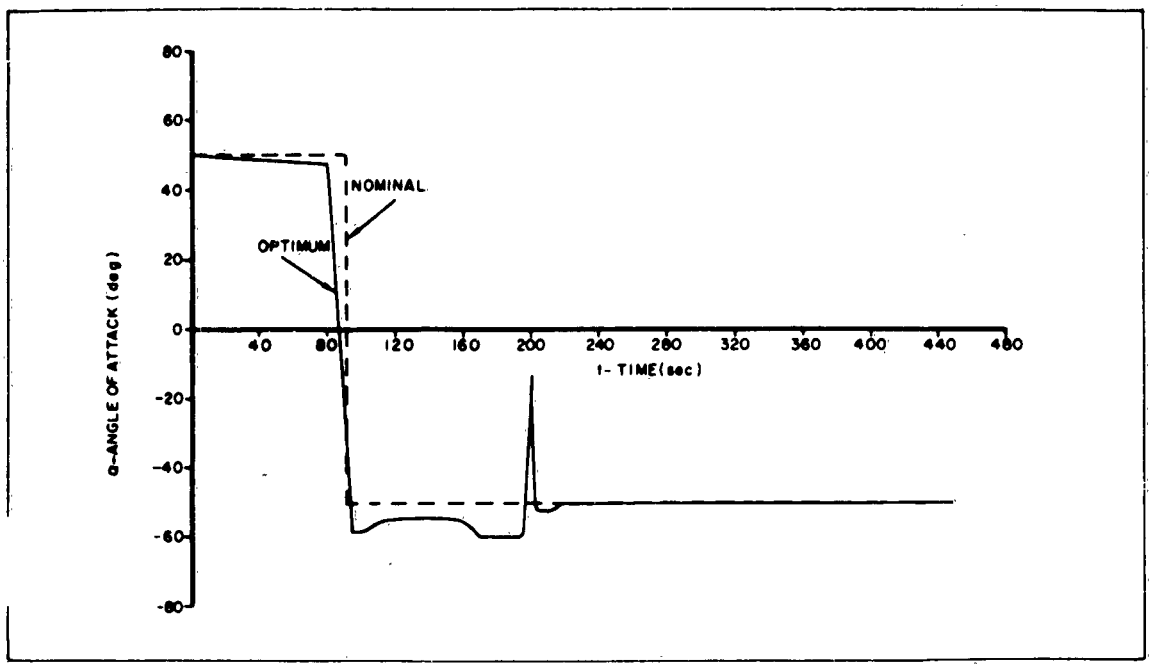


Figure 12 - α vs t for SORTIE Re-Entry which Minimizes Stagnation-Point Heating Rate Penalty Function

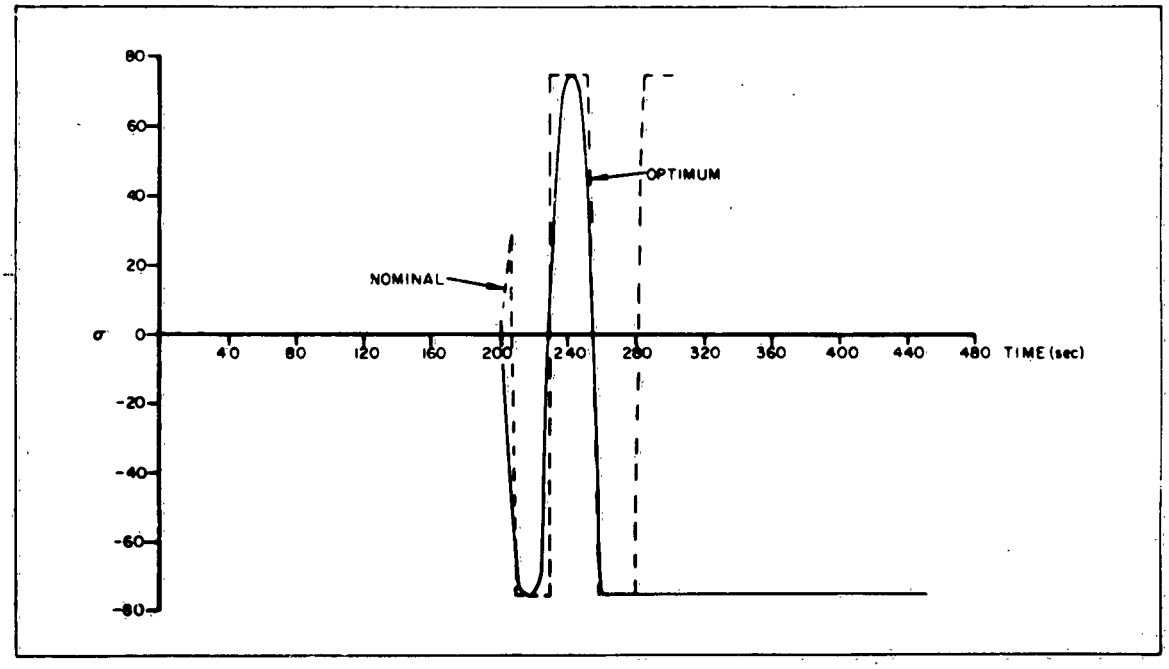


Figure 13 - σ vs t for SORTIE Re-Entry which Minimizes Stagnation-Point Heating Rate Penalty Function

The nominal σ program was modulated to minimize vehicle wander from the original heading. The optimum bank-angle program has had a significant change over the nominal one, especially during the time period of 180 to 230 seconds. This time period corresponds to peak stagnation-point heating rate for the nominal trajectory. Note that the bank angle can be used to modulate the component of lift in the plane of the velocity and radius vectors to maintain the desired relation of V_T and γ to minimize the stagnation-point heating rate. Since $\dot{q}$ peaked during this period on the nominal trajectory, the assumed bank-angle program was not a good one and large changes in this program could have been expected to reduce P to a minimum value. To reduce the peak $\dot{q}$, the bank angle would have to be changed such that the flight-path angle decreased as the velocity dropped. The velocity drop depends upon the drag of the vehicle, and thus, high values of angle of attack are needed to achieve maximum velocity changes. The increment of lift needed to change the flight-path angle a desired amount is achieved by the proper bank angle and bank-angle modulation determines the minimum value for the heating-rate penalty function.

9. RECOVERY AREA FLIGHT BOUNDARY FOR SORTIE

The recovery area for the SORTIE vehicle has been determined for the initial condition given by

V_o = 15,282 ft/sec
h_o = 151,617 ft
γ_o = -7.29°
β_o = 83.96°
θ_o = 81.31° (latitude = 8.69°)
ϕ_o = -32.34°

The initial conditions were arbitrarily selected from the re-entry trajectory given in Section 8.

Three optimum trajectories were calculated for these initial conditions: maximum downrange, θ^*, maximum, and minimum lateral latitude, ν. The stopping condition was altitude, h_f = 1,000 ft. Figure 14 shows these trajectories on a geographic Mercator projection and a "footprint" is sketched through the three points. Figure 15 shows the same trajectories on a Mercator projection of the auxiliary coordinate system μ vs ν, and Figures 16 and 17 show the control function histories α and σ vs t for the three extremal trajectories. The angle of attack for the maximum downrange trajectory is close to the value for maximum lift-to-drag ratio, namely α = 17.5°.

The maximum and minimum lateral latitude trajectories use essentially the angle of attack for maximum lift-to-drag ratio. The bank-angle program reaches values slightly less than 50 degrees and then gradually tapers off to zero when the turn angle approaches 90°. For comparison purposes, trajectories with constant bank angles were run and the dashed footprint in Figures 14 and 15 show the results of these runs. All of the constant-bank-angle trajectories show a spiralling tendency near the terminal portion of the flight. Two other trajectories are shown for the case of $\sigma = \pm 45°$, in which cases the bank angle was set equal to zero when the turn angle reached 90°. These two trajectories were used as the nominals to determine the maximum and minimum lateral latitudes. Figures 18 through 20 show the altitude, velocity and flight-path angle as a function of time for the three sets of trajectories.

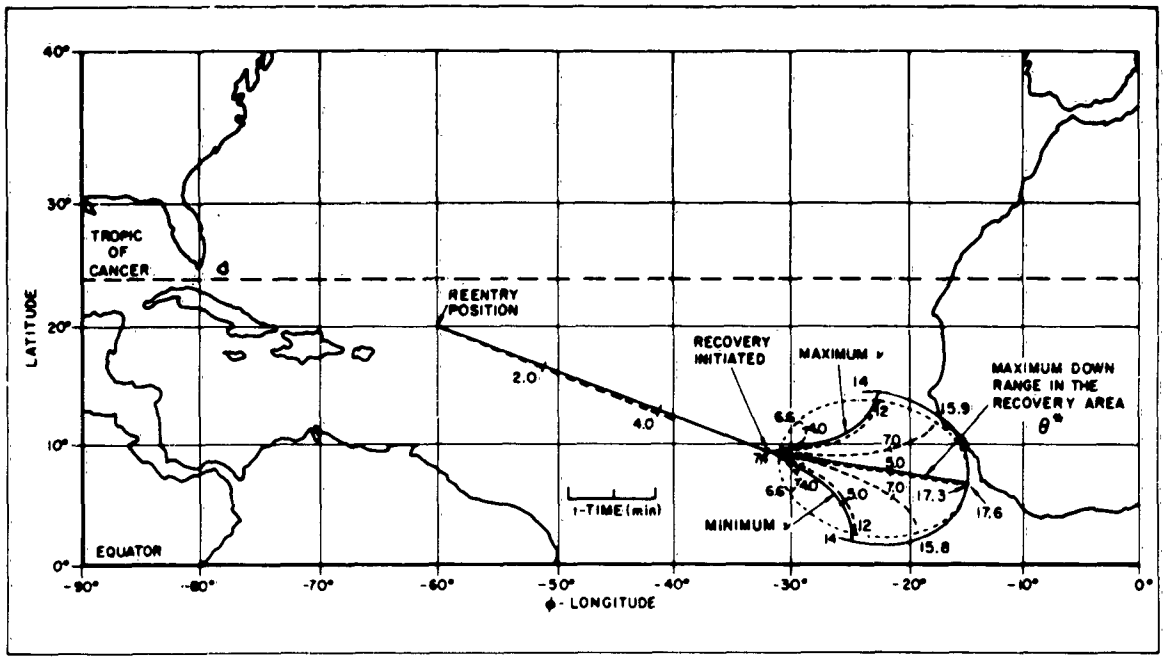


Figure 14 - Geographic Mercator Projection of Atlantic Missile Range Flight Boundary for the SORTIE Reentry and Recovery

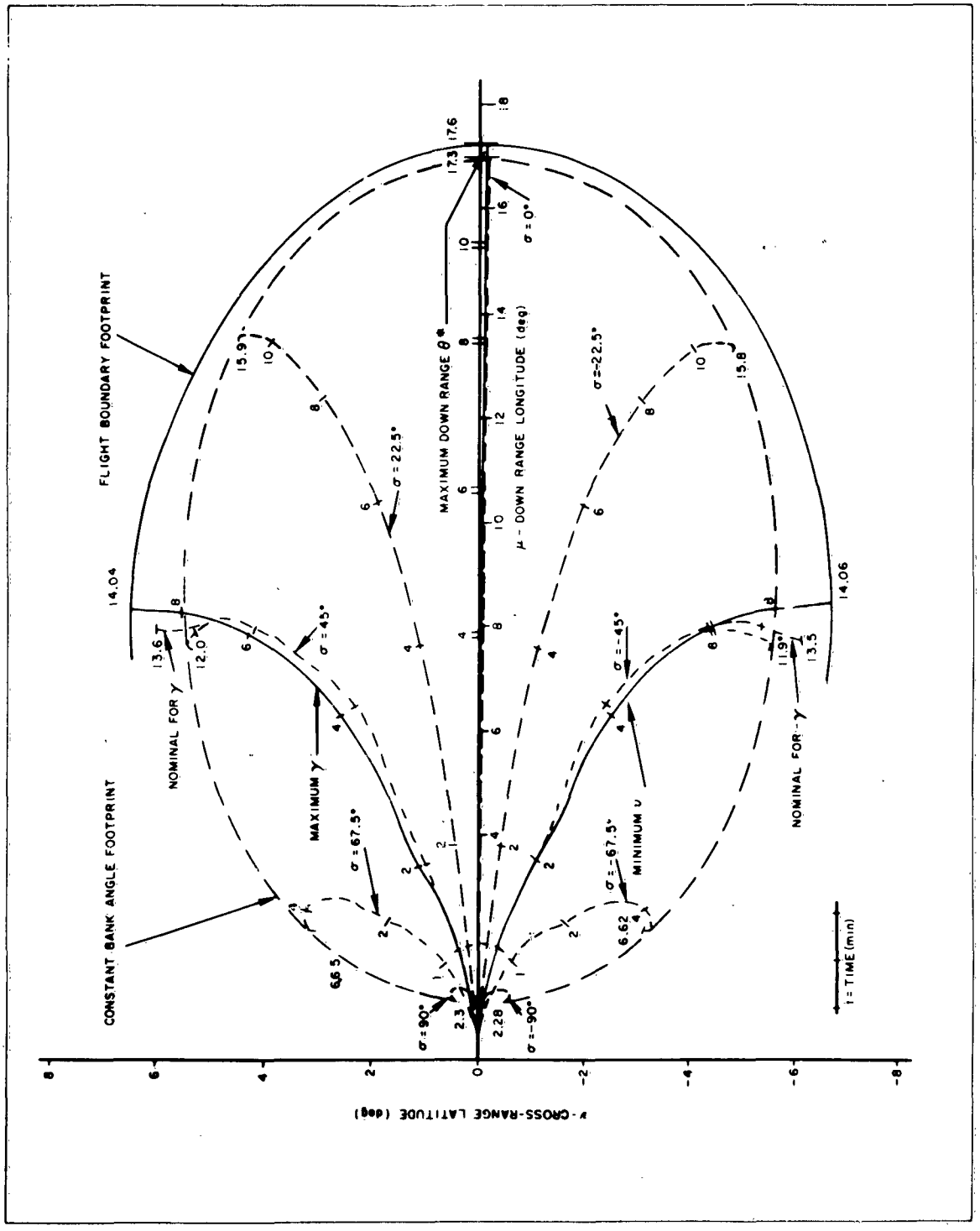


Figure 15 - Auxiliary Mercator Projection of Atlantic Missile Range Flight Boundary

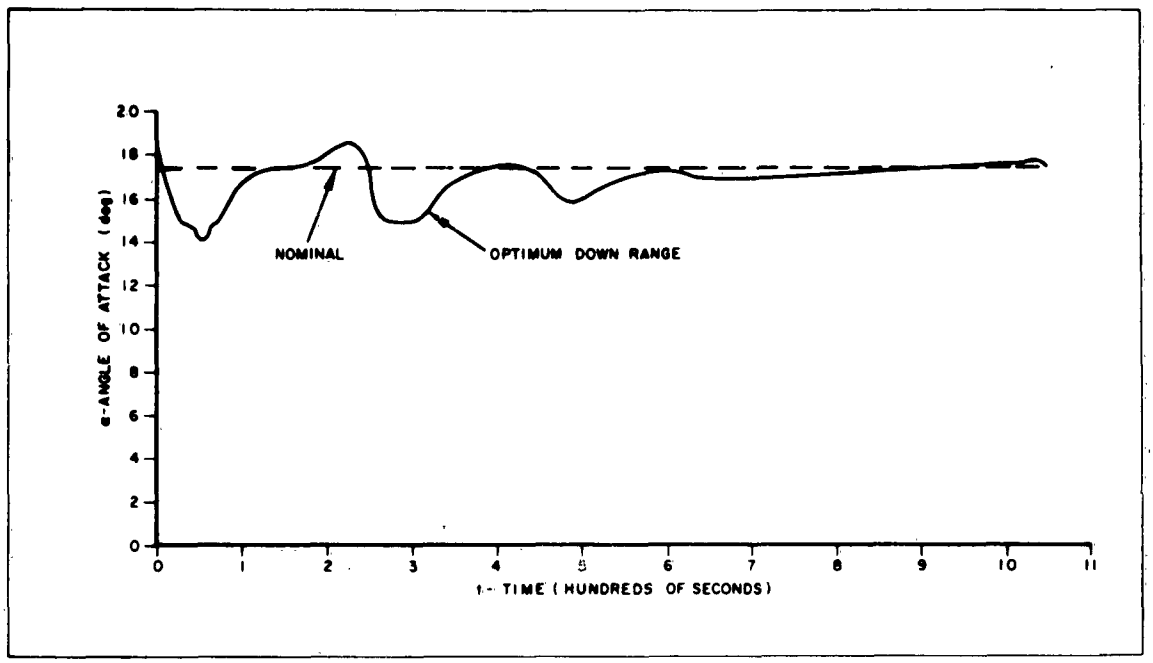


Figure 16 - α vs t for Maximum Downrange in Recovery Area

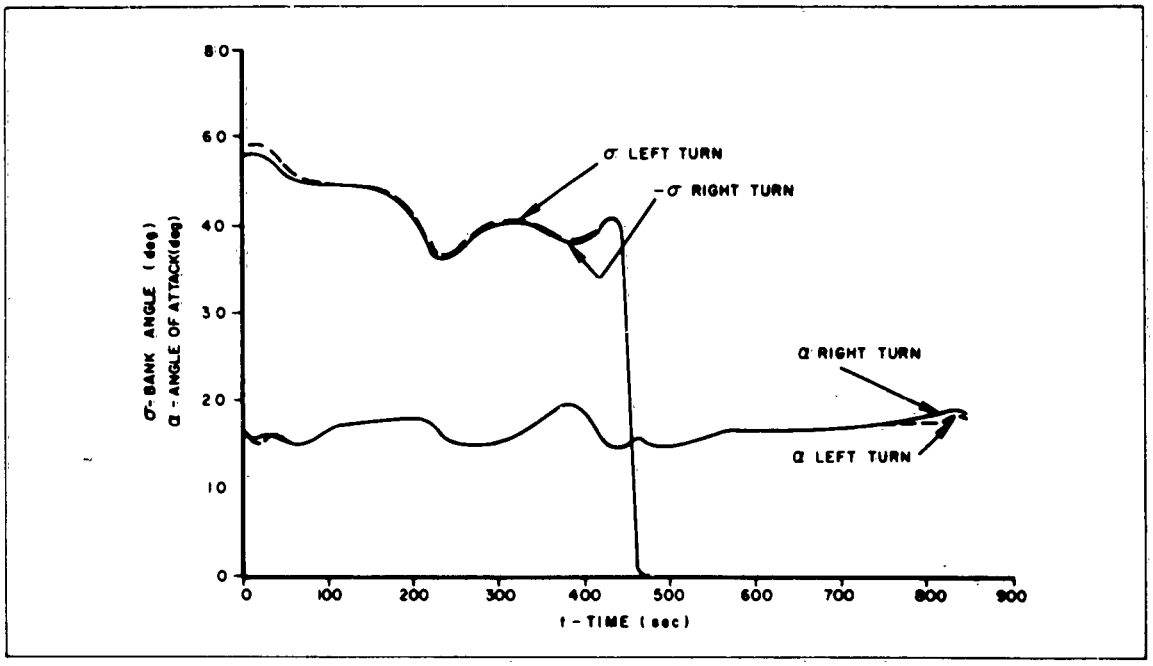


Figure 17 - σ and α vs t for Maximum and Minimum Latitude in Recovery Area

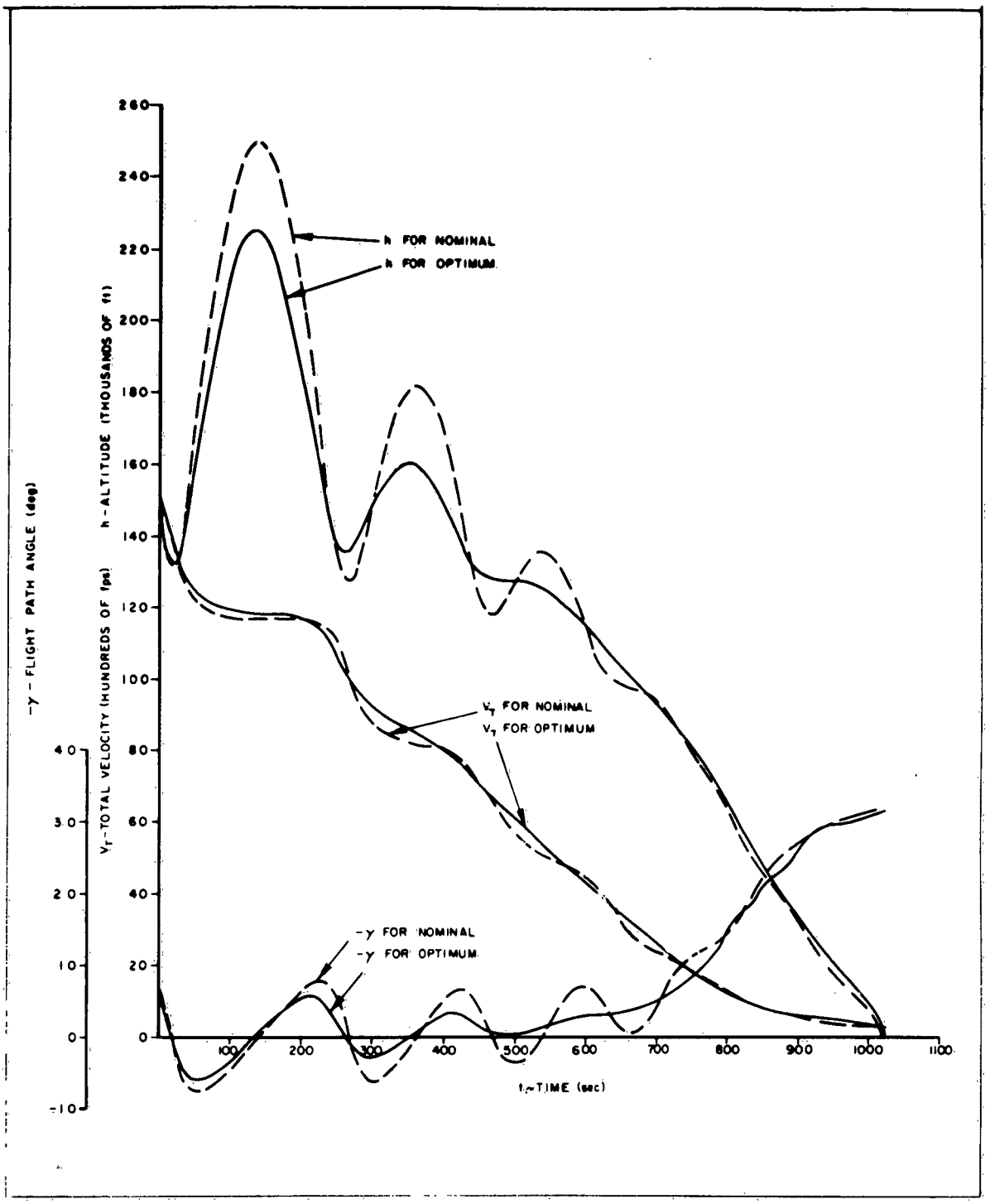


Figure 18 - h, V_T and γ vs t for Maximum Downrange in Recovery Area

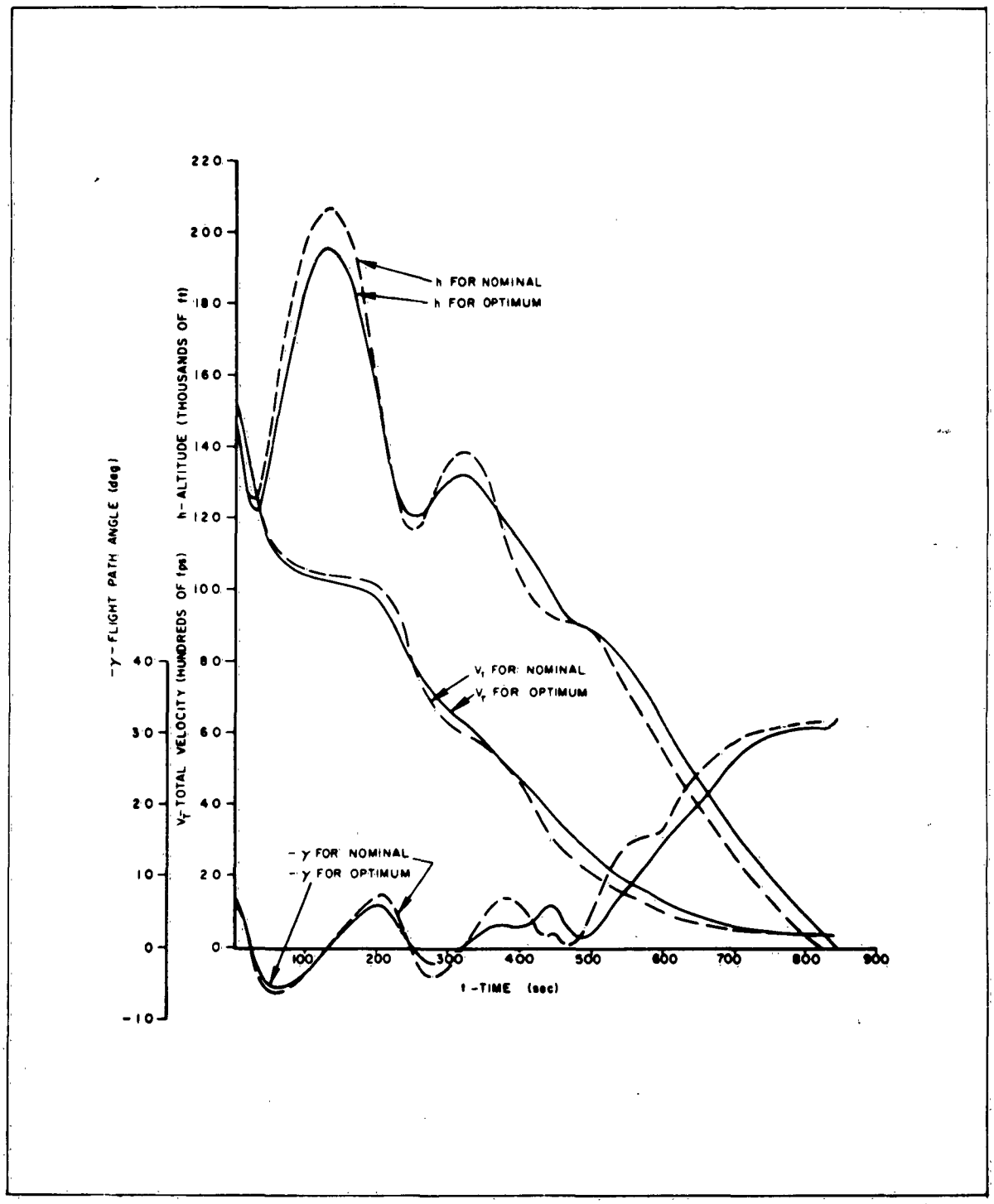


Figure 19 - h, V_T and γ vs t for Maximum Lateral Latitude in Recovery Area

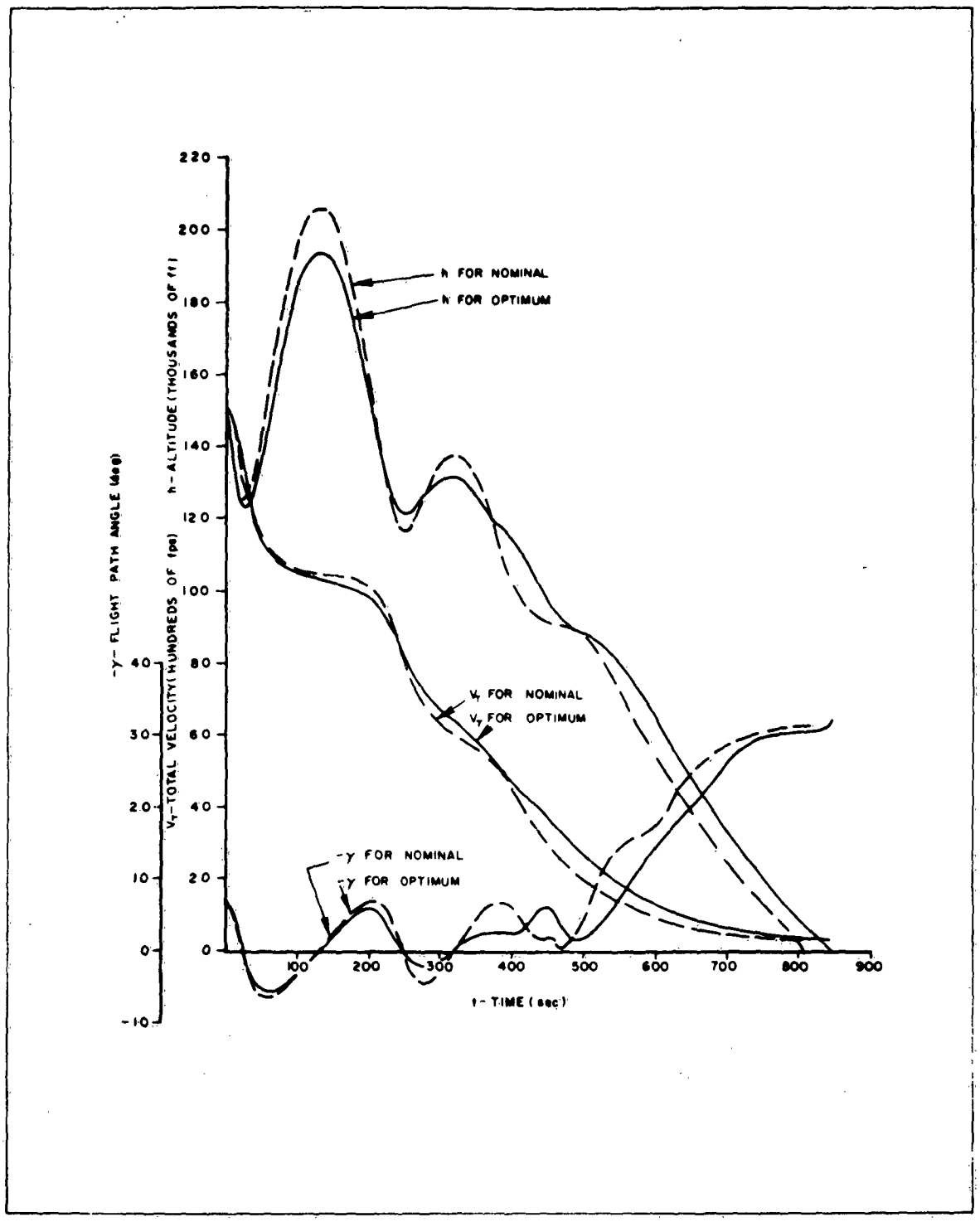


Figure 20 - h. V_T and γ vs t for Minimum Lateral Latitude in Recovery Area

The flight boundary near the initial point was not investigated since structural and other vehicle limitation data, which would determine the trajectory, were not available at the time of this investigation. The maximum downrange, θ^*, was determined from the nominal trajectory for the angle of attack for maximum lift-to-drag ratio of 17.5 degrees and a bank angle equal to zero. The nominal trajectory had a range of 1008 nmi (θ^* = 16.8 degrees) and the maximum downrange was 1026 nmi (θ^* = 17.2 degrees), an increase in range of about 18 nmi or about a 2% increase. Classical use of $(L/D)_{max}$ angle of attack gives a good approximation for the nominal trajectory to be used to determine maximum range capabilities. Maximum and minimum lateral latitude, ν and $-\nu$ respectively, were equal to

$$\nu_{max} = 6.59°$$

$$\nu_{min} = -6.60°$$

which were approximately 8.2% increase over the nominal values. Table 3 shows the summarized results.

TABLE 3
MAXIMUM AND MINIMUM LATERAL LATITUDES IN RECOVERY AREA

	$\nu_{nominal}$ - deg	ν_{max} - deg	$\Delta\nu$ - deg	$\frac{\Delta\nu}{\nu_{nominal}}$
Left Turn	6.10	6.59	.49	.0816
Right Turn	-6.10	-6.60	-.50	.0820

10. RESULTS

A demonstration problem has determined certain of the performance boundaries for a typical three-stage booster system and the SORTIE vehicle during reentry and recovery. Three phases of the computer program developed to solve three-dimensional trajectory optimization problems were demonstrated. These phases included the determination of trajectories for 1) vehicles with thrust, 2) re-entry glide vehicles with an in-flight constraint and 3) lateral maneuvers for a gliding vehicle. Starting from a nominal trajectory, each series of trajectories was computed using the steepest-ascent computational procedure. This computational procedure has shown that the method can determine in an efficient and in an effective manner the optimum trajectory satisfying not only the specified terminal conditions, but in-flight constraints as well. The efficiency of the computer program depends upon the ingenuity of the engineer and the nominal trajectory selected. However, even with the selection of a nominal which does not satisfy too closely the desired terminal conditions, the computer program is able to determine in a very short time the optimum trajectory.

The computer programs can greatly aid the engineer in solving difficult trajectory optimization problems. Careful analysis of each set of runs by the engineer can also point the way for a better nominal to be used in obtaining the optimum solution in fewer iterations.

These computer programs are especially valuable for the determination of solutions to problems with in-flight and terminal constraints. This study indicates that areas where significant gains can be achieved occur during the flight when the constraints are the overriding considerations of the problem. Thus, for vehicles reentering the earth's atmosphere at near orbital speeds and greater and during high-speed maneuvers when the angles of attack and bank angles may become quite large, the in-flight constraints play a very important role in determining the control-variable programs. Optimum paths in this region of flight lie in a very narrow band, and the control programs must be exact if the specified conditions of the problem are to be met. This

has been shown by both examples of the thrust-direction program for a typical three-stage booster system and the SORTIE re-entry glide with stagnation-point heating rate in-flight constraint.

Control programs determined by the cut-and-try approach do not, in general, satisfy the constraints imposed upon the problem and to satisfy all the constraints simultaneously would be impractical except for the fortuitous exceptional problems. The steepest-ascent technique can determine these control variable programs in an efficient and effective manner; starting with a reasonable nominal contral program each iteration approaches closer to the optimum, eventually satisfying simultaneously the imposed constraints and extremilizing the payoff quantity. This technique has been shown to be especially valuable in solving complex optimum programming problems.

APPENDIX

THRUST AND AERODYNAMIC COMPONENTS TRANSFORMATIONS

Let the x, y, z axis system be aligned in the vehicle as shown in Figure 21, with the two thrust vector offset angles of δ_T and i_T. The components of thrust along the three axes are given as follows:

$$\begin{aligned} T_x &= T \cos i_T \cos \delta_T \\ T_y &= T \cos i_T \sin \delta_T \\ T_z &= T \sin i_T \end{aligned} \tag{1}$$

where the magnitude of the thrust is given by T,

$$T = F(t) + (P_e - P_\infty) A_e \tag{2}$$

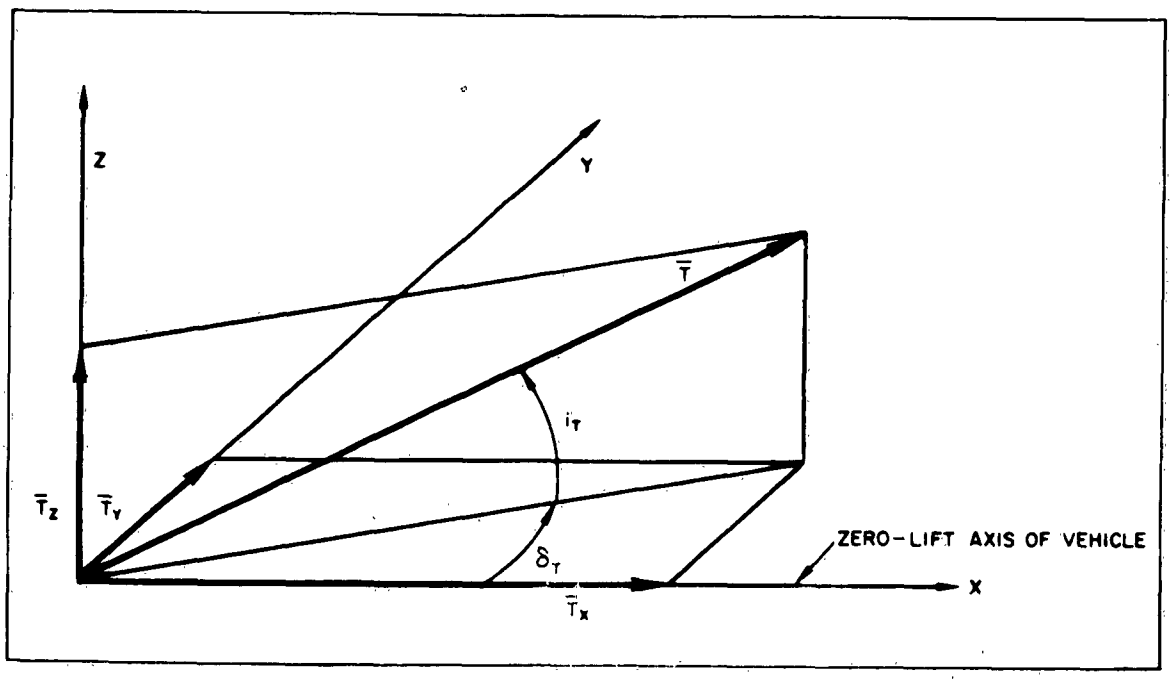


Figure 21 - Thrust Vector Orientation

F(t) is a given function of time and/or stages and is the basic thrust usually given for sea level pressure conditions. The second term of the equation represents the effect of altitude in the basic thrust. Where P_e is ambient pressure at sea level or reference altitude, P_∞ is the ambient pressure at altitude and A_e is the reference exit area associated with the rocket engine for each stage.

The components of the thrust vector given in body axis components must be transformed into components in the geographic coordinate system to be used in the optimization program. The overall transformation can be obtained as a product of four transformations, each one representing a rotation about a single axis.

The first rotation is about the body y axis through the angle-of-attack α:

$$\begin{bmatrix} \hat{L} \\ \hat{V}_T \\ \hat{y} \end{bmatrix} = \begin{bmatrix} \sin\alpha & 0 & \cos\alpha \\ \cos\alpha & 0 & -\sin\alpha \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} \tag{3}$$

where the hatted symbols represent unit vectors. This transformation rotates the body z and x axes into the lift direction, $\hat{L}$, and the velocity direction, $\hat{V}_T$.

The second rotation is about the velocity vector through the bank-angle σ:

$$\begin{bmatrix} \hat{\gamma} \\ \hat{V}_T \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \cos\sigma & 0 & -\sin\sigma \\ 0 & 1 & 0 \\ \sin\sigma & 0 & \cos\sigma \end{bmatrix} \begin{bmatrix} \hat{L} \\ \hat{V}_T \\ \hat{y} \end{bmatrix} \tag{4}$$

This transformation rotates the lift axis and the body y axis into the $\hat{\gamma}$ and $\hat{\beta}$ directions, where γ lies in a vertical* plane defined by the radial direction $\hat{r}$, and the velocity direction $\hat{V}_T$. The $\hat{\beta}$ vector thus lies in a horizontal plane.

The third rotation is about the $\hat{\beta}$ vector through the flight path elevation angle, γ:

$$\begin{bmatrix} \hat{r} \\ \hat{V}_H \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\gamma} \\ \hat{V}_T \\ \hat{\beta} \end{bmatrix} \tag{5}$$

This transformation rotates the $\hat{\gamma}$ and $\hat{V}_T$ vectors into the $\hat{r}$ and $\hat{V}_H$ vectors where $\hat{r}$ is the direction of the vertical* and $\hat{V}_H$ is the direction of the horizontal* component of the velocity.

The fourth and last rotation is about the radial vector $\hat{r}$ through the flight path heading angle, β:

$$\begin{bmatrix} \hat{r} \\ \hat{\theta} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\beta & -\sin\beta \\ 0 & \sin\beta & \cos\beta \end{bmatrix} \begin{bmatrix} \hat{r} \\ \hat{V}_H \\ \hat{\beta} \end{bmatrix} \tag{6}$$

This transformation rotates the $\hat{V}_H$ and $\hat{\beta}$ vectors into the $\hat{\theta}$ and $\hat{\phi}$ vectors of the geographic coordinate system.

The overall transformation matrix is therefore given by the product of the four transformation matrices of equations III-6, 5, 4, and 3:

$$\begin{bmatrix} \hat{r} \\ \hat{\theta} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} \tag{7}$$

*Due to the earth's oblateness $\hat{r}$ is not quite the vertical direction (See, for example, Reference 1).

where the a_{ij}'s are given below. The a_{ij}'s are as follows:

$$
\begin{aligned}
a_{11} &= \frac{u}{V_T}\cos\alpha + \frac{V_H}{V_T}\cos\sigma\sin\alpha \\
a_{12} &= -\frac{V_H}{V_T}\sin\sigma \\
a_{13} &= -\frac{u}{V_T}\sin\alpha + \frac{V_H}{V_T}\cos\sigma\cos\alpha \\
a_{21} &= \frac{v}{V_T}\cos\alpha - \sin\alpha\left(\frac{w}{V_H}\sin\sigma + \frac{v}{V_H}\frac{u}{V_T}\cos\sigma\right) \\
a_{22} &= \frac{v}{V_H}\frac{u}{V_T}\sin\sigma - \frac{w}{V_H}\cos\sigma \\
a_{23} &= -\frac{v}{V_T}\sin\alpha - \cos\alpha\left(\frac{w}{V_H}\sin\sigma + \frac{v}{V_H}\frac{u}{V_T}\cos\sigma\right) \\
a_{31} &= \frac{w}{V_T}\cos\alpha + \sin\alpha\left(\frac{v}{V_H}\sin\sigma - \frac{w}{V_H}\frac{u}{V_T}\cos\sigma\right) \\
a_{32} &= \frac{v}{V_H}\cos\sigma + \frac{w}{V_H}\frac{u}{V_T}\sin\sigma \\
a_{33} &= -\frac{w}{V_T}\sin\alpha + \cos\alpha\left(\frac{v}{V_H}\sin\sigma - \frac{w}{V_H}\frac{u}{V_T}\cos\sigma\right)
\end{aligned}
\tag{8}
$$

where

$$
\begin{aligned}
V_T &= \sqrt{u^2 + v^2 + w^2} \\
V_H &= \sqrt{v^2 + w^2}
\end{aligned}
$$

$$
\begin{aligned}
\cos\beta &= \frac{v}{V_H} \\
\sin\beta &= \frac{w}{V_H} \\
\cos\gamma &= \frac{V_H}{V_T} \\
\sin\gamma &= \frac{u}{V_T}
\end{aligned}
\tag{9}
$$

The aerodynamic force components may be written as

$$\hat{A} = L\hat{L} - D\hat{V}_T.$$

Hence, the components in the geographic coordinate system are given by the product of the three rotations of equations III-6, 5, and 4:

$$\begin{bmatrix} A_r \\ A_\theta \\ A_\phi \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \begin{bmatrix} L \\ -D \end{bmatrix} \tag{10}$$

where

$$b_{11} = \cos\sigma \cos\gamma$$

$$b_{12} = \sin\gamma$$

$$b_{21} = -\sin\gamma \cos\beta \cos\sigma - \sin\beta \sin\sigma$$

$$b_{22} = \cos\gamma \cos\beta$$

$$b_{31} = \cos\beta \sin\sigma - \sin\gamma \sin\beta \cos\sigma$$

$$b_{32} = \cos\gamma \sin\beta$$

REFERENCES

(1) Brown, R. C., Brulle, R. V., and Giffin, G. D., Six-Degree-of-Freedom Flight-Path Study Generalized Computer Program, WADD Technical Report 60-781, Part I, May, 1961.

(2) Kelley, Henry J., Kopp, Richard E., and Mayer, H. Gardner, Successive Approximation Technique for Trajectory Optimization, Proceedings of the IAS Symposium on Vehicle System Optimization, Garden City, New York, November 28-29, 1961, pp. 10-25.

(3) Konecci, E. B., Manned Space Cabin Systems. Chapter in "Advances in Space Science," Vol. I (edited by Ordway), Academic Press, 1959.

(4) Hankey, Wilbur L., Jr., Neumann, Richard D., Flynn, Evard H., Design Procedures for Computing Aerodynamic Heating at Hypersonic Speeds, WADC Technical Report 59-610, June 1960.

(5) Kelley, H. J., Gradient Theory of Optimal Flight Paths, Jour. Amer. Rocket Soc., Vol. 30, pp. 947-954, October 1960.

(6) Bryson, A. E., Denham, W. F., Carroll, F. J., and Mikami, K., Lift or Drag Programs that Minimize Re-entry Heating, Jour. Aerospace Sciences, Vol. 29, No. 4, April 1962.

(7) Bryson, A. E., and Denham, W. F., A Steepest-Ascent Method for Solving Optimum Programming Problems, Raytheon Report BR-1303, 1961, and Jour. Appl. Mech., Vol. 29, No. 1, March 1962.

(8) Bryson, A. E., Mikami, K., and Battle, C. T., Optimum Lateral Turns for a Re-entry Glider, Aerospace Engineering, Vol. 21, No. 3, March 1962. Also Raytheon Report BX-110.

(9) Koelle, H. H., Editor, Handbook of Astronautical Engineering, McGraw-Hill Book Company, Inc., 1961.

(10) Mikami, K., Battle, C. T., Goodell, R. S., and Bryson, A. E., Three-Dimensional Trajectory Optimization Study, Part I - Optimum Programming Formulation, ASD-TDR-62-295, Part I, May, 1962.

(11) Mikami, K., Battle, C. T., Goodell, R. S., and Bryson, A. E., Consultant, Three-Dimensional Trajectory Optimization Study, Part II, Computer Program and User's Manual, ASD-TDR-62-295, Part II, May, 1962.

(12) Mikami, K., Battle, C. T., Goodell, R. S., and Bryson, A. E., Consultant, (U) Three-Dimensional Trajectory Optimization Study, Part III, Demonstration Problem Report, ASD-TDR-62-295, Part III, May, 1962. CONFIDENTIAL REPORT

The Doliac Macro-Micro Control Logic, Its Synthesis, Evaluation, Potential and Problems

By

Daniel O. Dommasch

DODCO, Inc.
Blawenburg, New Jersey

THE DOLIAC MACRO-MICRO CONTROL LOGIC, ITS SYNTHESIS, EVALUATION, POTENTIAL AND PROBLEMS

Daniel O. Dommasch, DODCO, INC.

Abstract

The acronym, DOLIAC, is formed from the first letters of the title, DODCO OPTIMUM LIMITED INFORMATION ADAPTIVE CONTROLLER. The term macro-micro signifies that both the over-all guidance function (macro control) and the autopilot function (micro control) are accomplished within the same logical optimized system.

The DOLIAC logic concept of autopilot system optimization was developed by DODCO, INC., in 1958, and subsequently has been examined in detail by the Cornell Aero Lab. and other investigators. References pertaining to the logic are provided in the paper. The macro-micro additions to the logic were accomplished early in 1960 and have since been subject to continued research and development effort. The basic DOLIAC logic is obtained using either standard differential or variational calculus procedures and both the necessary and sufficient conditions for solution existence are satisfied by the analysis (which is based on the use of an inequality boundary constraint on system action). The research leading to the development of the DOLIAC logic has been conducted under Air Force contracts directed by Mr. P. C. Gregory of the Flight Controls Laboratories, ASD.

It is demonstrated that the Lagrange multiplier acting on an inequality restraint performs the same function as the slack variable in linear programming analyses and that the basic DOLIAC logic defines a simple non-linear gain switching system requiring error and error rate data for its functioning. Since system performance depends on the direct sensing of error rates, adequate sensing of these rates must be possible or system performance is sharply degraded. Thus, there is no point in employing the DOLIAC logic to try to provide a control precision which is beyond the precision of possible error measurements.

Although simple heuristically based system stability criteria for the DOLIAC logic are presented in the paper, these criteria are known to be too restrictive, and much work remains to be done in the area of analytic stability analysis. Although the non-linearities of the DOLIAC logic are of an apparently simple sort, their nature precludes the definition of an equivalent linear system and, therefore, precludes the direct use of existing linear stability criteria for analysis. Whereas the analytic treatment of DOLIAC stability is incomplete, extensive numerical studies on both digital and analog computing equipment reveal, without question, the superiority of the non-linear optimized logic over conventional linear system performance in many cases. In other cases,

the DOLIAC theory may be used to demonstrate that linear error and error rate systems themselves form a degenerate class of optimum systems which, in these cases, may represent the only practical optimums obtainable.

It seems apparent that the mere existence of the DOLIAC logic provides an affirmative answer to the question "Is optimum synthesis practical"; however, it is shown that the DOLIAC theory is, in and of itself, still rather crude, representing only a single step forward from conventional linear theory and there exists the potential for improvements in the future.

LIST OF SYMBOLS

C	Restraint limit
e	error quantity (desired - actual value)
e_n	error at point n, time t
e_{n+1}	error at point n + 1, time t + Δt
e_{n+2}	error at point n + 2, time t + 2 Δ t
$(e_{n+1})_c$	controlled error at point n + 1
$(e_{n+1})_{nc}$	uncontrolled error at point n + 1
$\dot{e}$	error rate
$\ddot{e}$	error acceleration
$\Delta\ddot{e}_c$	commanded error acceleration
f	function to be made stationary
f_L	stationary value of f along root locus curves in plane of f and commanded error acceleration
$\bar{f}$	Level line value of f
g	Equation of ray in the f , $\Delta\ddot{e}_c$ plane joining the maximum and minimum solutions of f_L

LIST OF SYMBOLS (Cont' d)

k	Inverse gain factor
K	Forward loop gain factor
m	derivative level of Taylor expansion, Lagrange multiplier ratio
n	as a subscript, point in time
r	response factor
t	time
Δt_p	prediction or lead time
Δt_s	sample time of sample data system
θ	pitch attitude
λ	Lagrange multiplier
ϕ	Restraint function

THE DOLIAC MACRO-MICRO CONTROL LOGIC, ITS SYNTHESIS, EVALUATION, POTENTIAL AND PROBLEMS

1.) INTRODUCTION

The word "optimum," in common with such words as "logical" and "objective," is normally subjective in its connotations to any given individual. Thus, most people consider themselves as logical and objective which is, of course, paradoxical, since any consideration of oneself is, by definition, subjective. The fact that a similar paradoxical situation exists concerning the word "optimum" requires us at the outset, if we want to deal with specifics rather than generalities, to limit our concept of optimum in such a way that we can obtain a mathematical statement capable of defining what we mean by "optimum." We must, furthermore, limit ourselves to consideration of practical (i.e., physically obtainable) optima which are respectful of the governing laws of physics, otherwise we are likely to obtain solutions to problems which exist abstractly but which have no real physical counterparts. These restrictions on acceptable methods of synthesis are not peculiar to any field of analysis, nor are they in any sense new, novel or original. Indeed, the need for restricting optimum solutions to obey the laws of nature was recognized long ago by both Euler and Lagrange (the historical fathers of mathematical optimizing procedures) and, as a matter of fact, the most widely used method of introducing restraints of a physical nature into maximum-minimum problems carries the name of Lagrange.

In the event that we are able to completely define the nature of a restraint either in the form of a differential equation (e.g., when applying the laws of motion) or in some other specific way (e.g., isoperimetric restraints), the approach to problem solution using the method of undetermined Lagrange multipliers is rather formalized although it is necessary to properly interpret the nature of the restraints. On the other hand, if the restraints are not so well defined, as in the case of inequality conditions, then there may be room for argument as to the applicability of the particular method of employing a given restraint and as to the meaning of the results obtained from a given application procedure. For the class of problems we are going to discuss, the restraints are of the inequality type, and the Lagrange multiplier takes on some of the characteristics of the "slack" variable encountered in the theory of linear programming. Thus, the physical significance of the multiplier is somewhat obscure, and this is one of the reasons why the new logic represents only a basic first-step in improvement over simple linear error and error rate feedback control theory. It does, however, represent an improvement and it does work. Although the basic DOLIAC logic is nonlinear in nature, in a degenerate form it includes the entire set of linear error and error rate feedback systems as a subset which is extremely interesting since this offers some proof of the thought that linear feedback may constitute an optimum solution to specific (if restricted) classes of problems.

With these introductory thoughts in mind, we now seek to define a practical optimum criterion, and to do this we need to ask just what actually can be accomplished by a real control system.

In general sense, any control logic system seeks to control some variable so that the error, defined as the difference between a desired value of the variable and the existing value, vanishes at all points and for all time. In a dynamic problem, this is an impossible assignment--unless we are willing to hedge a bit on the meaning of the term "vanish," so that practically speaking we can only require that the error be minimized on an absolute or magnitude basis. We, therefore, have the practical requirement that we should seek the minimum possible numerical value of error regardless of sign, and this is a requirement which is meaningful mathematically. We need to acknowledge the fact that all systems have inertial of one form or another so that, in the absence of infinite command authority, no instantaneous changes in system displacement state are possible. This means that there is nothing we can do to instantaneously change the displacement error pattern, but that we must wait a finite time for control action to have its effect on an existing error state. This is a physical limitation which cannot be circumvented, but rather must be accounted for, by modifying our criterion to state that all we can hope to accomplish is a minimization of absolute error sometime in the future, with this future time depending on the command authority at our disposal. In the absence of a change in command, and considering only a small time increment, Δt_p, we may relate the state of error in the future to the present error state via a Taylor expansion about a present time point "n," i.e., we may write that

$$(e_{n+1})_{nc} = e_n + \Delta t_p \dot{e}_n + \frac{\overline{\Delta t_p}^2}{2} \ddot{e}_n + \cdots + \frac{(\Delta t_p)^m \overset{m}{e}_n}{m!} \qquad (1:1)$$

In the presence of external disturbances which are random in nature, the higher derivatives are uncertain and, moreover, although we cannot instantaneously alter e_n, we can produce an <u>almost</u> instantaneous change in the error acceleration $\ddot{e}_n$. These factors together prevent the use of the full Taylor expansion as an accurate representation of the uncontrolled error history and, therefore, the series must be truncated at the $\ddot{e}_n$ level (as the highest order term) and Δt_p restricted in magnitude accordingly.

If we now postulate that a command is generated on the acceleration level, then, if the command is $k\Delta\ddot{e}_c$ where k is an inverse gain factor, we have, as a matter of definition,

$$(e_{n+1})_c = (e_{n+1})_{nc} + \frac{\overline{\Delta t_p}^2}{2} k\ddot{\Delta e}_c \qquad (1:2)$$

If there were no restraints imposed on the operation of the system, we could require that $(e_{n+1})_c$, the error at time $t + \Delta t_p$ in the presence of a command, be zero. In this event, the error command would be

$$\ddot{\Delta e}_c = 2(e_{n+1})_{nc} / k\overline{\Delta t_p}^2 \qquad (1:3)$$

in which it is apparent that $2/k\overline{\Delta t_p}^2$ is equivalent to the conventional forward loop gain factor which we will designate simply as K. Thus,

$$K = 2/k\overline{\Delta t_p}^2 \qquad (1:4)$$

and

$$\ddot{\Delta e}_c = K(e_{n+1})_{nc} \qquad (1:5)$$

If we now limit ourselves to simple error and error rate systems, then, from (1:1)

$$(e_{n+1})_{nc} = e_n + \Delta t_p \dot{e}_n \qquad (1:6)$$

so that

$$\ddot{\Delta e}_c = K(e_n + \Delta t_p \dot{e}_n) \qquad (1:7)$$

This is the conventional equation of a fixed gain error-and error rate feed back controller, and it is important to note that the equation is the result of our making the idealized assumption that control authority exists to properly implement a desired command. This assumption is seldom correct in practice, yet equation 7, when K is properly selected, frequently yields satisfactory control. We will return to this point later on in the paper.

In conventional linear theory, the term Δt_p is referred to as a "damping" factor. Such usage follows by analogy with the equation of motion of a linear system wherein the coefficient of the rate term represents the systems equivalent fluid damping constant. This usage, although conventional, is inexact and possibly misleading when we use equation 7 to control a system having non-linear dynamics. As a matter of fact, Δt_p is properly referred to as a "prediction time," although the term "lead time" is also applicable.

Since, historically, direct application of equation 7 is known to provide adequate results, one must suppose that the crudeness of its development is justified by the fact that it works, but by the same token it must also be supposed that improvements can be made without too much trouble, and that these improvements need not be sophisticated to provide substantial gains in performance.

The first improvement, which is more or less obvious ought to be made, is to provide information in some form or other that the command authority is limited. This statement is an inequality restraint on the problem, which introduces a Lagrange multiplier, but being that the restraint is indefinite, the multiplier cannot be defined directly by the optimum analysis for if it were, then the restraint would be isoperimetric and control authority would always be set at the limiting value. In this event, our problem would be determined beforehand and optimization would be impossible. This use of the Lagrange multiplier as a method of taking up the "slack" in the inequality restraint is analogous to the introduction of a "slack" variable in linear programming analysis to account for inequality type restrictions on the objective function.

As we shall show, the Lagrange multiplier, introduced as an indicator of an inequality adjoint condition, has an intimate relationship with the system gain and, therefore, is subject to all problems characteristic of the gain selection process. There is, however, an important difference which should be kept in mind. The introduction of a simple restraint renders the logic nonlinear and, therefore, no direct use can be made of linear theory for the establishment of the proper value of the multiplier. Thus, we emphasize at this point that despite the disarmingly simple form of the equations describing the DOLIAC logic, that this logic is intrinsically nonlinear and cannot be analyzed in general using root-locus procedures, the first method of Liapounoff, describing function techniques or similar procedures which sometimes are useful in the study of "mildly" nonlinear systems. On the other hand, there are some indications that conservative stability boundaries can rather readily be established for the DOLIAC type logic on a comparative basis. We shall say more about this point later on in the paper.

To complete our introduction, we make note of the fact that since our criterion for optimum performance is to minimize numerical error at some short time in the future, it is not possible to directly operate on the error and set the derivative $d(e_{n+1})_c / d\Delta e_c$ to 0, since this procedure as a mathematical definition is capable of finding only maximum (large plus values) or minimum (large negative values) of $(e_{n+1})_c$ rather than minimum numerical values.

This difficulty can be avoided by seeking the stationary values of the quantity $(e_{n+1})_c^2$; however, this modifies the basic criterion in a fashion which emphasizes large errors while de-emphasizing small errors and, accordingly, although the error squared criterion produces satisfactory control in some instances, it is inherently a poor regulatory criterion. Evidently, the artifice of modifying the criterion to suit the mathematics is hardly desirable. We could directly seek the maximums and minimums of absolute error; however, this involves some definition difficulties which can be avoided by seeking the stationary values of $1/e$. The maximums and minimums of this inverse function represent the closest approach of the error to a zero value regardless of sign of the error and permit the direct utilization of both the maximum and minimum solutions which evidently must exist for an arbitrary function. It is of interest to note that the use of the error squared criterion effectively eliminates one class of solutions immediately and leads only to the conventional error and error rate linear feedback hypothesis. On the other hand, the use of an inverse error function immediately introduces switching logic into the problem and, as well, a requirement for the selection of the best of two answers available in any given case. We thus obtain, without ambiguity, a direct and proper solution to the postulated optimum problem.

In the following section of this paper we shall develop the method of introducing inequality restraints into the reciprocal error stationary point analysis and this will lead directly to the basic DOLIAC logic concepts.

2.) SYNTHESIS AND ANALYSIS OF THE DOLIAC LOGIC

We have indicated that the DOLIAC logic is obtained directly by seeking the stationary points of an inverse error function subjected to an inequality restraint which limits the magnitude of possible control action. To provide a physical basis for the mathematical development, we need to discuss some basic concepts applicable to the problem.

The Lagrange multiplier technique of ordinary differential calculus and the technique of linear programming (which finds wide application in operations research for problems where linear restraints with "corners" are involved) are both based on the same geometric considerations. This geometry is easy to visualize in a three-dimensional problem space as illustrated by Figures 2:1 and 2:2.

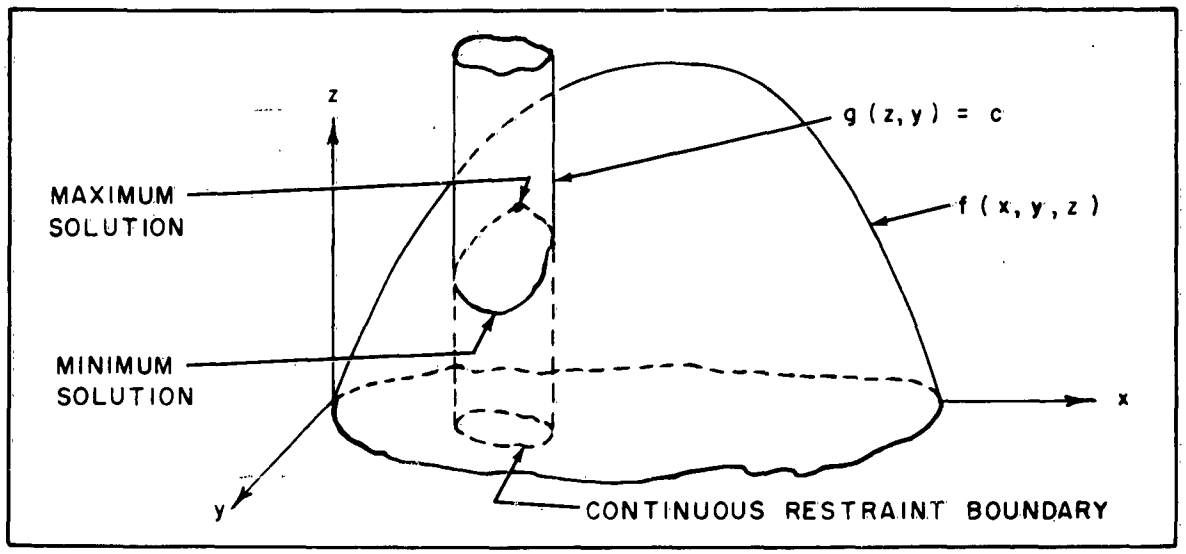


Figure 2:1. Lagrange Multiplier Technique

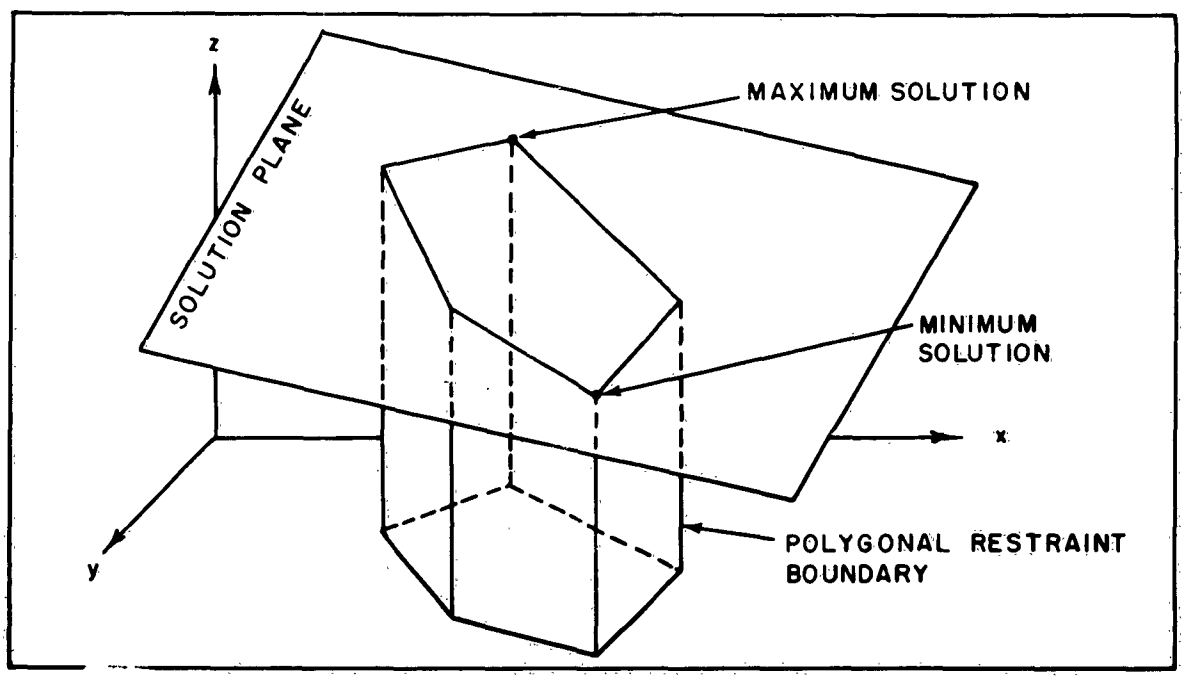


Figure 2:2. Linear Programming Technique

Figure 2:1 illustrates a standard Lagrange multiplier problem involving the determination of the maximum and minimum values of a function subject to the restraint that these stationary points shall lie on a cylindrical (not necessarily a circular one) surface $g(x,y) = C$. The geometric solution is apparent from the figure. Figure 2:2 presents a typical linear programming problem showing a plane of solution and a polygonal prismatic restraint which limits the problem. The maximum and minimum solutions are those corresponding to the maximum and minimum values of z on the solution plane which lie on the boundary of the polygonal cylinder.

Although a three-dimensional solution space has been considered for illustration purposes, the same geometric concepts apply to an "n" dimensional solution space, even though one cannot use a physical model to illustrate this general "n" dimensional case.

In the theory of linear programming, iterative techniques exist for handling inequality types of restraint conditions in a precise fashion. The commonly used procedure is to introduce so-called "slack" variables which transform the inequality restraints into apparent equalities; in other words, the slack variables take up the slack in the inequalities. The slack variables are eliminated from the solution during iteration on the problem matrix.

No apparently comparable technique exists for handling inequality restraints for problems involving continuous, but nonlinear functions. However, the Lagrange multiplier itself may be considered a slack variable which is determined by conditions external to the basic problem stipulation.

The development of the theory of optimum control synthesis methods involves the use of a simple inequality restraint, which is introduced via an indefinite Lagrange multiplier. The multiplier itself is not established by the analysis because the restraint is an inequality, and, indeed, were the restraint of any other type optimization would not be possible. Thus, the multiplier takes up the slack or uncertainty of the formulation and performs the function of a "trade off" parameter in the sense used in game theory. Although the imposition of the inequality restraint adds sufficient information to provide an order of magnitude of improvement in control system performance, further improvements should still be obtainable if the physical action of the multiplier can be established more precisely.

The particular problem we are immediately concerned with is summarized by the statements:

if, e_n = error at a point n

$\dot{e}_n$ = error rate at point n

$\ddot{e}_n$ = error acceleration at point n

then at some point Δt_p in the future, (n + 1) we predict that the error will be

$$(e_{n+1})_{nc} = e_n + \dot{e}_n \Delta t_p + \ddot{e}_n \frac{\Delta t_p}{2} + \ldots + \qquad (2:1)$$

provided no command is imposed.

If a command is generated at the acceleration level, then if this command is $k\Delta\ddot{e}_c$ where k is an inverse gain factor, the error we command at the point n + 1 is

$$(e_{n+1})_c = (e_{n+1})_{nc} + k\Delta\ddot{e}_c \frac{\overline{\Delta t_p}^2}{2} \qquad (2:2)$$

We seek to find that command value which makes $(e_{n+1})_c$ as numerically (not algebraically) small as possible and, therefore, seek the maximums and minimums of $1/(e_{n+1})_c$, while recognizing that $(e_{n+1})_c$ can be zero in an unrestrained case, and that this solution will be found unless the physical restraints imposed by the systems dynamics are recognized. We give recognition to these restraints by the inequality condition that a command $k\Delta\ddot{e}_c$ must be numerically less (regardless of sign) than some boundary value, C, i.e., $k\Delta e_c \leq C$, (where the sign of C depends on whether Δe_c is plus or minus). In reciprocal form, we write that $1/k\Delta e_c \geq C$, where the value of C has been changed accordingly. We thus define a function ϕ, such that

$$\phi = \frac{1}{k\Delta\ddot{e}_c} - C \geq 0 \qquad (2:3)$$

In accord with the classical Lagrange technique, we observe restraint (2:3) by forming the function

$$f = \frac{1}{(e_{n+1})_c} + \lambda\phi = \frac{1}{(e_{n+1})_{nc}} + \lambda\left(\frac{1}{k\Delta\ddot{e}_c} - C\right) \qquad (2:4)$$

and seek the stationary values of f, defined by $\partial f / \partial \Delta\ddot{e}_c = 0$. In this formulation the constant λC vanishes on differentiation so that its sign is of no importance; however, the actual sign of C is determined by the following considerations: If $\Delta\ddot{e}_c$ is positive, inequality 2:3 is satisfied as long as $\Delta\ddot{e}_c$ is sufficiently small. However, as $\Delta\ddot{e}_c$ grows, we ultimately find that C dominates, and for C being a positive real number, we reach a command limit on $\Delta\ddot{e}_c$. If $\Delta\ddot{e}_c$ is negative, C must be negative and the inequality reversed; thus, for $\Delta\ddot{e}_c$ negative

$$\phi = \frac{1}{k\Delta\ddot{e}_c} + C \leq 0 \qquad (2:5)$$

and, in general

$$\phi = \frac{1}{k\Delta\ddot{e}_c} \pm C \geq 0, \leq 0 \qquad (2:6)$$

We again recall that the sign of C is of no importance in the classical approach, since λC vanishes on differentiation.

In general, equation 2:4 may be written as

$$f = \frac{1}{(e_{n+1})_{nc} + \frac{k\Delta t_p^2}{2}\Delta\ddot{e}_c} + \frac{\lambda}{k\Delta\ddot{e}_c} \pm \lambda C \qquad (2:7)$$

Accordingly, f is a function of the single independent variable $\Delta\ddot{e}_c$, and only a single minimum condition can be imposed so that λ remains undetermined by the optimizing process, i.e., it depends on conditions external to equation 2:7.

At any given point λC being a constant, serves merely to move the entire function curve upwards or downwards along the f axis in a plane defined by the orthogonal axes f and $\Delta\ddot{e}_c$, and this is the reason why the term λC contributes nothing to a maximum-minimum solution.

If we were to plot f as a function of $(e_{n+1})_{nc}$, we would find that there exist two solutions for any given combination of $(e_{n+1})_{nc}$, λ and k with these solutions being located on either side of a discontinuity of the function $1/(e_{n+1})_c$. This discontinuity represents the unrestrained solution at which $(e_{n+1})_c$ is zero. Such a solution is normally far from desirable in a real system which possesses inertia since to obtain it for reasonably high gains, one must force an overshoot at a point n + 2, approximately as large as the error at the point n.

The locations of the roots of the optimizing equation, by themselves, do not fix the command, since the command must be selected from the two available possibilities. The selection criterion is based on the fact that control is applied steadily so that we seek the minimum error at all times. Thus, if e_n is the error at n and $(e_{n+1})_c$ is the predicted error, we select the value of $\Delta\ddot{e}_c$ leading to the minimum numerical (not algebraic) value of $e_n + (e_{n+1})_c$.

The root loci of the two possible solutions in the f, $\Delta\ddot{e}_c$ space and their movements relative to the unrestrained solution are of interest and can be defined by the following analysis.

Let

$$\lambda_{min.} = -2/\Delta t_p^2 \qquad (2:8)$$

and define a ratio m by the statement

$$m = \lambda/\lambda_{min.} \qquad (2:9)$$

so that m varies directly with λ. Also, let

$$K = 2/k\Delta t_p^2 \qquad (2:10)$$

so that K varies inversely with k, then the condition, $\partial f/\partial\Delta\ddot{e}_c = 0$, yields the optimum command equation

$$\Delta\ddot{e}_c = \frac{K(e_{n+1})_{nc}}{\pm\sqrt{1/m} - 1} \qquad (2:11)$$

which has two roots.

For various values of $(e_{n+1})_{nc}$ and m and K equation 2:11 specifies the two root locations along the $\Delta\ddot{e}_c$ axis. The vertical displacement of the function f at these points is then defined by the fact that,

$$f = \frac{1}{(e_{n+1})_{nc} + k\Delta\ddot{e}_c\overline{\Delta t_p}^2/2} + \frac{\lambda}{k\Delta\ddot{e}_c}$$

or in terms of m and k

$$f = \frac{1}{(e_{n+1})_{nc} + \Delta\ddot{e}_c/K} - \frac{mK}{\Delta\ddot{e}_c} \qquad (2:12)$$

and on substituting for $\Delta\ddot{e}_c$, the function along the root locus curves f_L is defined by

$$f_L = \frac{1}{(e_{n+1})_{nc}\left\{1 + \frac{1}{\pm\sqrt{1/m} - 1}\right\}} - \frac{mK}{K(e_{n+1})_{nc}\left(\pm\sqrt{1/m} - 1\right)} \qquad (2:13)$$

which simplifies to

$$f_L = \frac{1 + m \mp 2\sqrt{m}}{(e_{n+1})_{nc}}$$

For the cases, $(e_{n+1})_{nc} = \pm 1$, K = 1, plots of

$$\frac{1}{(e_{n+1})_c} = \frac{1}{(e_{n+1})_{nc} + \Delta\ddot{e}_c/K}$$

and

$$f_L = \frac{1 + m \mp 2\sqrt{m}}{(e_{n+1})_{nc}} \qquad (2:14)$$

as functions of $\Delta \dot{e}_c$, appear as in Figure 2:3, with equation 2:14 defining the loci of the roots of the control equations as m varies from 1 to ∞. Note that it can be shown that the values of m are limited to the range 1 to ∞, with the region $m < 1$ being outside of the area where valid solutions can be proven to exist, and also note that the root loci curves are not plotted in the plane commonly used for root locus plots in linear control theory.

The selection of the root to be used in any given case is determined by the criterion that the one yielding the minimum numerical value of $e_n + (e_{n+1})_c$ is proper, and this statement may be shown equivalent to the stipulation: Use the numerically larger value of $\Delta \dot{e}_c$ if

$$0 < \frac{(e_{n+1})_{nc}}{e_n} < (m - 1) \qquad (2:15)$$

otherwise, use the smaller.

It is not known whether the focal nature of the "m" rays revealed in Figure 2:3 is significant, nor if the level line passage through this point is theoretically meaningful; however, all numerical work conducted to date seems to indicate that the m = 9 value and its associated 2 to 1 gain switch provides consistently excellent results.

It is readily demonstrated that the level line of Figure 2:3 is independent of gain and of $(e_{n+1})_{nc}$. This is accomplished as follows. We first recall equation 2:14,

$$f_L = \frac{1 + m \mp 2\sqrt{m}}{(e_{n+1})_{nc}}$$

and find f_L at the limit, m = 1 (which requires use of the minus root), so that we get, in general at m = 1, letting f_L be the function value along the level line,

$$f_L = \frac{4}{(e_{n+1})_{nc}} \qquad (2:16)$$

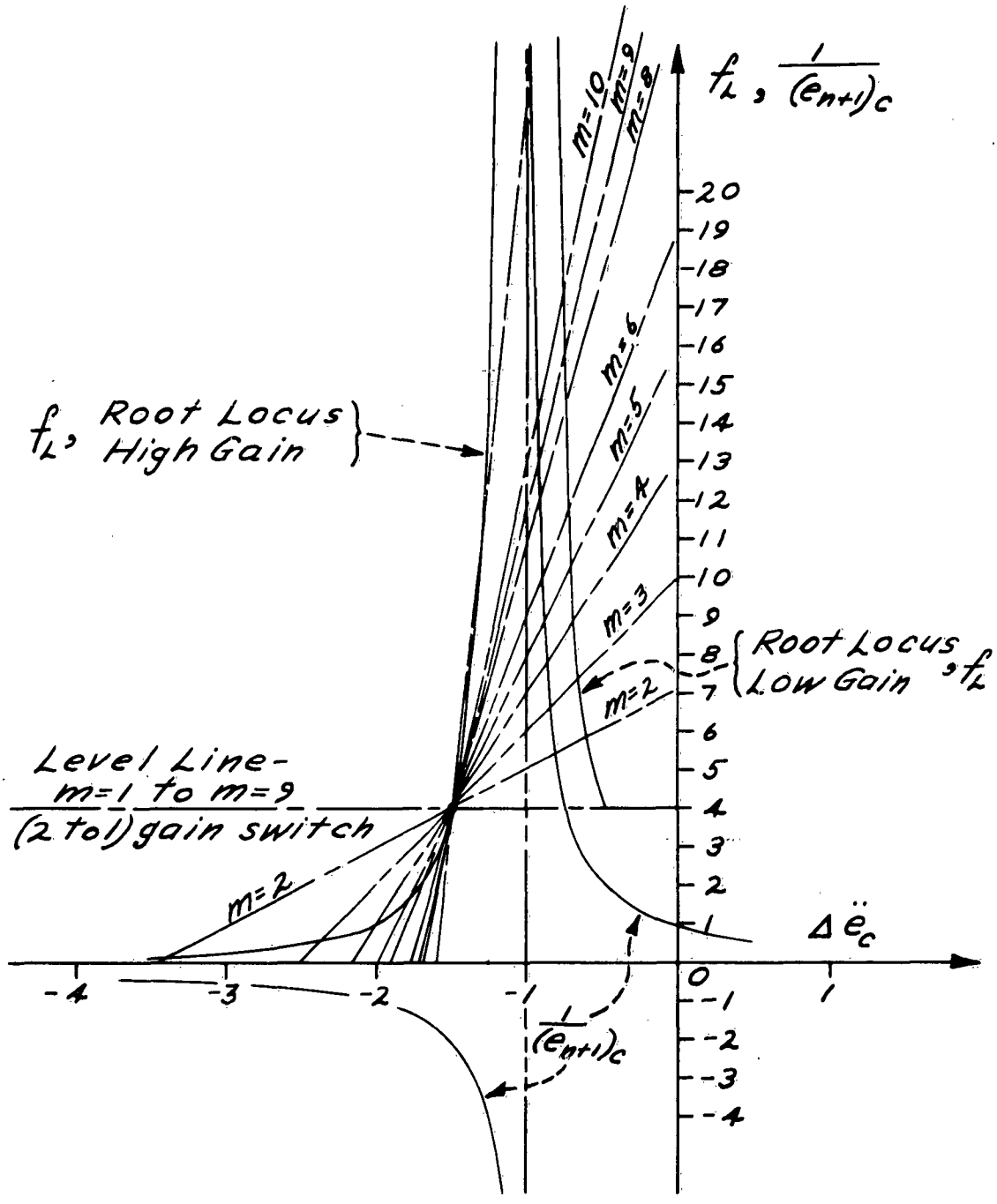


FIGURE 2:3. ROOT LOCI FOR VARIABLE m IN f_L, $\Delta \ddot{e}_c$ PLANE; K = 1

This same value of f_L^* is obtained for m = 9, since for m = 9 (accepting the high gain root), regardless of the value of K,

$$\bar{f}_L = \frac{10 - 2\sqrt{9}}{(e_{n+1})_{nc}} = \frac{4}{(e_{n+1})_{nc}} \qquad (2:17)$$

which completes our proof. For this f_L^* value,

$$\Delta\ddot{e}_c = -(3/2)K(e_{n+1})_{nc}$$

To establish if the bundle of m rays always penetrates the same point, regardless of the K and $(e_{n+1})_{nc}$ values, it is necessary to show that this point is common to all rays.

The general equation for any ray is

$$g = \frac{1 + m - 2\sqrt{m}}{(e_{n+1})_{nc}} + \frac{2(m-1)}{K(e_{n+1})_{nc}^2}\left\{\Delta\ddot{e}_c - \frac{K(e_{n+1})_{nc}}{\sqrt{1/m} - 1}\right\} \qquad (2:18)$$

if all rays pass through the focal point at $\Delta\ddot{e}_c = -(3/2)K(e_{n+1})_{nc}$, then substitution of this value of $\Delta\ddot{e}_c$ should provide that $g = f_L^*$, and this is found to be the case; therefore, in general, all rays pass through the point

$$\bar{f} = \frac{4}{(e_{n+1})_{nc}}, \quad \Delta\ddot{e}_c = -(3/2)K(e_{n+1})_{nc}$$

which is the high gain root for m = 9, which defines the two-to-one gain switching condition. This alternatively proves that the m = 1 ray (or level line) passes through the same point since it is a member of the family defined by 2:18.

Keeping in mind the relative locations of the root loci curves and the unrestrained solution value for $\Delta\ddot{e}_c$ (at the discontinuity of the curve $1/(e_{n+1})_c$), we see that λ serves to keep both the high and low gain roots away from the unrestrained solution with the high gain case representing a greater gain than called for by the unrestrained case and with the low gain case calling for a lesser gain than

represented by the unrestrained case. The separation of the root loci curves from the unrestrained solution is a measure of the undetermined parameter C (i.e., C and λ are directly related) and the sign of C is seen to be established by the root selection process. Thus, C is seen to be the equivalent of the slack variable of linear programming, inversely the multiplier λ serves the same purpose as C, since C and λ are not independent.

If we presume a perfect system such that a given command $\Delta\ddot{e}_c$, could be perfectly fulfilled, the unrestrained solution would represent a boundary of neutral stability, with each command $\Delta\ddot{e}_c$ resulting in overshoot in error at the (n + 2) point equal to the error at (n). In an actual system, this boundary is reached by increasing K of a linear error and error rate system until neutral stability is achieved. Assuming that this K is shown (and it can be found by standard linear methodology), then the damping of an optimum control system using this same value of K should be a function only of λ or its equivalent m. It would appear, therefore, that a stability criterion can be devised for the optimum micro control logic by extension of linear theory, with the recognition that the high gain roots will lie beyond the unrestrained stability boundary in all cases. At this point it appears (intuitively) that the m = 9 root has some special significance; however there is at present no abstract proof to substantiate the intuitive reaction. Numerical data are available which establish only that m = 9 is an excellent average or basic value for the switching logic.

A review of macro-micro control theory is pertinent, since this theory provides variations not considered above.

It can be demonstrated that the introduction of a response factor relating a macro command to a micro (or autopilot) function is equivalent to the inclusion of the term $\ddot{e}_n$ in the relation for $(e_{n+1})_{nc}$, i.e.,

$$(e_{n+1})_{nc} = e_n + \Delta t\dot{e}_n + \frac{\Delta t^2}{2}\ddot{e}_n \qquad (2:19)$$

and that the response factor "r" is given as $r = (\ddot{e}_n/\dot{\theta}_n)\,\Delta t/2$ where $\dot{\theta}$ is the micro control rate and $\ddot{e}$ is the error acceleration of the macro controlled quantity.

When r is included in the analysis, we get that

$$m = \lambda r \Delta t^2/2 \qquad (2:20)$$

$$K = 2/kr\Delta t^2 \qquad (2:21)$$

It is clear that for a given value of λ, r must be limited to those values which produce a value of m greater than one.

In a macro control case, the generation of a command is removed from the macro control quantity level by at least one step and, therefore, $\ddot{e}$ cannot be instantaneously changed. Thus, for fixed λ, there can exist a variable switching ratio. Under direct control (micro control), $\dot{e}_n$ can be changed instantaneously and, therefore, cannot be used to alter gain switching characteristics.

It has been found that in a sample data system, the prediction time must be greater than the sample time (preferably at least 5 times as great), so that the system stability is associated with the ratio of conditions expected at two different times in the future, i.e., at $t + \Delta t_p$ and at Δt_s, where Δt_p = prediction time and Δt_s = sample time. In an analog system the ratio of prediction to sample time is infinite so that such systems display better damping than sample data systems.

Aside from the effects of the ratio $\Delta t_p / \Delta t_s$, it appears that K and m may be independently established and that K should be equal to or greater than the value leading to neutral stability of an average linear system.

This completes our basic though brief analytic examination and development of the DOLIAC logic concept, and we can now list a few pertinent conclusions.

(1) The logic can properly be called "optimum" in a mathematical sense since it was devised using a mathematically proper optimum definition.

(2) Although we have not carried forth the development here, it is shown in the book, "Principles Underling Systems Engineering" Pitman Publishing Corporation, 1962, that the sufficient conditons for the existence of a solution are satisfied by the DOLIAC logic. Thus, both the necessary and sufficient conditions are fulfilled.

(3) The logic is practical, since it does not require a complex network of sensors for its operation, nor more than an elemental computing system for its implementation. Moreover, it requires a minimum amount of input data and, therefore, may be referred to as a "limited information system."

(4) Because the Lagrange multiplier behaves like a slack variable, one must still face the problem of proper gain selection.

(5) The theory is nonlinear because of its optimum switching criterion, and, therefore, stability studies have, up to this time, had to be carried forth using full digital and analog simulation methodology. Studies carried forth at Cornell Aero Laboratories and at DODCO have demonstrated that linear stability analysis techniques do not apply. Indeed, the logic is stable in regions where conventional linear theory indicates it to be unstable.

(6) The theory presented represents only a first step beyond linear theory and, therefore, future improvements are possible; however, the direction which should be followed is not defined at this time.

In the next and final section of this paper, we shall discuss the more pragmatic nature of the DOLIAC logic, and the nature of the results obtained from extensive application studies of the DOLIAC logic in its micro and macro-micro forms.

3.) RESULTS AND CONCLUSIONS

Over the course of the past four years the group at DODCO, as well as other investigators, have extensively analyzed the nature of the DOLIAC logic from various standpoints, and these investigations have shown that the logic is inherently "lazy" in that it does not switch gain unless drift occurs or external disturbance and/or abrupt changes in command are introduce. Thus, for long periods of time the system lies partially dormant operating principally at one gain setting; however, even in such regulatory phases, as drift develops, short bursts of high gain operation occur in accord with the switching logic requirements. Under idealized circumstances, therefore, it is possible to achieve system states which call for almost completely linear operation of the DOLIAC logic, and studies based wholly on such circumstances can produce extremely

misleading results, since under these circumstances one can readily find a describing function which apparently represents the system, but which has, in fact, no meaning outside of the simplified set of operating circumstances assumed for analysis. Similarly, if the problem given to the system is idealized, one can be led to the false conclusion that stability is predictable using the Evans root locus technique, but again the application is quite meaningless (in a general sense) since the switching characteristics depend on what occurs ta any given time and, therefore, the poles and zeros of any assumed linear equivalent system move about as functions of time (and incidentally as functions of the particular error state).

As mentioned in the introduction, the fact does remain that the DOLIAC logic will assume a quasi-linear state whenever possible, and this indicates that simple error and error rate logic can indeed be optimum under some circumstances. The difficulty here, however, is that these circumstances can never continuously be obtained in any practical application, and the main reason for using optimum or adaptive logic is to account for real rather than purely idealized situations. Thus, the superiority of the DOLIAC logic over linear logic, per se, resides only in the fact that the DOLIAC logic, because of its switching law, can instantaneously and "logically" shift gain whenever it needs to.

At this point it is of interest to note that if we sought to minimize the quantity $(e_{n+1})_c^2$ as a criterion for optimum performance, subject to the same type of inequality constraint introduced in Section 2, then, by simple manipulation, we at once obtain a command logic of the form

$$\Delta\ddot{e}_c = K(e_{n+1})_{nc}$$

which is readily verified by the reader. This is the standard linear error and error rate control equation. Now, since this equation is derivable from an error squared criterion (recall that the error squared criterion tends to diminish emphasis on regulation as contrasted to control), we see that another way to separate the DOLIAC and linear logics is purely on a criterion basis. Thus, if we tamper with the criterion, we can eliminate the switching logic and end up with what is still mathematically an optimum solution, but not to the problem we are obliged to solve. We conclude that linear theory, therefore, seeks to minimize error squared and if this is desirable, then linear theory should be employed.

It is now pertinent to ask, how is the DOLIAC logic related to other switching logics, of which several have been proposed? The answer to this is simple. The DOLIAC switching logic is a result of applying a specific criterion and a specific root selection process both of which are entirely independent of the specific control problem. Thus, this logic results from definition and analysis rather than from empiricism and extrapolation. In other words, the DOLIAC logic has been designed on the basis of purely fundamental considerations, and as such is related to conventional control logic only after the fact. Its resemblance to conventional logic is conicidental--not purposeful, and because this is true, the DOLIAC developments serve to justify the efforts in classical control theory.

To the control engineer, one of the major theoretical drawbacks of the DOLIAC logic resides in the fact that its development provides no clue as to its stability characteristics. This is a theoretical problem requiring further study; however, the application of linear control logic to non-linear dynamic systems is equally displeasing in a theoretical sense, and since no physical system is really linear, the practical differences in the application of the DOLIAC logic and linear theory do not constitute a problem of real magnitude. As a matter of fact, a simple rule of thumb exists which, although conservative, makes practical application of the DOLIAC logic rather simple. The rule is "If one analyzes the DOLIAC logic as a constant gain linear system and it is stable for the low gain, then the full logic is also stable although a linear study will show that constant high gain operation would be unstable." This provides a simple method of selecting the gain constant K. Additionally, the ratio of prediction to sample times should be five or more. A system designed on the basis of these considerations will perform more effectively than any stable linear logic; however, there are some limitations to observe.

The switching logic and the location of the switching lines in an error-error rate phase plane are entirely dependent on the measured error to error rate ratio and should this ratio be improperly determined, the systems performance suffers. The most critical sensor is the one which is used to find the highest level error derivatives, and as a rule of thumb, the natural frequency of this sensor should exceed the system normal response frequency by an order of magnitude. Because of this restriction the DOLIAC logic should not be applied to the control of systems wherein reasonably accurate measurements of error rate are not possible. As a matter of fact, basic error and error rate linear systems exhibit this same characteristic and, frequently, poor sensor information may be worse than none. The sensor problem is not one of resolution but rather of phase lag since the logic can readily handle consistent residual errors,

but phase lags of anywhere near 90^{o} cause trouble unless the prediction of time Δt_p and the basic gain factor K are reduced, with a reduction of both of them leading to poorer command following ability.

It should now be of interest to briefly review the applications of the DOLIAC logic which have been investigated to date and to note the successes of the logic together with a significant failure.

The first application was to a simple micro control problem (auto pilot logic) involving a prefilter on command inputs. The analysis was by means of digital simulation of control of a non-linear vehicle throughout a complete environmental spectrum including the subsonic, transonic and supersonic regions both at high and low altitudes and subject to various classes of command input. Both perfect and imperfect sensors were examined. The results were extremely satisfactory and superior to those obtained by linear error and error rate comparison systems, and also superior to other types of gain switching logics which were simulated. The results obtained were subsequently checked at CAL using analog simulation of parts of the environmental pattern; duplication of conclusions was obtained. Additionally, autopilot operation was checked digitally without a prefilter and satisfactory results were again obtained.

The second application involved a macro-micro control problem, namely, the control of skin temperature of a re-entry vehicle by means of altering the flight path through a directly coupled autopilot. The basic error signal in this case being the difference in desired and sensed skin temperature. A sensed response factor "r" was used in the logic. Results were entirely satisfactory.

The third application was an attempt to control ICBM type flexible missiles for which accurate measurements of pitch rate by even complex sensor networks were not possible. Results were unsatisfactory since satisfactory error rate data could not be provided to the logic, and less sophisticated control configurations proved superior under simulation.

The fourth application was to combine longitudinal, lateral range and skin temperature control of a re-entry vehicle. An entirely satisfactory unified descent control system was evolved capable of handling, adaptively, and environmental pattern encountered. Error display information for pilot control was also generated by the logic.

The fifth application was to automatic landing descent and flare control, and an entirely satisfactory macro-micro control logic was synthesized.

The sixth application was to an automatic trim control system designed to operate in parallel with a human pilot, and study of the interaction of the control logic and the pilot is now being carried forth at ASD. A human pilot is included in the ASD simulation.

The seventh application has been to macro-micro control of explicit point and explicit orbit guided trajectories, again with entirely satisfactory results.

The eighth application was to quasi-optimum performance guidance and control of spaceplane type vehicles using limited sensors. Insufficient work has been conducted to fully evaluate this system; however, data obtained to date are favorable.

The ninth application was to the problem of macro-micro control of orbital docking maneuvers, and the results of the extensive simulation study of this system have been extremely satisfactory.

Thus, so far, we have a batting average of eight successes in nine application studies with one failure, with the reason for the failure being quite clear.

It should be recognized that because we have so far been unable to find an entirely proper stability criterion (the rule of thumb presented earlier is known to be too conservative), that the study of the DOLIAC logic and its applications have been largely pragmatic in nature and that the stipulations presented here are in no way based on abstraction or idealization, but rather upon the very practical criteria: "Does it work?" and "How does it work compared to other systems?" The answers to these questions have been "Yes, it does work in all cases where the theory makes sense" and "on a comparative basis it seems to work better than other systems we have compared it to." By no means have we compared all possible systems, nor do we in any sense feel that the present logic is an ultimate form, rather we feel that it represents a first step forward, that is, it doubtlessly can be improved upon, that it does call for more study and that it provides an affirmative answer to the question, "Is optimum synthesis practical?" Space has not permitted more than a summary of the several thousands of pages of analyses and data which have resulted from study of the DOLIAC logic. However, the basic theoretical elements are contained in this paper and listed below are a group of selected references which may be consulted for further information.

Finally, we want to note that the present logic is the result of evolution and that the early DODCO attempts at optimum adaptive synthesis reported on in reference 3 were vastly different from those discussed in this paper. Also, we want to express our thanks to ASD and to Capt Raymond Rath and Mr. P. C. Gregory for their initial interest in this effort and for their continued support of a rather complex area of research which has had its moments of extreme frustration.

REFERENCES

1. Rynaski, E. G. and Schuler, J. M., "The DODCO Optimal Control System - A Critical Evaluation," IRE paper presented at JACC Conference, New York University, June 28, 1962.

2. Rynaski, E. G. and Schuler, J. M., Application and Evaluation of Certain Adaptive Control Techniques in Advanced Flight Vehicles. Vol. II, The DODCO Theory and Application of Optimum Limited Information Adaptive Flight-Control Techniques, ASD TR 61-104, WPAFB, Ohio, 1961.

3. Gregory, P. C. (Ed.) Proceedings of the Self-Adaptive Flight Control Symposium, WADC TR 59-49, March, 1959, pgs. 216-251.

4. Mishkin, E. and Braun, L. (Jr.), Adaptive Control Systems, McGraw Hill, 1961.

5. Mykytow, W. J., Davis, H. M. et al (Ed.), Proceedings of Symposium on Aerothermoelasticity, ASD TR 61-645, WPAFB, Ohio, Nov., 1961, pgs. 468-497.

6. Dommasch, D. O. and Barron, R. L., The Theory and Application of Optimum Limited Information Adaptive Flight Control Techniques, WADC TR 59-468, WPAFB, Ohio, October 1959.

7. Dommasch, D. O. and Laudeman, C. W., Principles underlying Systems Engineering, Pitman, 1962.

8. Laudeman, C. W. and Muzyk, R. A., Accuracy Requirements for Micro Control Sensors and Logic Usable during Orbital Docking Maneuvers, DODCO TR Nr 123, June 1962.

9. Dommasch, D. O., Macro and Micro Control of the Descent of Advance Flight Vehicles Capable of Superorbital Speeds, ASD TR Nr. 61-23, WPAFB, Ohio, November 1960.

10. Stewart, W. G., The Examination of the Performance of an Optimum Flare-Landing Controller Applied to Advanced Flight Vehicles, DODCO TR Nr 96, May 1961.

11. Dommasch, D. O. and Stewart, W. G., Simulation of Operation of the DODCO Automatic Trim Control System Installed in the Cessna T-37 Aircraft Including the Effect of Sensor Inaccuracies and of Pilot Control Coupling, DODCO TR Nr. 107, November 1961.

ASD-TDR-63-119

The Approximate Realization of Optimum Control Systems

By

S. S. L. Chang
F. J. Alexandro, Jr.

New York University
New York City, New York

THE APPROXIMATE REALIZATION OF OPTIMUM CONTROL SYSTEMS

S.S.L. Chang
F.J. Alexandro, Jr.

ABSTRACT

The paper describes two ways of approximate realization of Pontryagin and Bellman theories. The more sophisticated method involves combining the features of both theories and use of both digital and analog feedback. The less sophisticated but more economical method employs simple analog nonlinear elements. Both methods are applicable to high order systems with random as well as deterministic inputs. A comparison of the limitations and similarities of the two methods are given.

Both methods display bang-bang characteristics for large inputs, but linear stable characteristics when the error signal is low. A reserve control force is provided (in both methods) to counteract random distrubances, and minor variations in plant transfer function.

The more economical method is applied to the positioning and stabilization of an inertial platform which has a fourth order transfer function and appreciable load disturbances. Analog studies show that the approximate optimum control not only out-performs linear controls in every way by a wide margin, but is also much simpler to implement and requires less weight.

INTRODUCTION

Two types of difficulties are encountered in applying optimum control theories:

1. Physical difficulties
   a. the existence of random noise and disturbance.
   b. the lack of knowledge and/or the change in plant dynamics.
2. Computational difficulties
   a. the two point boundary value problem in maximum principle.
   b. the requirement of large storage capacity in dynamic programming.

The present paper deals with methods of overcoming these difficulties. For the convenience of utilizing an earlier report without costly revision, the paper is made up of two integral parts with independent equation and figure numbers. Part I (by Chang) is on the general philosophy and method. Part II (by Alexandro and Chang) is on a special application in which the general philosophy of Part I is realized in a very simple and practical manner. It should be noted that while the special problem dealt with in Part II is that of optimum control of an inertial platform, the method is applicable to any system with no more than two dominant lagging time constants.

PART I

GENERAL PHILOSOPHY AND METHOD

Systems with Random Noise and Disturbance

A system with random disturbances can be described by the set of equations

$$\dot{\underline{x}} = \underline{f}(\underline{x},\underline{u},\underline{v},t) \tag{1}$$

where $\underline{x}$, $\underline{u}$, and $\underline{v}$ are the state vector, control vector, and disturbance vector respectively, and $\underline{f}$ is a vector function. Generally, $\underline{x}$ and $\underline{f}$ have the same dimensions, while the dimensions of $\underline{u}$ and $\underline{v}$ are different. The function $\underline{f}$ has the proper continuity properties so that given $\underline{x}(t_o)$, and $\underline{u}(t)$, $\underline{v}(t)$ for all $t \geq t_o$, $\underline{x}(t)$ is uniquely determined.

The control vector $\underline{u}$ can be selected at will within given constraints either as a function of $\underline{x}$ or as a function of t. Following Bellman, we shall refer to such a selection as a control policy. In a deterministic system, $\underline{v}(t) = 0$. Given the control policy, and $\underline{x}(t_o)$ the subsequent $\underline{x}(t)$ is completely determined. Therefore we can use $\underline{x}(t)$ to describe the system. In a system with random disturbances, however, $\underline{v}(t)$ is not zero nor predictable. Given the control policy and $\underline{x}(t_o)$, the subsequent $\underline{x}(t)$ is unknown.

The only predictable aspect of the system is then a probabilistic one; Let $\Delta\tau$ represent a small volume element in state space about $\underline{x}_1$, then the probability of finding the representative point $\underline{x}(t)$ of the system in $\Delta\tau$ is proportional to the volume of $\Delta\tau$:

$$p(\underline{x}(t) \in \Delta\tau) = \rho(\underline{x}_1, t)\,\Delta\tau \tag{2}$$

The probability density ρ is a function of two variables: the point $\underline{x}_1$ under consideration and time t. The uncertain state of the system is represented by $\rho(\underline{x}_1, t)$. If as t increases, ρ converges towards a small neighborhood near the desired point $\underline{x}_d$ in state space, the system is a convergent and satisfactory system. If ρ spreads out the system is not satisfactory. The speed and compactness of the convergence of ρ towards $\underline{x}_d$ are measures of the merit of the system.

Partial Differential Equation of ρ

In most control systems, the random disturbances are small compared to the control forces, and (1) can be written as

$$\underline{\dot{x}} = \underline{f}(\underline{x},\underline{u},0,t) + \sum_i v_i \frac{\partial \underline{f}}{\partial v_i} \tag{3}$$

Given the control policy, $\underline{u}$ is a function of $\underline{x}$ and t, and $\underline{f}(\underline{x},\underline{u},0,t)$ can be written as $\underline{F}(\underline{x}, t)$.

Eq. (3) can be rewritten as

$$\dot{\underline{x}} = \underline{F}(\underline{x},\ t) + \sum_i v_i \frac{\partial f}{\partial v_i} \tag{4}$$

Eq. (4) describes the system with a single sample of $\underline{v}$. Considering all possible $\underline{v}$, the variation of probability density ρ with t at any particular location $\underline{x}_1$ can be described by

$$\frac{\partial \rho}{\partial t} = \left(\frac{\partial \rho}{\partial t}\right)_o + \left(\frac{\partial \rho}{\partial t}\right)_v \tag{5}$$

where $(\partial\rho/\partial t)_o$ gives the part due to mean motion $\underline{F}$, and $(\partial\rho/\partial t)_v$ gives the part due to random disturbance $\underline{v}$.

The probability density can be thought of as the substance or mass of a fluid. The conservation equation in hydrodynamics gives

$$\left(\frac{\partial \rho}{\partial t}\right)_o + \underline{\nabla} \cdot (\rho \underline{F}) = 0 \tag{6}$$

The effect of the random disturbances is to cause a diversion of the system from its expected path. The sum total of these diversions is a "diffusion" of the probable states of the system.

We note that in Eq. (3) $\underline{v}$ is a linear additive term. If a certain $\underline{v}(t)$ occurs for an interval δt, the representative point $\underline{x}$ is displaced by an amount $\delta\underline{x}$ which is independent of $\underline{x}$. We can

therefore imagine that the entire distribution pattern $\rho(\underline{x}, t)$ is shifted by $\delta\underline{x}$. The new ρ at $\underline{x}_1$ is therefore identical with the undisturbed ρ at $\underline{x}_1 - \delta\underline{x}$. The change in ρ is

$$\delta\rho = \rho(\underline{x}_1 - \delta\underline{x}, t) - \rho(\underline{x}_1, t)$$

$$= \sum_i \left(\frac{\partial\rho}{\partial x_i}\right)(-\delta x_i) + \frac{1}{2}\sum_i\sum_j \frac{\partial^2\rho}{\partial x_i \partial x_j}(\delta x_i)(\delta x_j) \qquad (7)$$

$$+ \ldots$$

Taking ensemble average of (7) gives

$$\delta\rho = \frac{1}{2}\sum_i\sum_j \frac{\partial^2\rho}{\partial x_i \partial x_j}\overline{(\delta x_i)(\delta x_j)} \qquad (8)$$

The ensemble averages of the odd order terms vanish because $\underline{v}$ and $-\underline{v}$ are equally likely, and the fourth and higher order terms are negligible.

In control systems, the disturbances generally have much wider bandwidths than that of the system. In terms of time, the correlation times of v_i are much shorter than the system time constants. We can select a δt which is much shorter than the system time constants but longer than the correlation time of v_i, and calculate the ensemble averages $\overline{(\delta x_i)(\delta x_j)}$ occasioned by $\underline{v}$ in time δt.

Let the initial time be denoted t_1. Eq. (3) gives

$$\delta\dot{\underline{x}}(t) = \sum_i \frac{\partial f}{\partial v_i} v_i(t) \tag{9}$$

$$\delta\underline{x}(t) = \int_{t_1}^{t} \delta\dot{x}(\tau)\, d\tau = \sum_i \frac{\partial f}{\partial v_i} \int_{t_1}^{t} v_i(\tau)\, d\tau \tag{10}$$

The partial derivatives can be left outside the integral sign because they do not change appreciably in any interval which is much shorter than the system time constants. It follows from Eqs. (9) and (10)

$$\frac{d}{dt}\left[(\delta x_i)(\delta x_j)\right] = (\delta\dot{x}_i)(\delta x_j) + (\delta\dot{x}_j)(\delta x_i)$$

$$= \sum_k \sum_l \frac{\partial f_i}{\partial v_k} \frac{\partial f_j}{\partial v_l} \int_{t_1}^{t} \left[v_k(t)\, v_l(\tau) + v_k(\tau)\, v_l(t)\right] d\tau$$

Taking ensemble average of the above equation gives

$$\frac{d}{dt}\overline{(\delta x_i)(\delta x_j)} = \sum_k \sum_l \frac{\partial f_i}{\partial v_k} \frac{\partial f_j}{\partial v_l} \int_{t_1}^{t} \left[\Phi_{kl}(\tau - t) + \Phi_{kl}(t - \tau)\right] d\tau \tag{11}$$

where Φ_{kl} is the correlation function of v_k and v_l. Substituting λ for $t-\tau$ in Eq. (11) and integrating the result from t_1 to $t_1+\delta t$ gives

$$\overline{(\delta x_i)(\delta x_j)} = \sum_k \sum_l \frac{\partial f_i}{\partial v_k} \frac{\partial f_j}{\partial v_l} \int_{t_1}^{t_1+\delta t} \int_0^{t-t_1} [\phi_{kl}(-\lambda) + \phi_{kl}(\lambda)]\, d\lambda\, dt$$

$$= \sum_k \sum_l \frac{\partial f_i}{\partial v_k} \frac{\partial f_j}{\partial v_l} \int_0^{\delta t} [\phi_{kl}(-\lambda) + \phi_{kl}(\lambda)]\, (\delta t - \lambda)\, d\lambda \qquad (12)$$

Since the correlation time is very short,

$$\int_0^{\delta t} [\phi_{kl}(-\lambda) + \phi_{kl}(\lambda)]\, (\delta t - \lambda)\, d\lambda$$

$$= \int_0^{\delta t} [\phi_{kl}(-\lambda) + \phi_{kl}(\lambda)]\, \delta t\, d\lambda \qquad (13)$$

$$= \delta t \int_{-\infty}^{\infty} \phi_{kl}(\lambda)\, d\lambda = \delta t\, \Phi_{kl}(0)$$

where $\Phi_{kl}(j\omega)$ is the cross-spectral density of v_k and v_l. Eq. (12) becomes

$$\overline{(\delta x_i)(\delta x_j)} = D_{ij}\, \delta t \qquad (14)$$

The factors D_{ij} are defined as

$$D_{ij} = \sum_k \sum_l \frac{\partial f_i}{\partial v_k} \frac{\partial f_j}{\partial v_l} \phi_{kl}(0) \tag{15}$$

Substituting (14) into (8) gives

$$\delta\rho = \tfrac{1}{2} \sum_i \sum_j \frac{\partial^2 \rho}{\partial x_i \partial x_j} D_{ij} \, \delta t$$

Therefore

$$\left(\frac{\partial \rho}{\partial t}\right)_v = \frac{\delta\rho}{\delta t} = \tfrac{1}{2} \sum_i \sum_j D_{ij} \frac{\partial^2 \rho}{\partial x_i \partial x_j} \tag{16}$$

Finally, by substituting (6) and (16) into (5) we obtain

$$\frac{\partial \rho}{\partial t} = -\nabla \quad (\rho \underline{F}) + \tfrac{1}{2} \sum_i \sum_j D_{ij} \frac{\partial^2 \rho}{\partial x_i \partial x_j} \tag{17}$$

Eq. (17) is a partial differential equation which gives the variation of ρ with time. Its general solution is quite beyond our present stage of development in mathematics. However, by examining a few special cases, we can gain some insight on the effects of random disturbances on systems of different designs.

One Dimensional Solutions

Consider the sample case with a state variable x of only one dimension. Equation (17) is reduced to

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho F) - \frac{D}{2}\frac{\partial^2 \rho}{\partial x^2} = 0 \tag{18}$$

Example 1. The function F is given by

$$F = -ax$$

and a, D are constants. In steady state Eq. (18) becomes

$$a\frac{d}{dx}(\rho x) + \frac{D}{2}\frac{d^2\rho}{dx^2} = 0 \tag{19}$$

Integrating (19) gives

$$a\rho x + \frac{D}{2}\frac{d\rho}{dx} = C_1 \tag{20}$$

where C_1 is a constant independent of x. Note that in order to satisfy the equation

$$\int_{-\infty}^{\infty} \rho dx = 1 \tag{21}$$

both the product ρx and $\frac{d\rho}{dx}$ must vanish at infinity. Therefore $C_1 = 0$,

and (20) becomes

$$\frac{1}{\rho}\ \frac{d\rho}{dx} + \frac{2ax}{D} = 0 \qquad (22)$$

Integrating the above expression gives

$$\rho = C_2\, e^{-ax^2/D} \qquad (23)$$

where C_2 is a constant and can be determined from (21). Eq. (23) represents a packet of r.m.s. distance $\sqrt{D/2a}$ from the origin. The system converges eventually into this packet.

Example 2.

$$F(x) = 0 \quad \text{for} \quad x > 0$$

There is no restoring force on the positive side of x. It will be shown in general that there is no steady state solution and the probability density flows in the positive x direction until all is exhausted.

The total probability of the system above some positive value of x, say b is given by

$$p(x > b) = \int_b^\infty \rho(x,t)\ dx$$

As F = 0, Eq. (18) gives

$$\frac{dp}{dt} = \int_b^\infty \frac{\partial \rho}{\partial t}\, dx = \frac{D}{2} \left.\frac{\partial \rho}{\partial x}\right|_b^\infty = -\frac{D}{2}\left(\frac{\partial \rho}{\partial x}\right)_{x=b} \qquad (24)$$

Since dp/dt represents the rate of increase of $p(x > b)$, conservation law requires that the expression on the right hand side of Eq. (24) represents the rate of flow of the probability density passing the point x = b. As b is arbitrarily selected, this relation holds for all positive values of x:

$$\text{Rate } (x) = -\frac{D}{2}\left(\frac{\partial \rho}{\partial x}\right) \qquad (25)$$

The integrated rate of flow is

$$\int_0^\infty \text{Rate } (x)\, dx = -\frac{D}{2} \int_0^\infty \frac{\partial \rho}{\partial x}\, dx = \frac{D}{2}\left[\rho(o) - \rho(\infty)\right] \qquad (26)$$

In order to satisfy (21), $\rho(\infty) = 0$. While (26) does not rule out transient flow in the -x direction at some particular location, it does prove that on the whole the probability density moves out in the +x direction until $\rho(0) = 0$.

Practical Implications - Measure to Counteract Disturbances

While the above calculations are for a one-dimensional problem about x = 0, the situation is basically the same in a multi-dimensional problem about the expected or undisturbed path. In a

bang-bang control system, the switching surfaces can be followed only by exerting maximum available control force. Once the representative point is sent beyond the switching surface by the disturbances, there is no available control force to restore it to the expected path. The situation is quite similar to that of Example 2. The probability is overwhelming that the system does not follow the expected path or even stay close to it.

A corrective measure is suggested by Example 1. If the controller is so designed that a large part but not all the control force is used to follow the expected trajectory, a linear corrective region can be set up about the latter. The required linear range is no more than $\sqrt{D/a}$. The probability density ρ is then focussed into a small packet about the expected path or $\underline{x}(t)$. Unless the disturbances are extremely severe, the required margin in control force is comparatively small. It is a small price to pay to gain surer control.

Measure to Counteract Uncertainties in Plant Dynamics

There are many papers published on the subject of adaptive control, model feedback systems etc., and our discussion on this subject will be brief. Two methods stand out as effective measures to counteract plant uncertainties:

1. Estimating the plant dynamics from past response, and using the latest estimate for the computation of optimum trajectory.

2. Employing minor-loop feedback to force the system response to a preselected model.

Of the two methods, (1) is usually slower and more costly in comparison, but for slow-varying systems with a wide range of variation, it has the advantage of making better or more effective utilization of the control forces. For these reasons, (2) perhaps should be employed first, and (1) is used in addition only in case of necessity.

In order to make (2) effective, the required minor-loop gain is usually quite high. Thus lead networks are needed with high noise as a result. An alternative is to use the directly-measured highest time-derivative as the feedback signal. This has the advantage of lead stabilization but without the difficulties resulting from differentiating instrument noise.

Mathematical Difficulties and A Counter Measure

There are two basic methods of optimal control:[1] Bellman's dynamic programming and Pontryagin's maximum principle. Their mathematical difficulties, which are mentioned in the introduction, are well known and do not need further explanation. The proposed counter measure utilizes features of both theories. Consider the problem

1. For an introductory description of these methods see Chapters 9 and 12 of S. S. L. Chang, "Synthesis of Optimum Control Systems," McGraw Hill, New York, 1961.

of sending the system from a fixed initial point $\underline{x}_1$ to a fixed final point $\underline{x}_2$ with a minimum of cost. The least cost from each point $\underline{x}$ in the neighborhood of $\underline{x}_2$ can be computed by dynamic programming. The constant cost surfaces in the neighborhood of $\underline{x}_2$ are then stored in the computer. By assuming an initial covariant vector $\underline{\psi}$, an optimum trajectory can be computed to go from $\underline{x}_1$ to some point $\underline{x}_3$ in the neighborhood of $\underline{x}_2$. With a fast computer, a few of these trajectories can be computed before a decision needs to be made. The one with least total cost is then selected. The basic idea is illustrated in Fig. I-1.

Because only the cost values of points in a small neighborhood of $\underline{x}_2$ needs to be computed and stored, the required storage capacity is usually not prohibitive. Because the final point of the Pontryagin trajectory is a finite neighborhood rather than a single point, the trial process in the selection of an initial covariant vector is also not overly difficult. In any case the overall difficulty can be minimized by a sensible selection of the size of the neighborhood about $\underline{x}_2$.

A Proposed Digital Control System

A combination digital-analog controller which embodies all the above discussed counter measures is illustrated in Fig. I-2. Using the method of Fig. I-1, a nearly optimum trajectory is computed

by the digital computer on the basis of a maximum control force which is no more than a large fraction (say 90%) of the actual available value. There are two outputs from the computer: the optimal control force $\hat{\underline{u}}$ and the desired acceleration $\hat{\ddot{\underline{x}}}$. The actual control force applied to the plant is $\hat{\underline{u}} + \delta\underline{u}$, where $\delta\underline{u}$ is the corrective force to keep $\ddot{\underline{x}} - \hat{\ddot{\underline{x}}}$ small.

Besides the immediate feedback signal $\ddot{\underline{x}}$, the actual $\underline{x}$ is also measured and this information is fed back to the computer. From the measured $\underline{x}$ and $\hat{\underline{u}} + \delta\underline{u}$, the computer estimates the system dynamics and uses this information and the measured $\underline{x}$ to revise the optimal path data $\hat{\underline{u}}$ and $\hat{\ddot{\underline{x}}}$. The parameter $\underline{\alpha}$ in Fig. I-2 is used to indicate that the system dynamics may change.

There is another reason for using the acceleration signal for inner loop feedback: In order for such feedback to be effective, the loop gain must be high, and time delay in measuring the feedback variable must be kept at a minimum. Suppose the controlled plant is a space vehicle. Its acceleration can be measured directly and instantly right on the vehicle, but position and velocity have to be determined in roundabout ways.

The general method illustrated in Fig. I-2 is overly elaborate or too costly for most applications. However, if the plant has only two dominant lagging time constant (including $1/\rho$ which represents the limit of $T \to \infty$), and its transfer function is known, the counter-

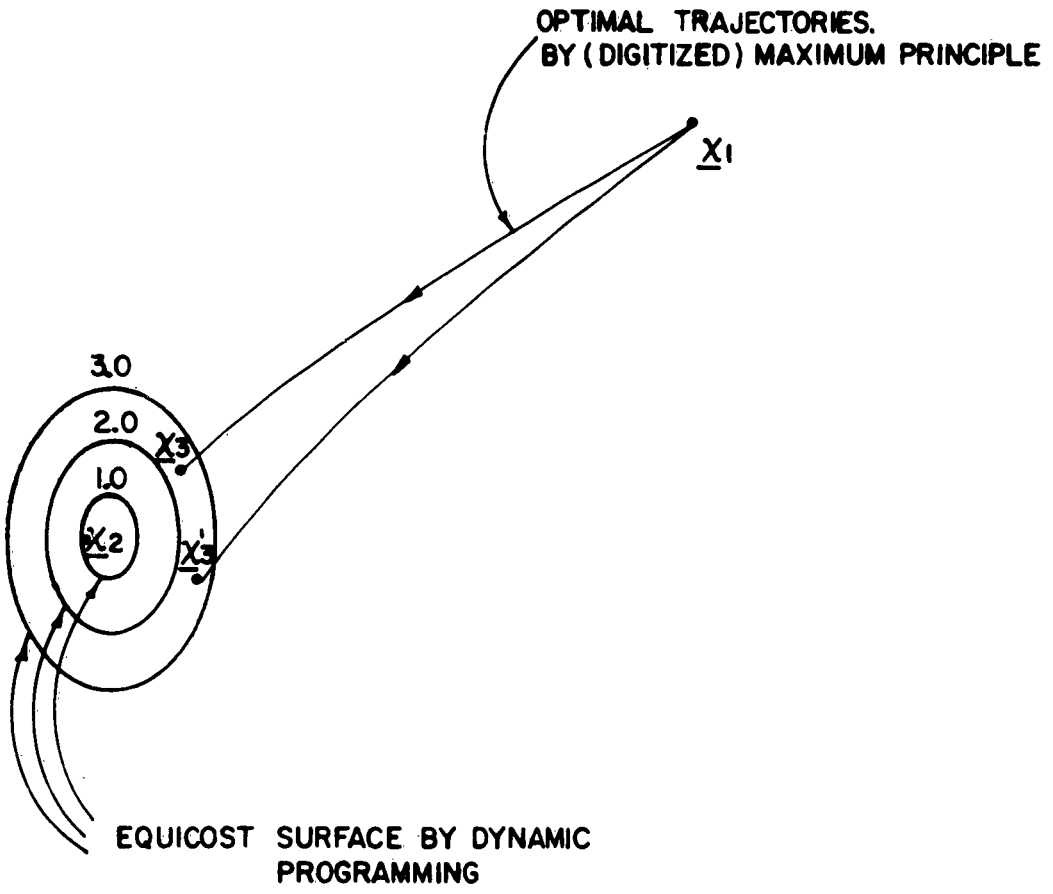


Figure I-1. Computation of Optimal Trajectory

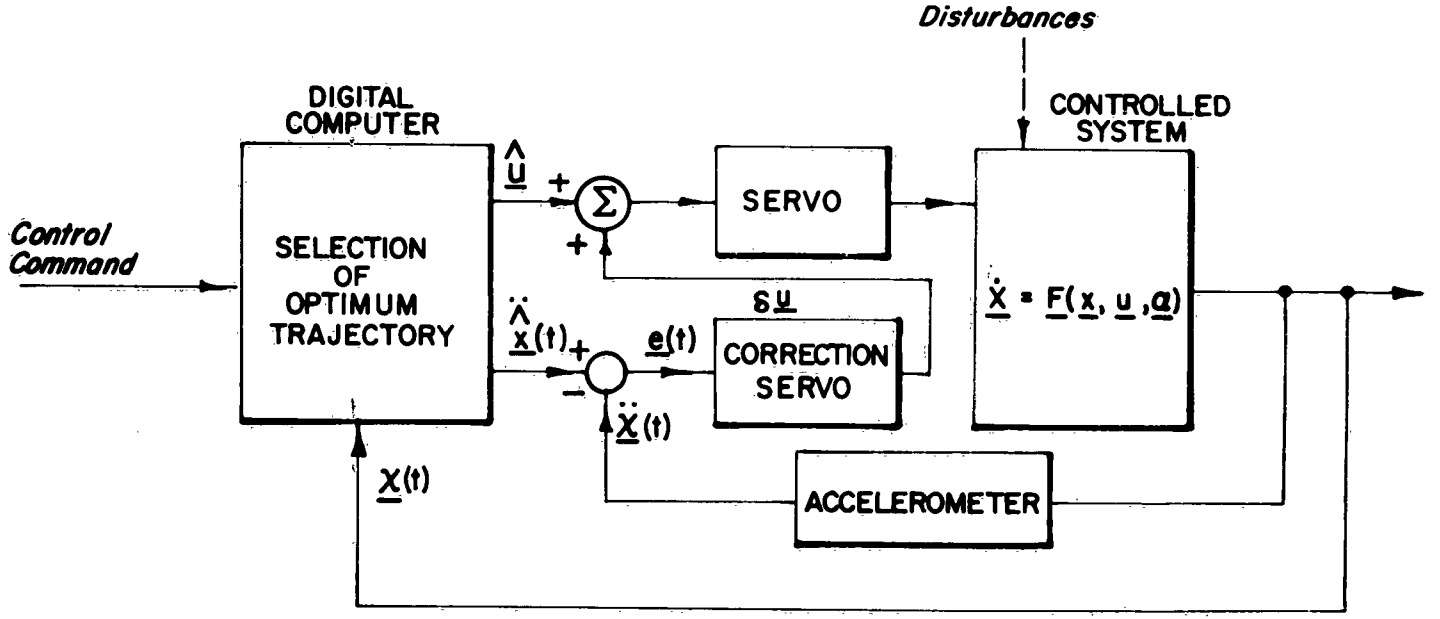


Figure I-2. Computer-Controlled Nonlinear System

measures discussed above can be realized by a very simple and practical controller using only analog elements. The simpler model will be described in Part II.

PART II

APPROXIMATELY OPTIMUM CONTROL OF AN INERTIAL PLATFORM WITH RANDOM DISTURBANCES

In this part of the paper we will discuss a number of techniques which can be used to modify the bang-bang principle in order to overcome some common problems such as finite gain, random inputs, and minor time constants, namely:

1. Automatic dual mode.
2. Provision for reserve torque
3. Means for incorporating lag compensation.

An example will also be given in which the methods are incorporated into an approximate bang-bang system to control an inertial platform.

Automatic Dual Mode System

The problem of positioning an inertial load which may be subject to large inputs of displacement was considered by McDonald in 1950.[1] He demonstrated that the system could be made to respond to a step input in the minimum time by using a relay controller which reverses the torque in accordance with the sign of the expression:

$$\frac{2T}{J} e + \dot{e} \mid \dot{e} \mid \quad ;$$

where T = Maximum torque output ,

J = Inertia of load ,

e = Error .

The dividing line is shown in Fig. 1. By controlling the torque in this manner the system will follow the optimum trajectory in the phase plane and only one torque reversal is required in order to bring the output to the correct position.

The disadvantage of the relay controller is that due to small time delays present, a limit cycle will normally exist; with the result that the performance of the system will

be marred by the presence of small oscillations in the output. In addition, small load disturbances will cause much larger errors in a relay system than they would in a linear system.

To remedy the defects of the relay system, McDonald proposed a dual mode servo system[2] which would combine the advantages of a relay system and a linear system. For large inputs the torque is controlled by a relay so as to follow the optimum trajectory in the phase plane. When the magnitude of the error is reduced below a certain value, the relay is switched out and replaced by a linear amplifier. Thus optimum switching can be employed for large errors, while avoiding both the limit cycle inherent in a relay system, and the ineffectiveness of the relay servo in compensating for small random load disturbances.

It is not necessary to switch between a relay and a linear amplifier in order to obtain the desirable characteristics of the dual mode system; for a high gain servo system subject to torque saturation can be designed to provide automatically the characteristics of the dual mode system. To achieve automatic dual mode operation the switching boundary used for the relay system must be modified somewhat to account for the effect of finite gain and to provide a linear region. In order to force the system to follow the optimum trajectory

when the gain is not infinite, it is necessary that torque saturation occur at the optimum trajectory. The system will then follow the optimum trajectory until the linear region is reached. Thus zero torque must occur somewhat ahead of the optimum trajectory to insure that full torque is obtained at the instant the optimum trajectory is reached. This can be accomplished by constructing the neutral boundary (locus of points in the e vs $\dot{e}$ plane for which the torque is zero) shown in Fig. 2. This neutral boundary is constructed by shifting the left half of the optimum trajectory to the left by an amount Δ, and the right half of the optimum trajectory to the right by an amount Δ; where Δ is equal to the value of error which will produce torque saturation.

Thus:

$$\Delta = T/K \quad ; \tag{2}$$

where

T = Maximum Torque Output ,

K = Gain from Summing Point to Torque Output .

The two halves of the curve are then joined by the straight line whose equation is $e + \tau\dot{e} = 0$; where τ is chosen so that the straight line is tangent to the two parabolas; thereby providing a smooth transition between the nonlinear and linear

portions of the curve. Thus by using this neutral curve, we have obtained our goal of having the error reduced to zero in the minimum time, while still maintaining the desirable characteristics of a linear system.

The equation for the neutral curve is given in three parts:

$$\begin{aligned} \frac{2T}{J}\, e + \dot{e}^2 &= -\frac{2T}{J}\,\Delta \quad ; \quad e < -e_1 \quad , \\ e + \tau\dot{e} &= 0 \quad ; \quad -e_1 \leq e \leq e_1 \, , \\ \frac{2T}{J}\, e - \dot{e}^2 &= \frac{2T}{J}\,\Delta \quad ; \quad e > e_1 \quad , \end{aligned} \tag{3}$$

where e_1, the value of e at which the linear and parabolic portions of the curve join, and τ, the time constant in the linear region, can be determined as follows. At the point of intersection to the left of the origin:

$$-\frac{2T\tau}{J}\,\dot{e}_1 + \dot{e}_1^{\,2} = -\frac{2T}{J}\,\Delta \quad , \tag{4}$$

where $\dot{e}_1$ is the value of $\dot{e}$ at the point of intersection.

$$\dot{e}_1^{\,2} - \frac{2T\tau}{J}\,\dot{e}_1 + \frac{2T}{J}\,\Delta = 0 \quad , \tag{5}$$

$$\dot{e}_1 = \frac{\frac{2T\tau}{J} \pm \sqrt{\left(\frac{2T\tau}{J}\right)^2 - \frac{8T\Delta}{J}}}{2} \quad . \qquad (6)$$

If the linear portion of the curve is to be tangent to the parabolic protion then $\dot{e}_1$ must have only one value. Therefore,

$$\left(\frac{2T\tau}{J}\right)^2 = \frac{8T\Delta}{J} \quad ; \qquad (7)$$

where $$\tau = \frac{2J\Delta}{T} \quad , \qquad (8)$$

and $$\dot{e}_1 = \frac{T\tau}{J} = \sqrt{\frac{2T\Delta}{J}} \quad , \qquad (9)$$

and $$e_1 = 2\Delta \quad . \qquad (10)$$

The equations (8) and (10) give the values of τ and e_1 in (3). This neutral curve can be generated by producing a nonlinear function of $\dot{e}$ such that:

$$f(\dot{e}) = \frac{J}{2T}\dot{e}^2 + \Delta \quad ; \quad \dot{e} > \sqrt{\frac{2T\Delta}{J}} \quad ,$$

$$f(\dot{e}) = \sqrt{\frac{2J\Delta}{T}}\,\dot{e} \qquad ; \qquad \sqrt{\frac{2T\Delta}{J}} > \dot{e} > -\sqrt{\frac{2T\Delta}{J}} \quad ,$$

$$f(\dot{e}) = -\frac{J}{2T}\,\dot{e}^2 + \Delta \quad ; \quad \dot{e} < -\sqrt{\frac{2T\Delta}{J}} \quad . \qquad (11)$$

The function $f(\dot{e})$ is then added to e as shown in Fig. 3. Since the torque is now controlled by $e + f(\dot{e})$, the torque is zero when $e + f(\dot{e}) = 0$ and hence the correct neutral curve is generated.

The function $f(\dot{e})$ is not difficult to generate since it consists of a linear region and a region characterized by a square law relation.

While it may seem somewhat arbitrary to have chosen τ so that the parabolic and linear portions are tangent, there are some definite advantages. Not only does this insure a smooth transition between the linear and nonlinear range, but this choice guarantees that in the linear range the system will be stable with a damping factor of 0.707. This can be easily shown as follows; consider the system shown in Fig. 3. In the linear region,

$$G(s) = \frac{(1 + \tau s)\,K}{Js^2} \quad , \qquad (12)$$

where $G(s)$ = open loop transfer function.

The control ratio (closed loop transfer function) is

$$\frac{C}{R}(s) = \frac{G(s)}{1 + G(s)} \quad ,$$

$$= \frac{\frac{(1 + \tau s) K}{Js^2}}{\frac{(1 + \tau s) K}{Js^2}} \quad ,$$

$$= \frac{(1 + \tau s) K}{Js^2 + (1 + \tau s) K} \quad . \qquad (13)$$

The poles of $\frac{C}{R}(s)$ are the solutions of the equation:

$$Js^2 + (1 + \tau s) K = 0 \quad ,$$

or

$$s^2 + \sqrt{\frac{2K}{J}}\, s + \frac{K}{J} = 0 \quad . \qquad (14)$$

This can be written as

$$s^2 + 2 \xi \omega_0 s + \omega_0^2 = 0 \quad ;$$

where

$$\omega_0^2 = \frac{K}{J} \quad ,$$

$$\therefore \xi = \frac{1}{2\omega_0} \sqrt{\frac{2K}{J}} = 0.707 \quad . \qquad (15)$$

Reserve Torque

The foregone discussion assumes that the system is of second order; in reality there will always be other time constants present in any practical system. When there are other time constants present, an exact solution to the problem of positioning the output of the system in the minimum time is made considerably more complicated. In this case it would be necessary to generate neutral curves and surfaces in n-dimensional phase space, where n is equal to the order of the system. However if the gain of the system is not too high the effect of the other time constants can be accounted for by modifying the neutral curve in the phase plane so as to produce control which is approximately optimum.

Consider the case of the neutral curve that is not modified. As a result of the storage effect of the minor time constants, the actual trajectory BCD in the phase plane would overshoot the trajectroy AO which passes through the origin. This is shown in a somewhat exaggerated manner in Fig. 4. Since the full torque is already expended, there is no torque available to return the system to AO; with the result that more than one torque reversal is necessary in order to reach the equilibrium point in the phase plane. This situation can be remedied by altering the neutral curve slightly

so that it no longer has the same shape as the trajectory in the phase plane. If the neutral curve is made somewhat "flatter", then less than the saturation torque is sufficient to move the system along the upper saturation limit. The actual trajectory will approach the neutral curve, with the result that only one reversal of torque is required to reach the equilibrium point. This is shown (also exaggerated) in Fig. 5.

Notice that with this modification the applied torque is not the maximum available torque over the entire trajectory, but is reduced to a value slightly below maximum for part of the trajectory. Full torque is applied only when the actual trajectory is being forced toward the saturation limits, after which slightly less than full torque is applied.

Thus we have provided a "reserve torque" which is used only for a portion of the trajectory. Since this "reserve torque" is small, full torque or almost full torque is always applied. This system is only slightly slower than an optimum system, in which full torque is applied for the entire trajectory.

This reserve torque is also extremely valuable when there are random disturbances. Suppose a disturbing torque occurs when the system is in the unsaturated zone. Without

any reserve torque the system can be brought back to the unsaturated zone only if the disturbing torque is in the same direction as the control torque. With reserve torque the system is brought back to the unsaturated zone irrespective of the direction of the disturbing torque.

Saturable Lag Network

In most practical servo systems it is generally desirable to have a high static gain so as to keep the error due to load disturbances low. However, as the gain is made larger the minor time constants become more troublesome; not only because they complicate the problem of reaching the equilibrium position in the minimum time, but also because they may cause the system to become unstable in the linear region.

A system could be made to have high static gain, and still be stable in the linear range, by adding integral compensation as shown in Fig. 6. However, the problem of obtaining optimum control for large step inputs is made even more complicated. Since no useful purpose is served by allowing the output of the integrator to exceed the value required for full output torque, the difficulty is resolved by limiting the output of the integrator.

As the load disturbances are generally much smaller

than the maximum output torque, the integrator output can be limited at a value considerably less than that necessary to provide full output torque. With this arrangement, a constant load disturbance, which is not sufficiently large enough to cause the integrator to saturate, will not produce a steady state error. However, when a large step input of displacement is applied, the output of the integrator will be much smaller than either $\dot{e}$ or e, and hence will have a negligible effect on the performance.

An Example

In order to demonstrate the feasibility of the methods presented in this paper, they were applied to the design of a control system to stabilize an inertial platform. A system of this type is generally used to maintain the platform at a fixed position within a few seconds of arc, in the presence of load disturbances; except when the platform is moved to a new position, at which time the input is a step of many degrees of arc. Hence the system has the following characteristics:

1. Extremely low error for constant load torque up to 30% of full output torque.
2. Minor time constants are presented.
3. The system is subjected to step inputs of displacement

hundreds of times larger than necessary to produce torque saturation.

The system is shown in Fig. 7. In the linear region:

$$G(s) = \frac{4 \times 10^4(0.1s+1)(0.033s+1)}{s^3(2 \times 10^{-3}\ s+1)^2} = \frac{3.33 \times 10^7(s+10)(s+30)}{s^3(s+500)^2} \qquad (16)$$

The root locus plot is shown in Fig. 8. For the above value of K the closed loop poles are in the left half plane and provide a reasonable margin of stability. However, as the system is only conditionally stable, instability will arise for inputs large enough to cause saturation, if only linear compensation is used.

The integrator output was limited at a value which provides slightly more than 30% of full output torque, thus satisfying the requirement of low error for load torques up to 30% of full output torque. The function $f(\dot{e})$ was first generated to provide automatic dual mode operation without provision for "reserve torque".

The system was simulated on an analog computer. Fig. 9 illustrates the response to a step input of displacement one hundred times larger than necessary to produce saturation. The corresponding phase plane plots with inputs of different

magnitudes are illustrated in Fig. 10. These curves show that two reversals of torque are required before the linear region is reached. The extra torque reversal is a consequence of the two minor time constants located at $T = 2 \times 10^{-3}$ sec.

The performance of the previous system was improved by providing "reserve torque". This was accomplished by multiplying $f(\dot{e})$ by a factor of 1.13 before adding it to e, thus producing the necessary "flattening" of the neutral curve. The step response (Fig. 11) and phase plane plot (Fig. 12) show that only one torque reversal is now required in order to reach the linear region. It is to be noted that, even with this provision for "reverse torque," full torque is provided for almost the entire trajectory, and the settling time in Fig. 10 is about 30% shorter than that in Fig. 8.

Step responses and phase plane plots were also obtained for a second system employing linear compensation without limiting the integration (Fig. 13a and 14); for a third system employing linear compensation with the integration limited (Fig. 13b and 15); and for a fourth system using the non-linear neutral curve, but with the integration not limited (Fig. 13 c and 16). Comparison of these curves with those obtained for the previous system clearly illustrates the superiority of the system which employs the non-linear curve and has the integration limited.

These curves show that the second system is unstable for large inputs, and that the third and fourth systems require a much longer time, and several more torque reversals before settling.

The importance of the integration in the system is illustrated by the response to a step input of torque disturbance. If the integration is removed then a step input of torque disturbance results in a steady state error (Fig. 17a); while no steady state error results when the integration is present (Fig. 17b).

The performance of a dual mode system is superior to a relay system since the automatic dual mode system behaves as a linear system when subjected to a random load disturbance which does not exceed the maximum torque output; with the result that the error is very much less than would be obtained if the system were of the relay type. This can be seen from inspection of Fig. 18 which shows the response of the automatic dual mode system to a torque disturbance which is an arbitrary function of time.

CONCLUSION

This paper has presented several simple methods for obtaining approximately optimum control in a practical system

with minor constants. These give approximately the response of an optimal bang-bang system with large inputs, while still maintaining the desirable response of a linear system to random torque disturbances. The required control function, which consists of a linear region and a region characterized by a square law relation, can be generated by the use of a few diodes.

The results of an Analog Computer study demonstrate the value of the methods proposed in the paper. Comparisons of the performance of a system incorporating these methods with the performance of several other systems illustrate how the proposed methods improve the stability and reduce the settling time.

NOTE: The research contained in this report has been sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research (USAF), Washington 25, D.C.; under Grant No. AF-AFOSR-62-321 and its predecessor Contract No. AF 49(638)-586.

REFERENCES

1. McDonald, D., "Non-linear Techniques for Improving Servo Performance." Proceedings of N.E.C., vol. 6 (1950), pp. 400-21.

2. McDonald, D., "Intentional Non-linearization of Servomechanisms," Proceedings of Symposium on Non-linear Circuit Analysis, Polytechnic Institute, Brooklyn, N. Y., vol. 2 (1953), 11 402-11.

3. Mathews, K. C., Bol. R. C., "The Application of Non-linear Techniques to Servomechanisms," Proceedings of N.E.C., vol. 8 (1952), pp. 10-21.

4. Hopkins, A. M., "A Phase Plane Approach to the Compensation of Saturating Servomechanisms," A.I.E.E. Transactions, vol. 70, pt. 1 (1951), pp. 631-39.

5. Buland, R. N., Furumoto, N., "Dual-Mode Relay Servos," A.I.E.E. Transactions, vol. 79, pt. II (1960), pp. 405-11.

6. LaSalle, J. P., "The Time Optimal Control Problem," Contributions to the Theory of Non-linear Oscillations, Princeton University Press, vol. 5 (1960), pp. 1-24.

7. Aoki, M., "Stochastic Time Optimal Control Systems," A.I.E.E. Transactions, vol. 80, pt. 2 (1961), pp. 41-44.

8. Chang, S. S. L., Synthesis of Optimum Control Systems, McGraw-Hill Book Co., New York, 1961.

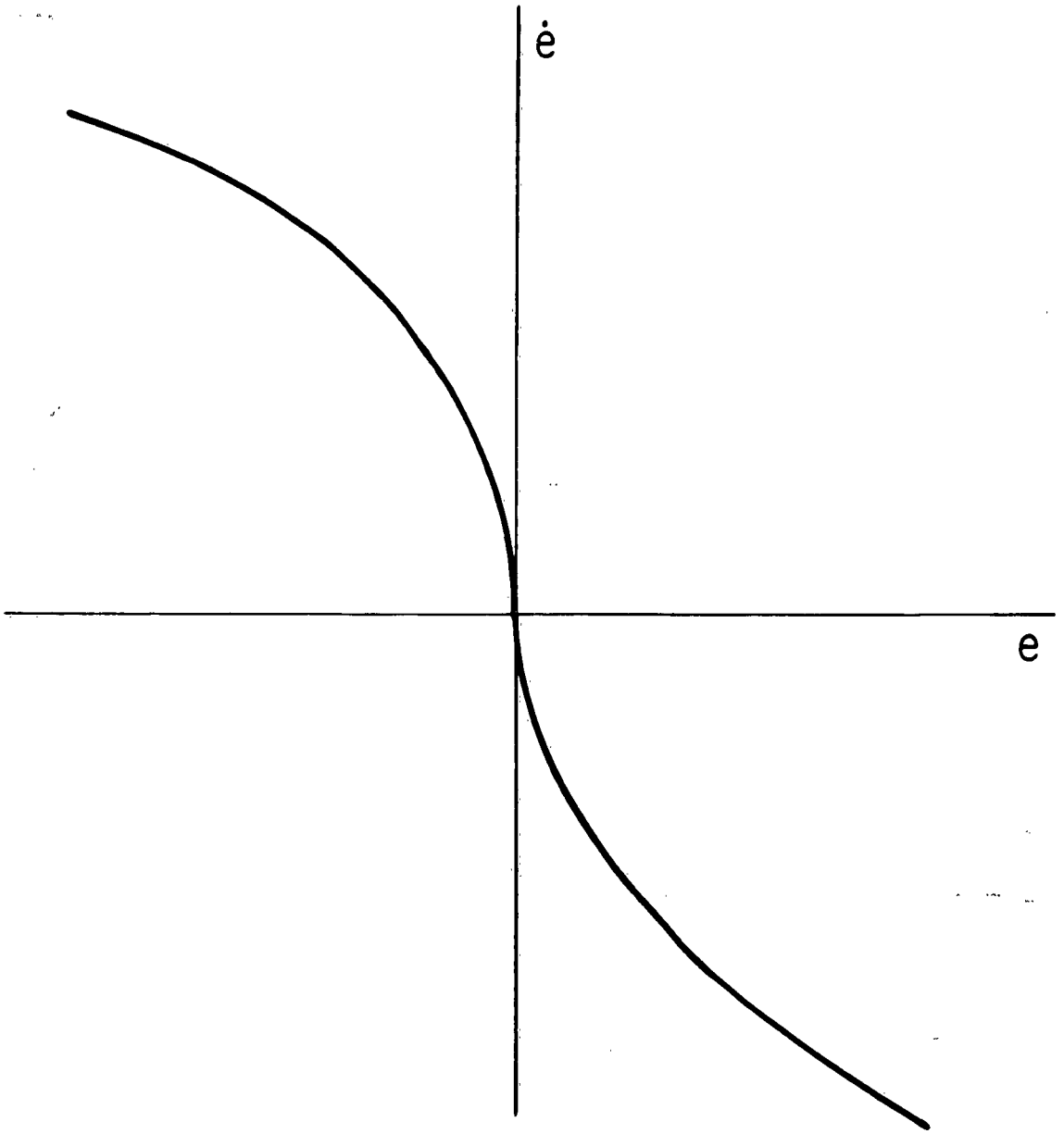


Figure 1. Switching Curve $\frac{2T}{J}e + \dot{e}\,|\dot{e}| = 0$

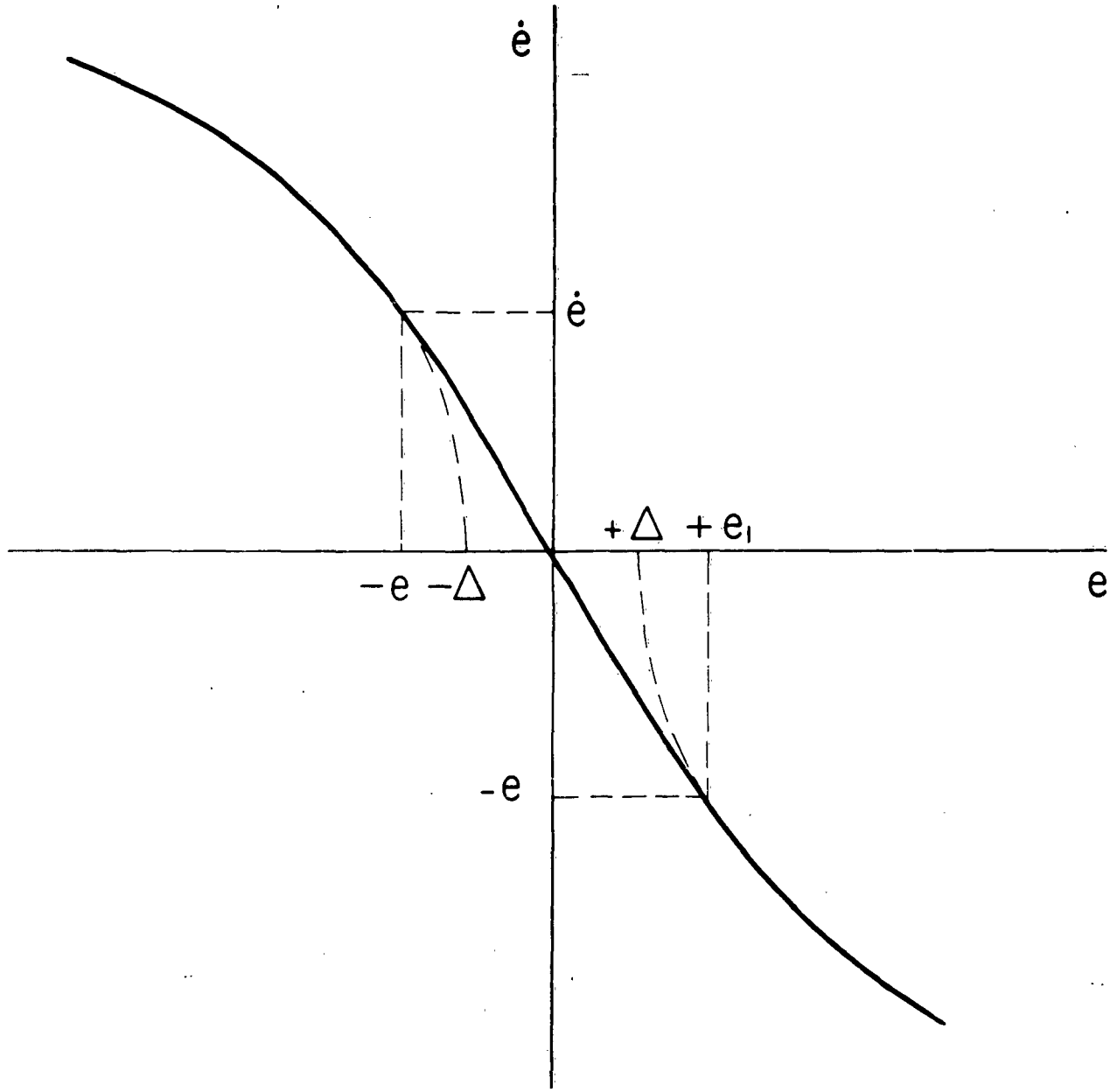


Figure 2. Neutral Curve for Dual Mode System With Finite Gain

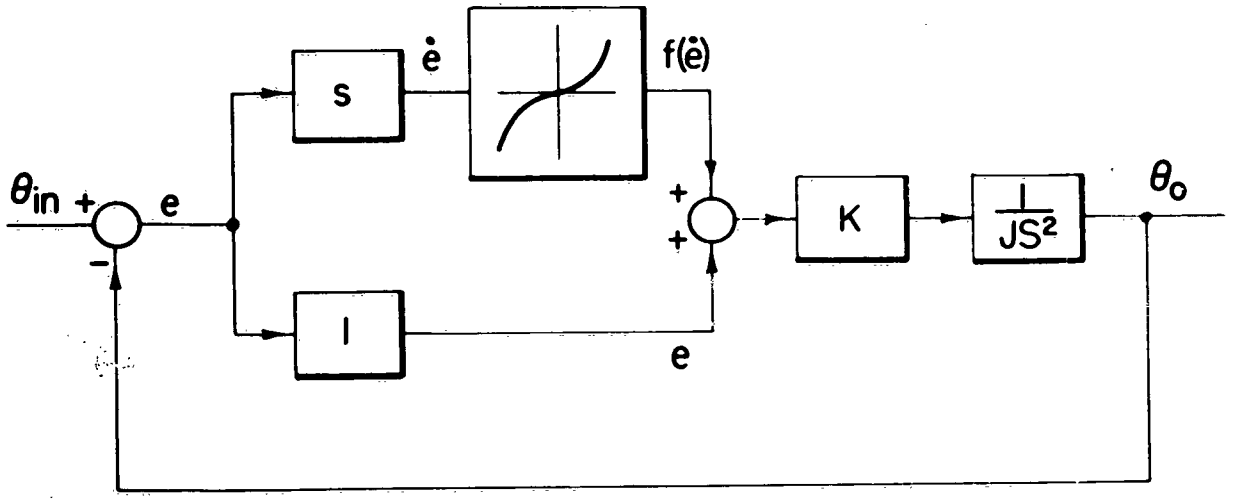


Figure 3. Second Order Automatic Dual Mode System

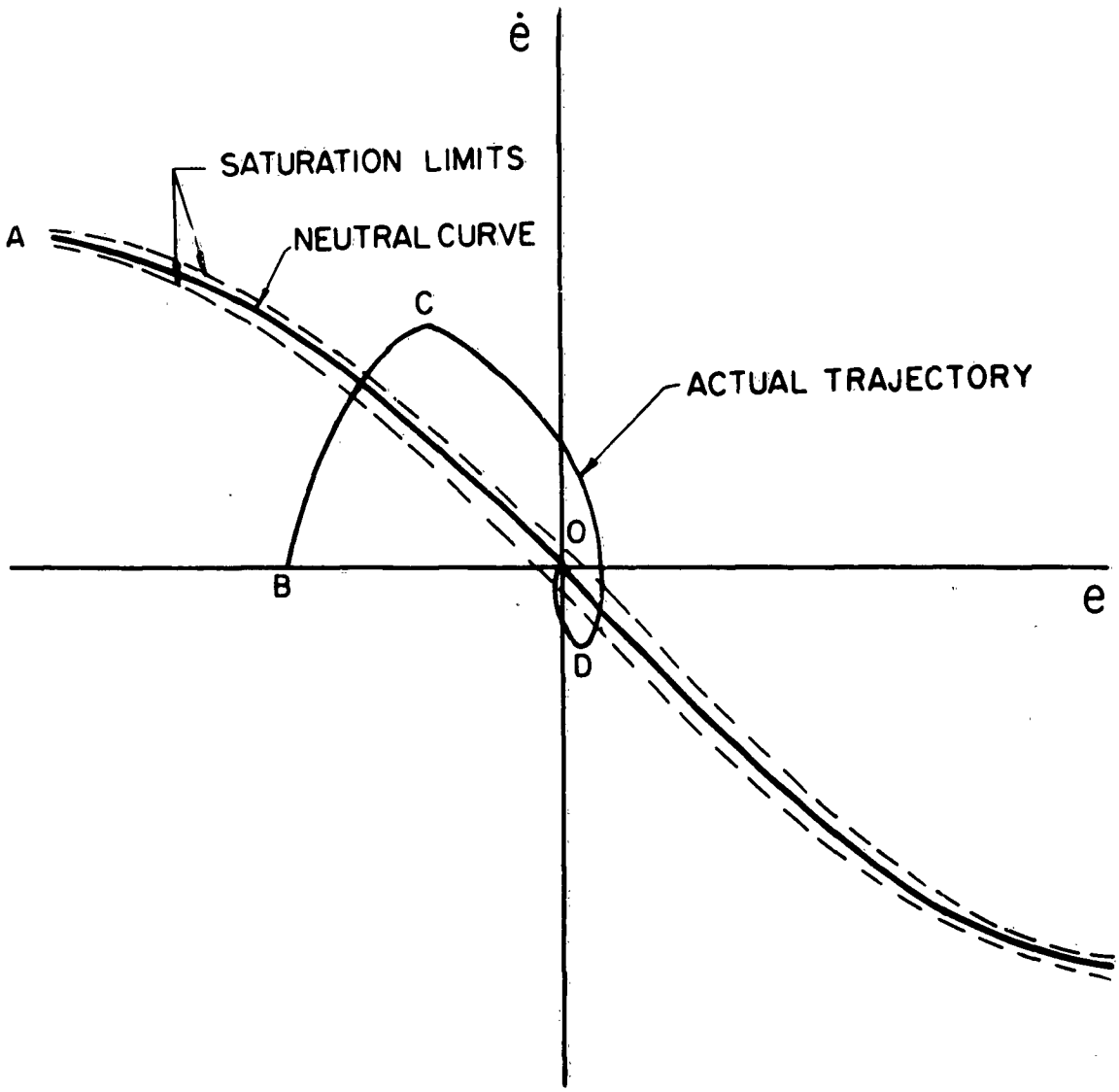


Figure 4. Effect of Minor Time Constants on Phase Plane Trajectory

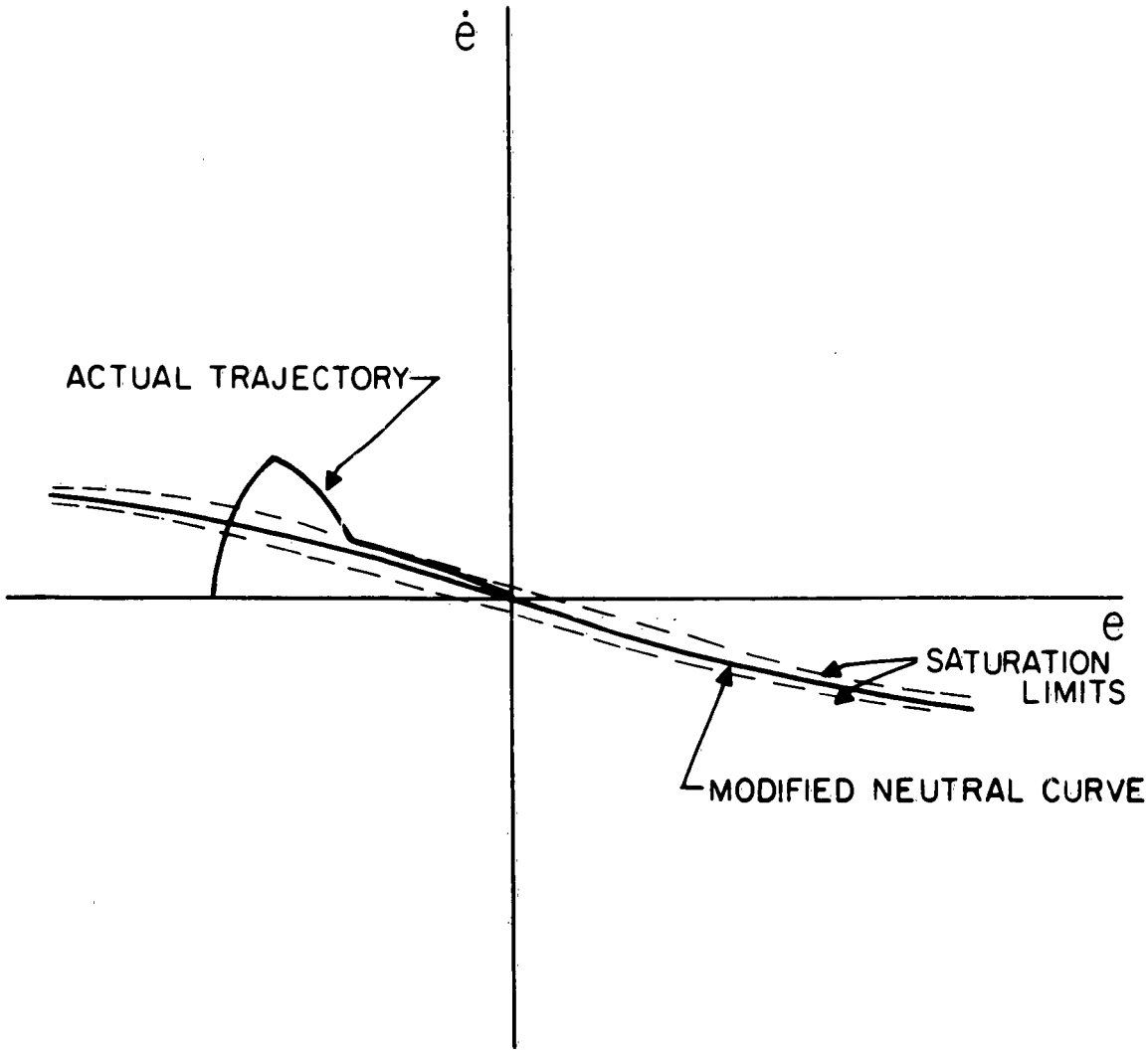


Figure 5. Modified Neutral Curve
(To Obtain "Reserve Torque")

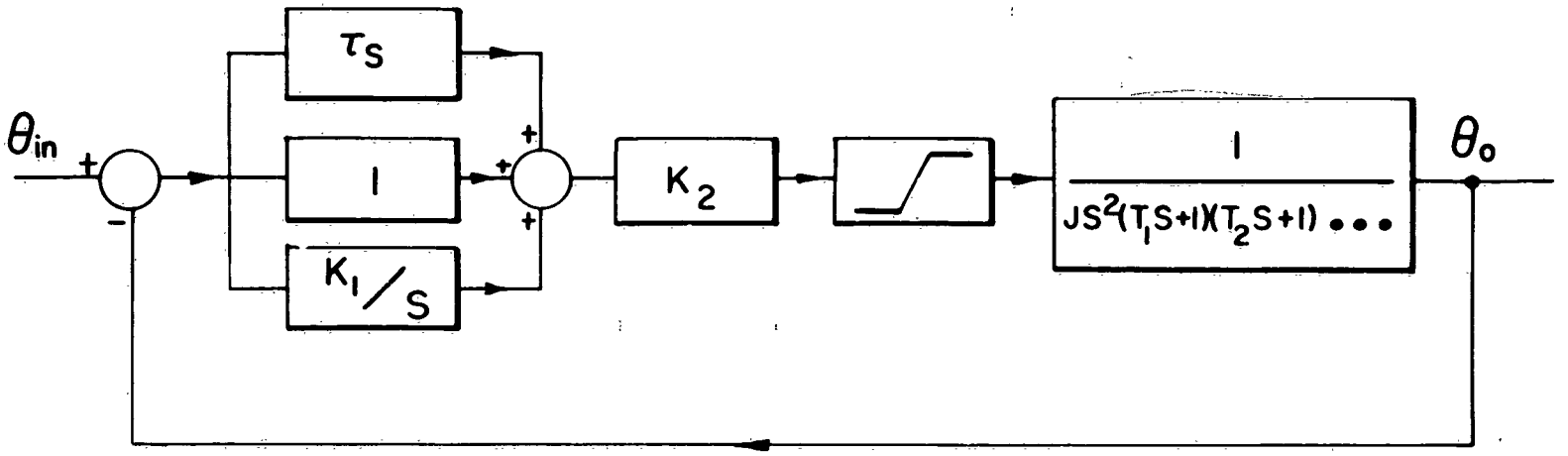


Figure 6. System With Integration Added

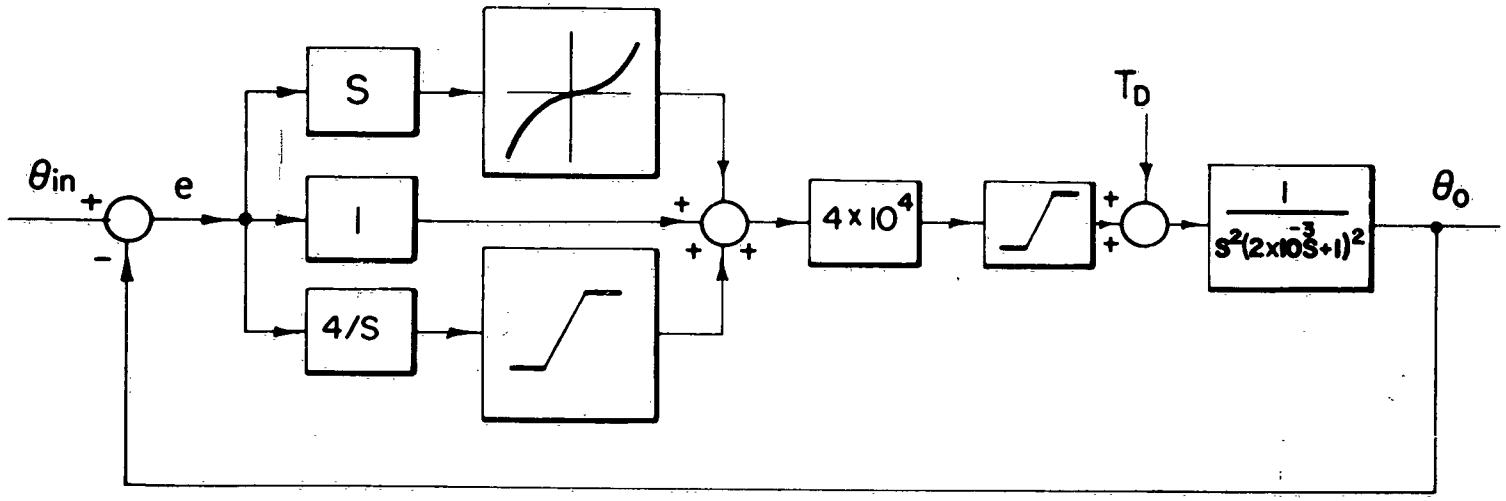


Figure 7. System Used for Analog Computer Study

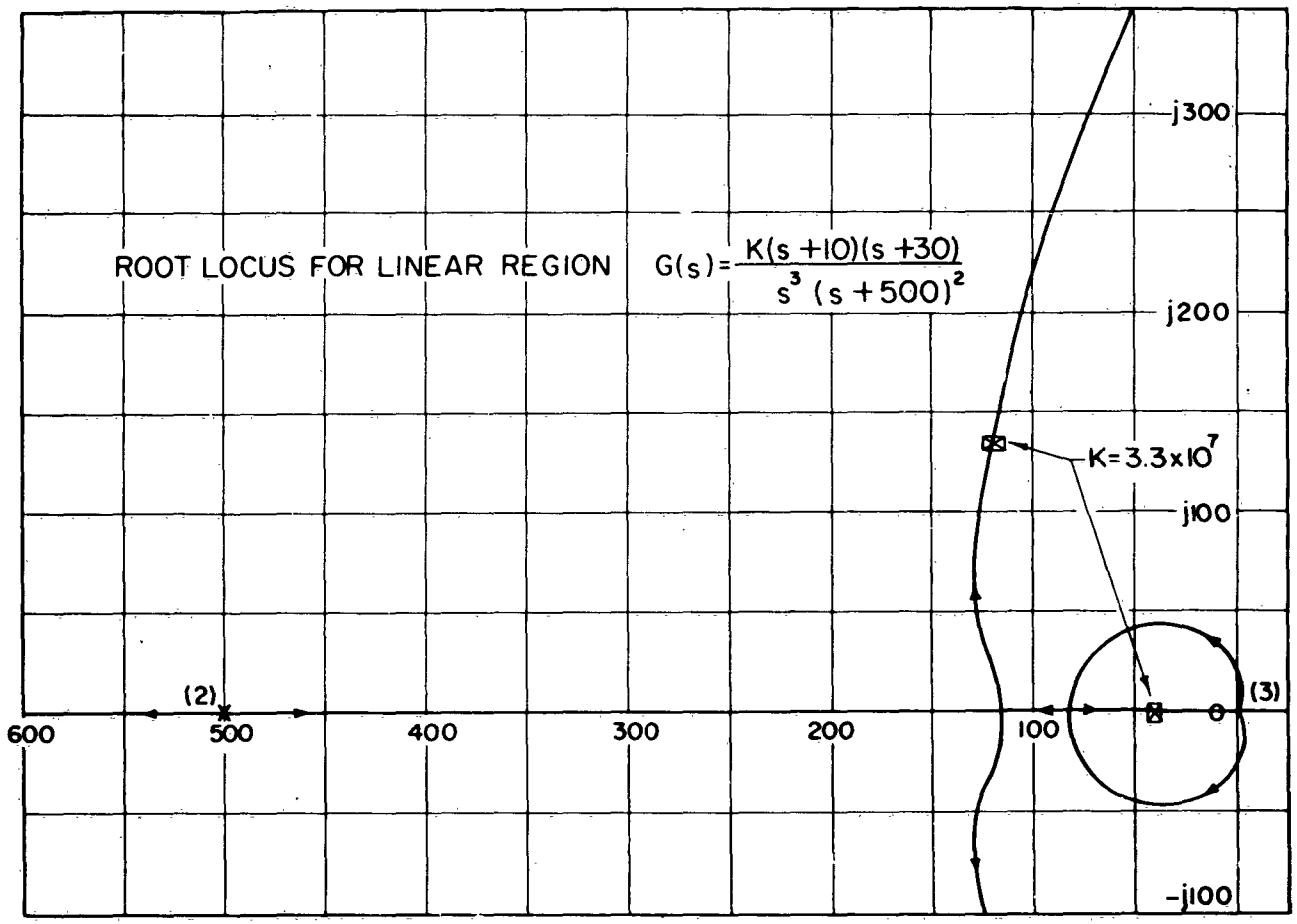


Figure 8. Root Locus for Linear Region of System Used for Analog

Computer Study. $G(s) = \dfrac{K(s + 10)(s + 20)}{s^3 (s + 500)^2}$

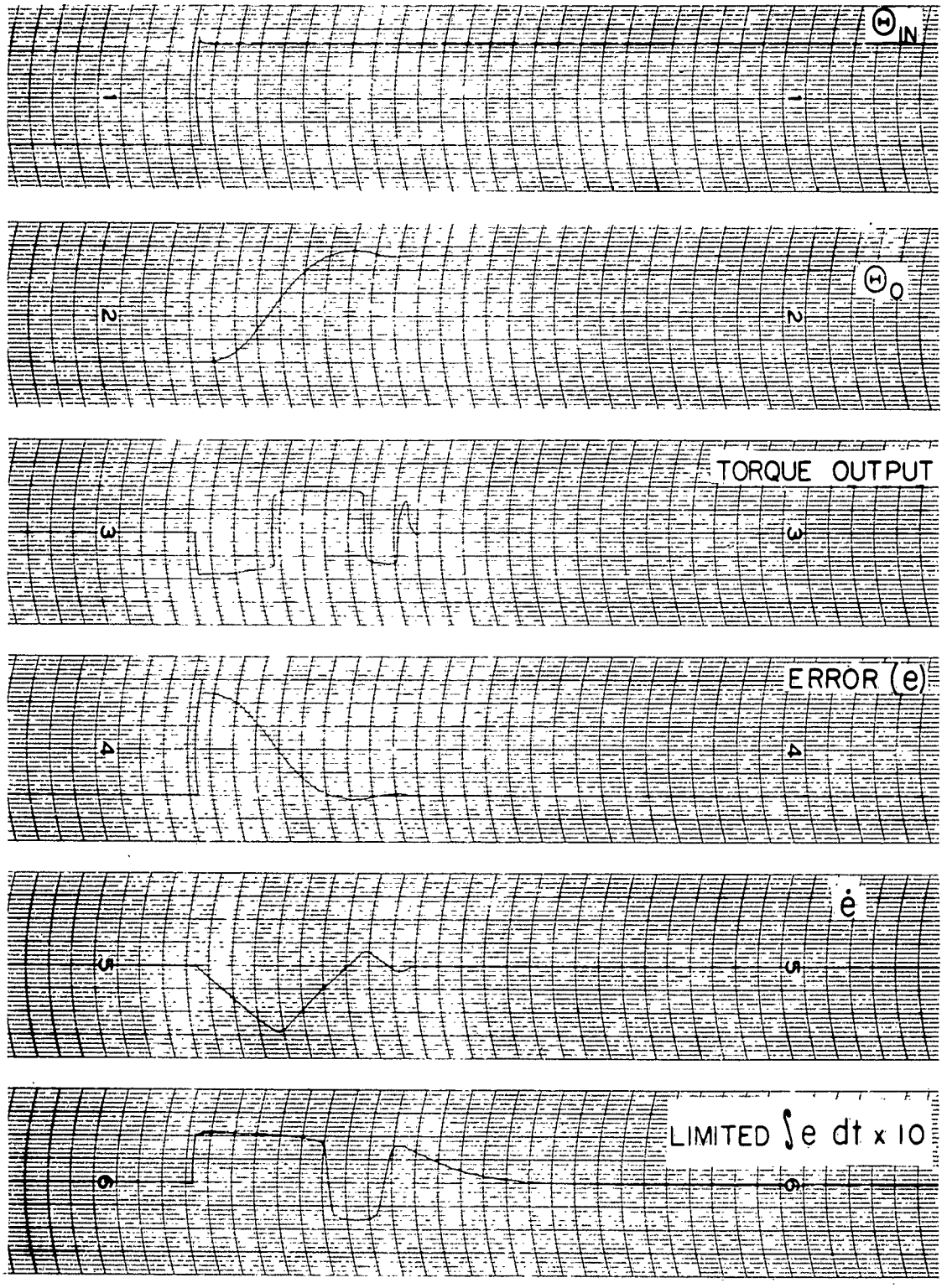


Figure 9. Step Response of System Without Reserve Torque

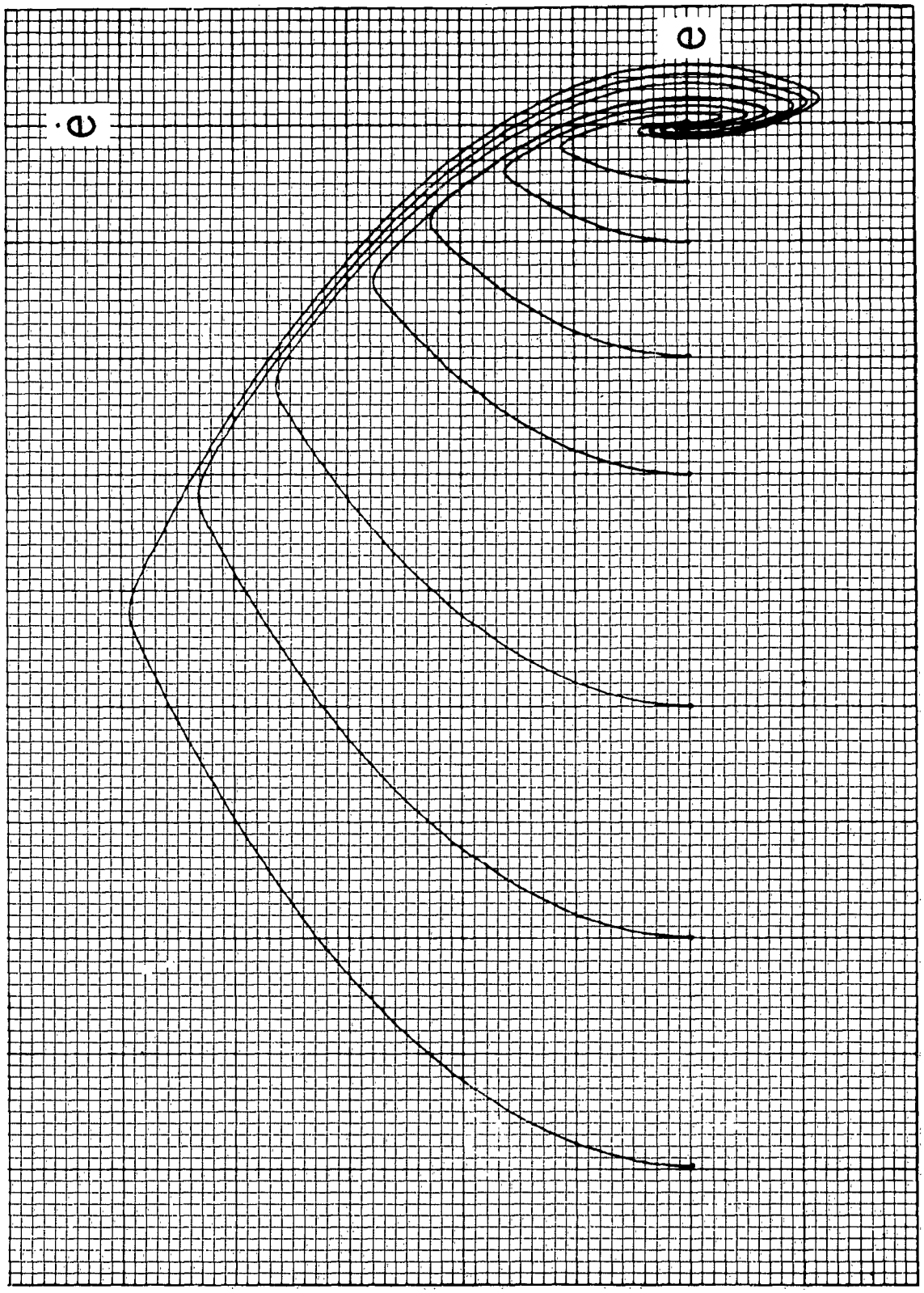


Figure 10. Phase Plane Plot for System Without Reserve Torque

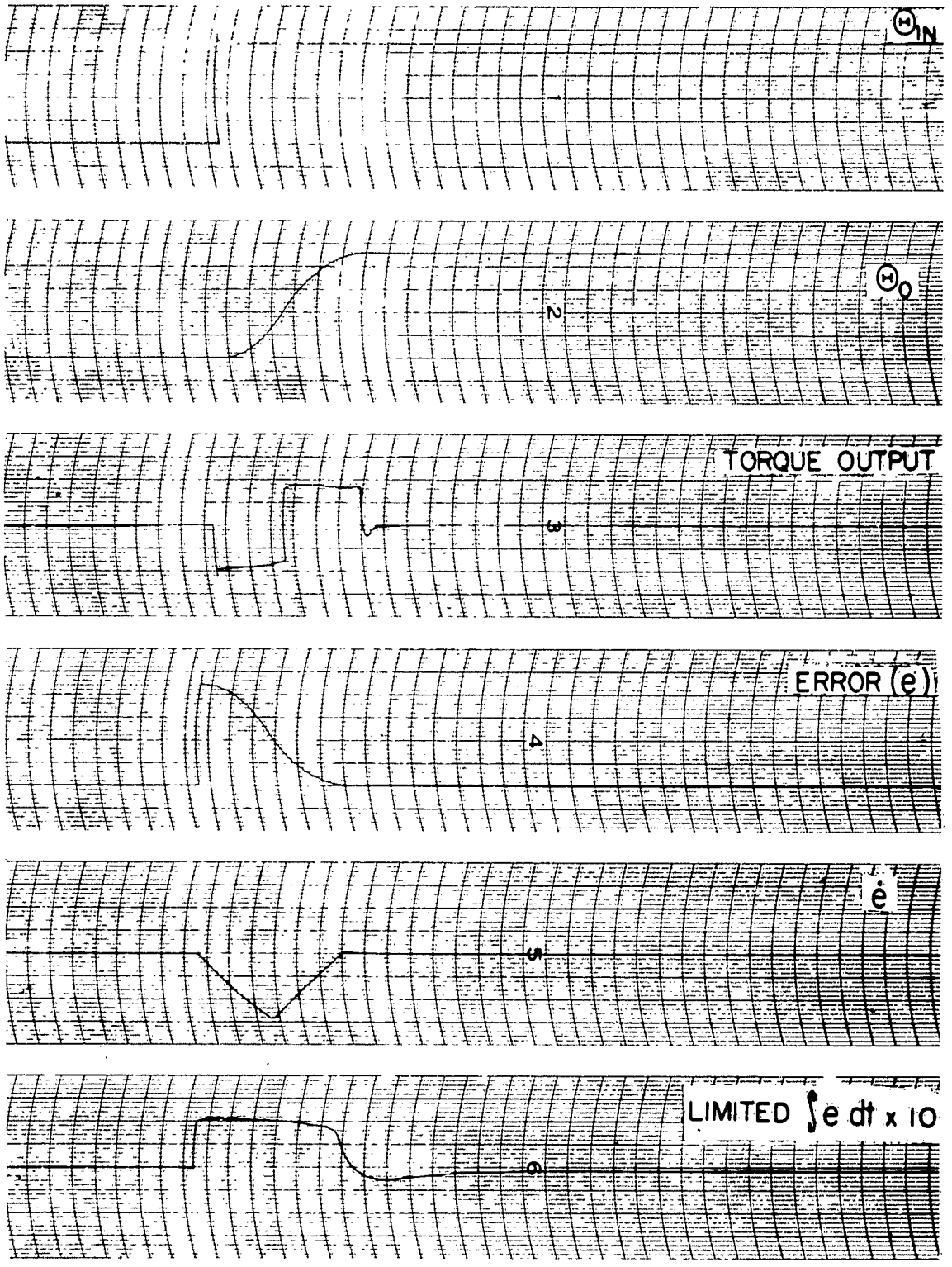


Figure 11. Step Response of System Incorporating Reserve Torque

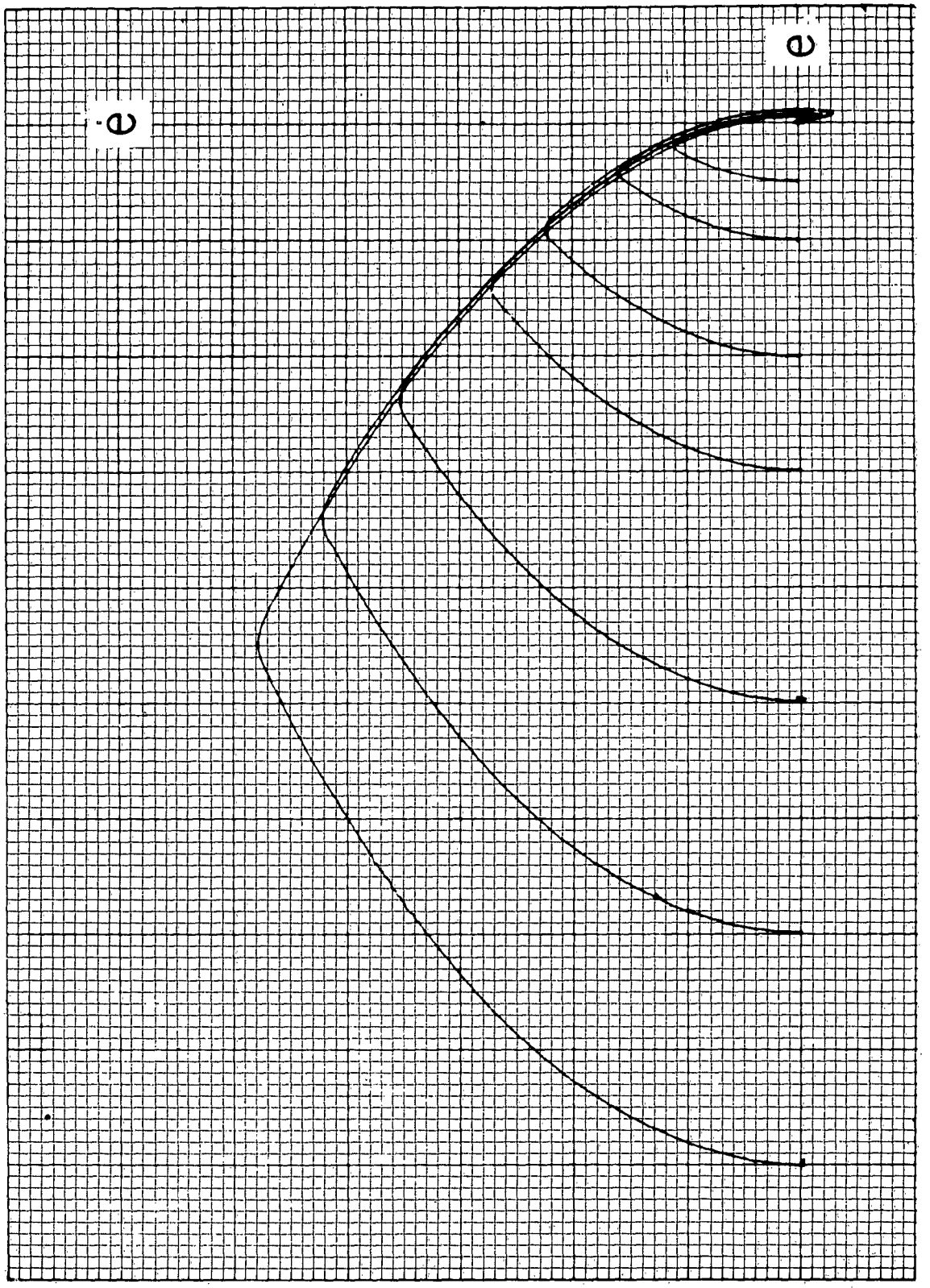
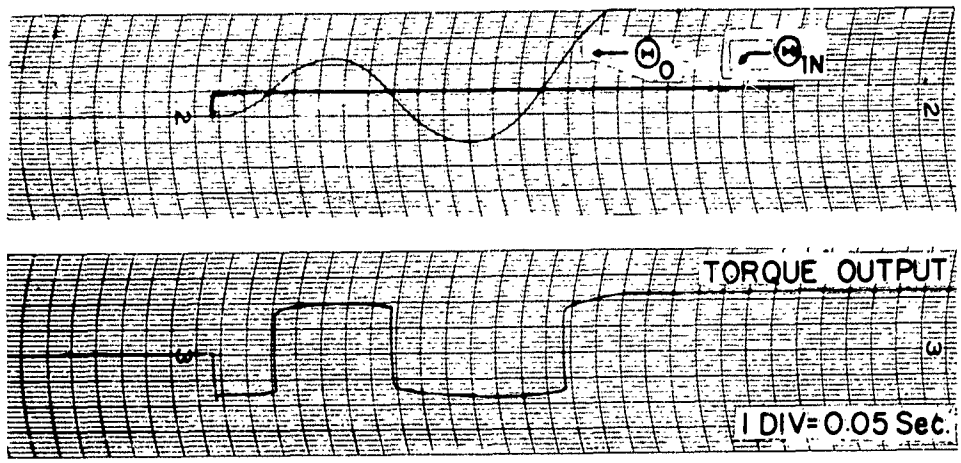
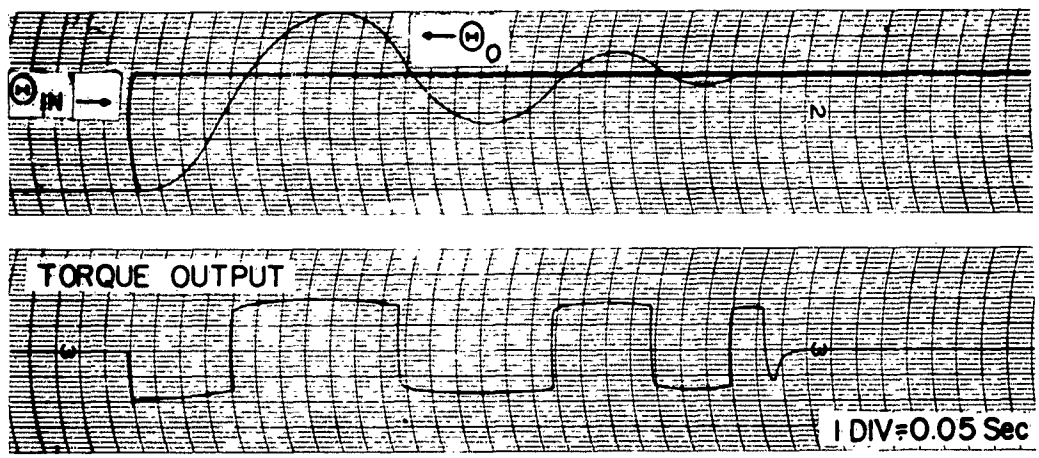


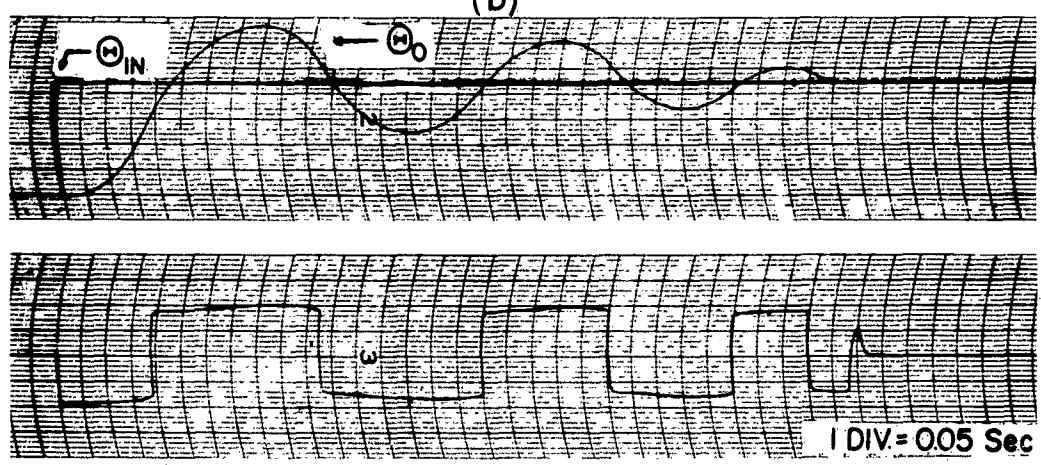
Figure 12. Phase Plane Plot of System Incorporating Reserve Torque



(a)



(b)



(c)

a) Linear Compensation and Integration Not Limited
b) Linear Compensation and Integration Limited
c) Non-Linear Compensation and Integration Not Limited

Figure 13. Step Response

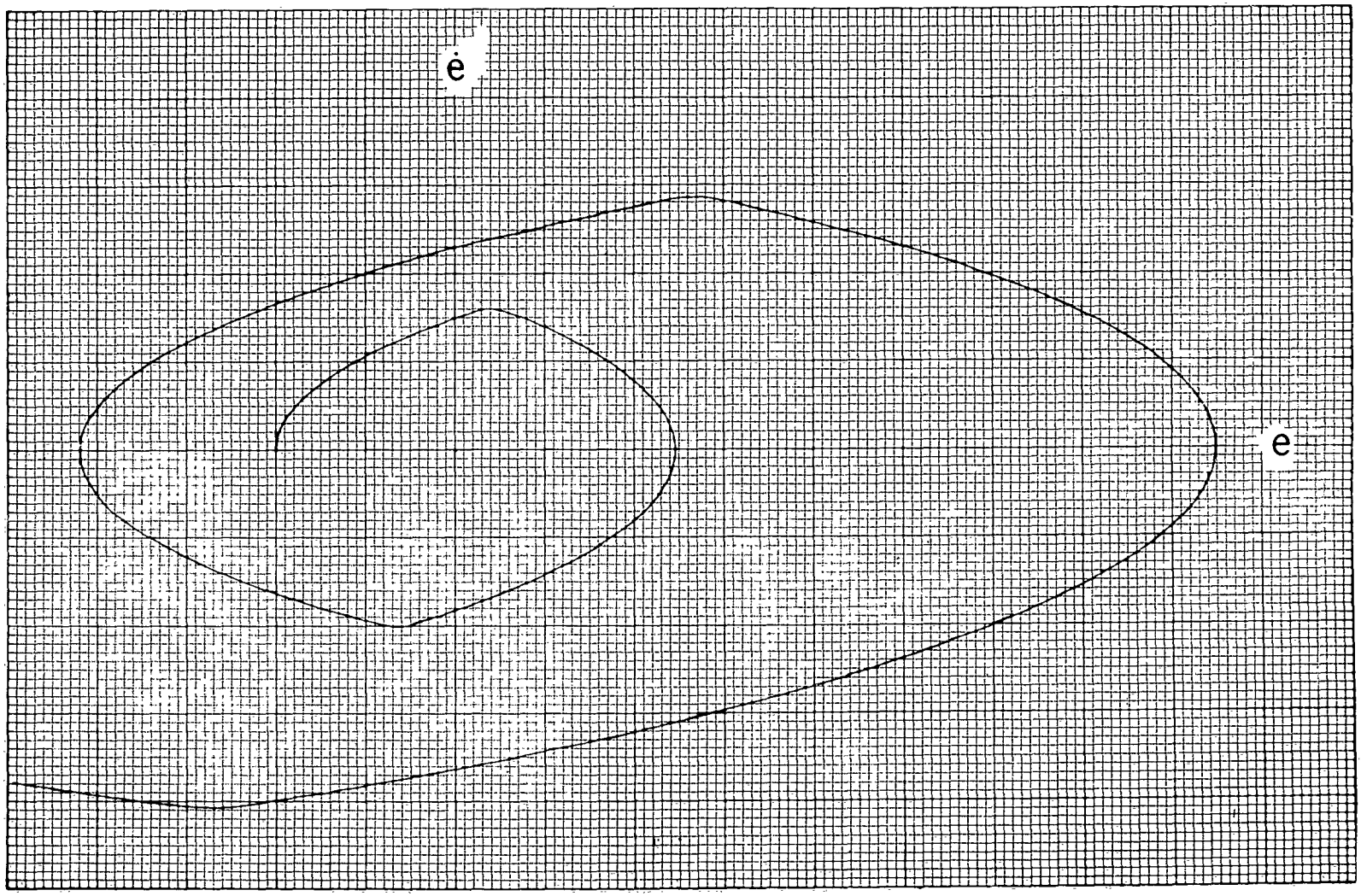


Figure 14. Phase Plane Plot for System With Linear Compensation and Integration Not Limited

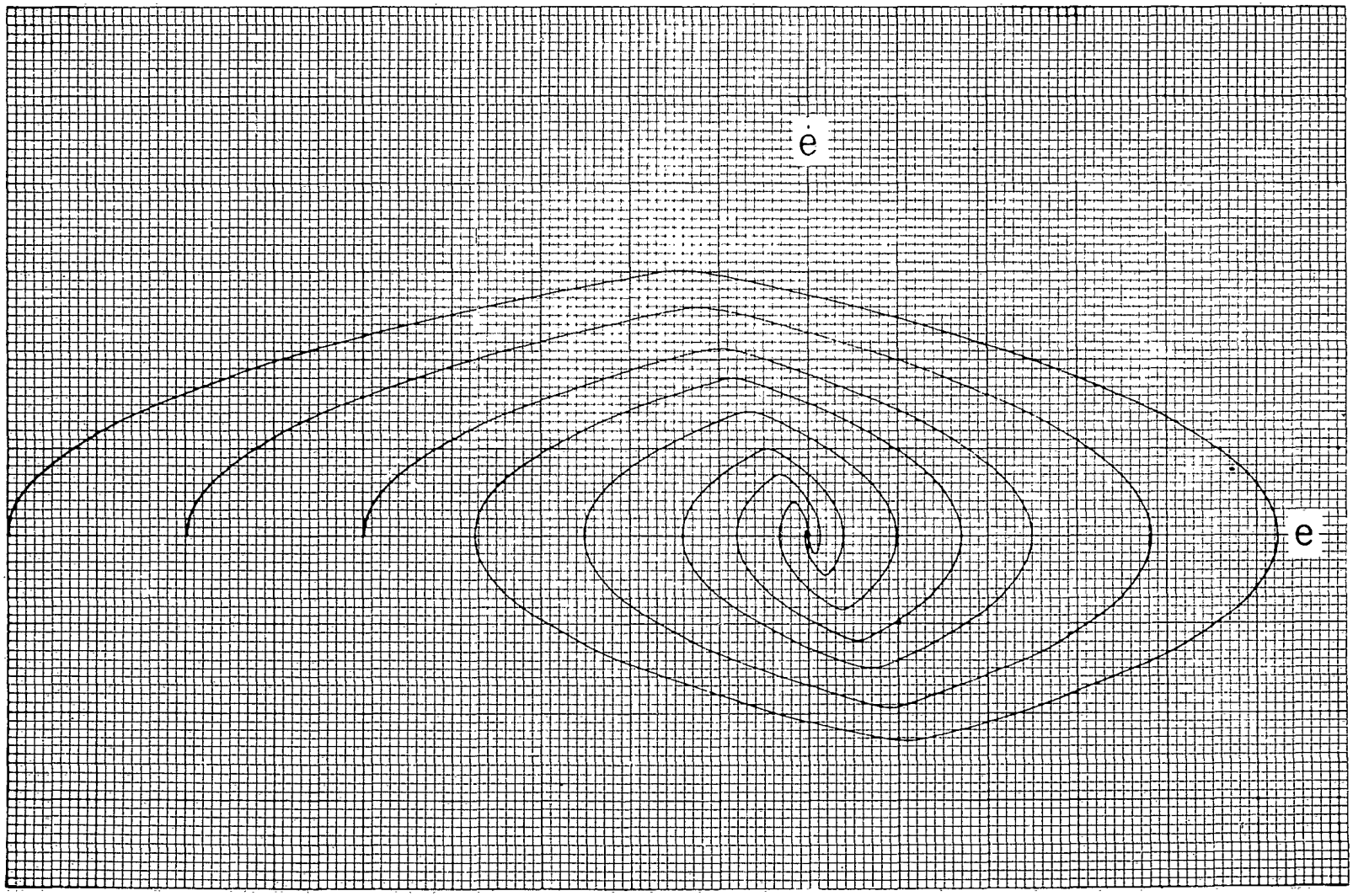


Figure 15. Phase Plane Plot for System With Linear Compensation and Integration Limited

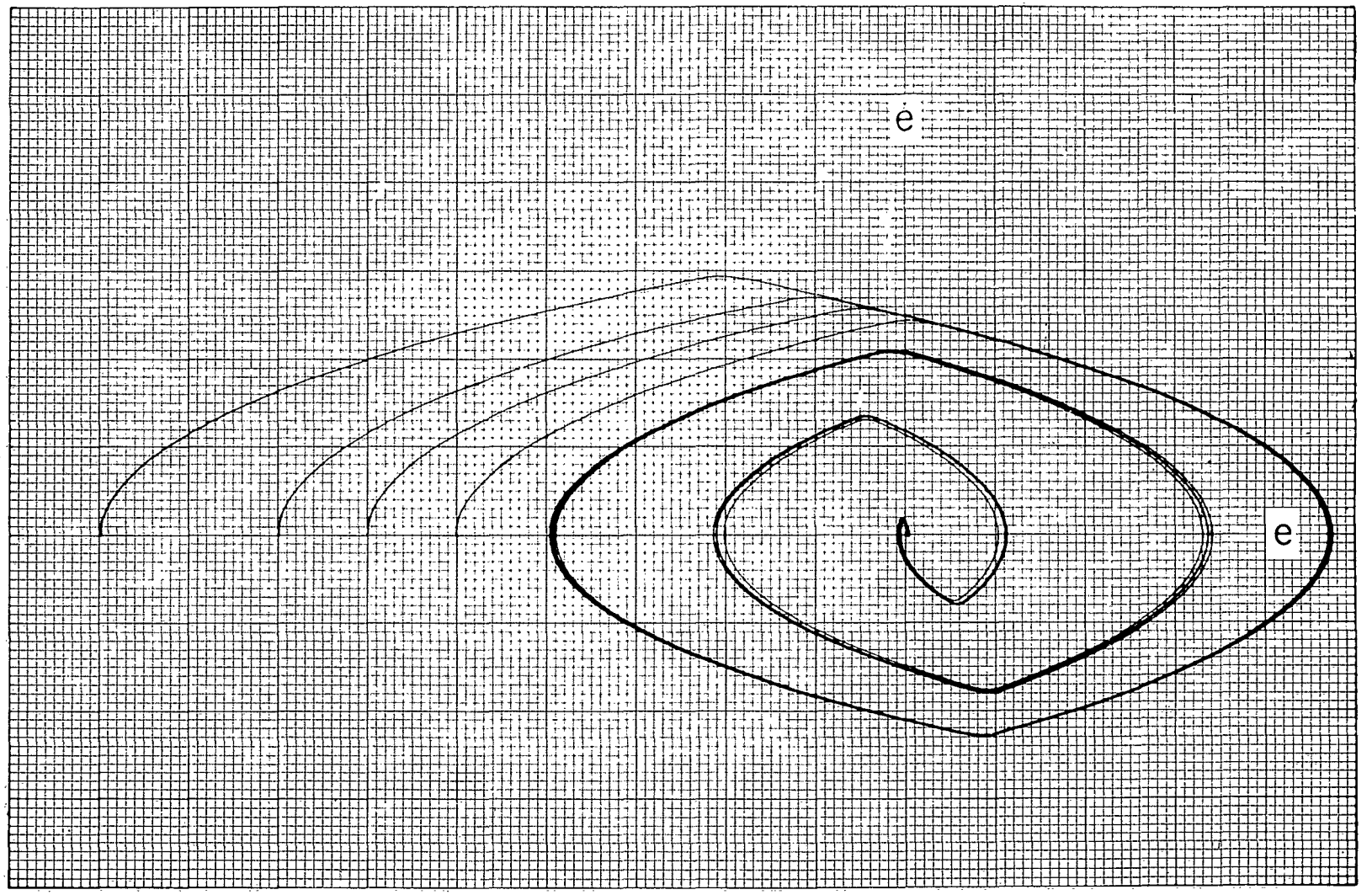
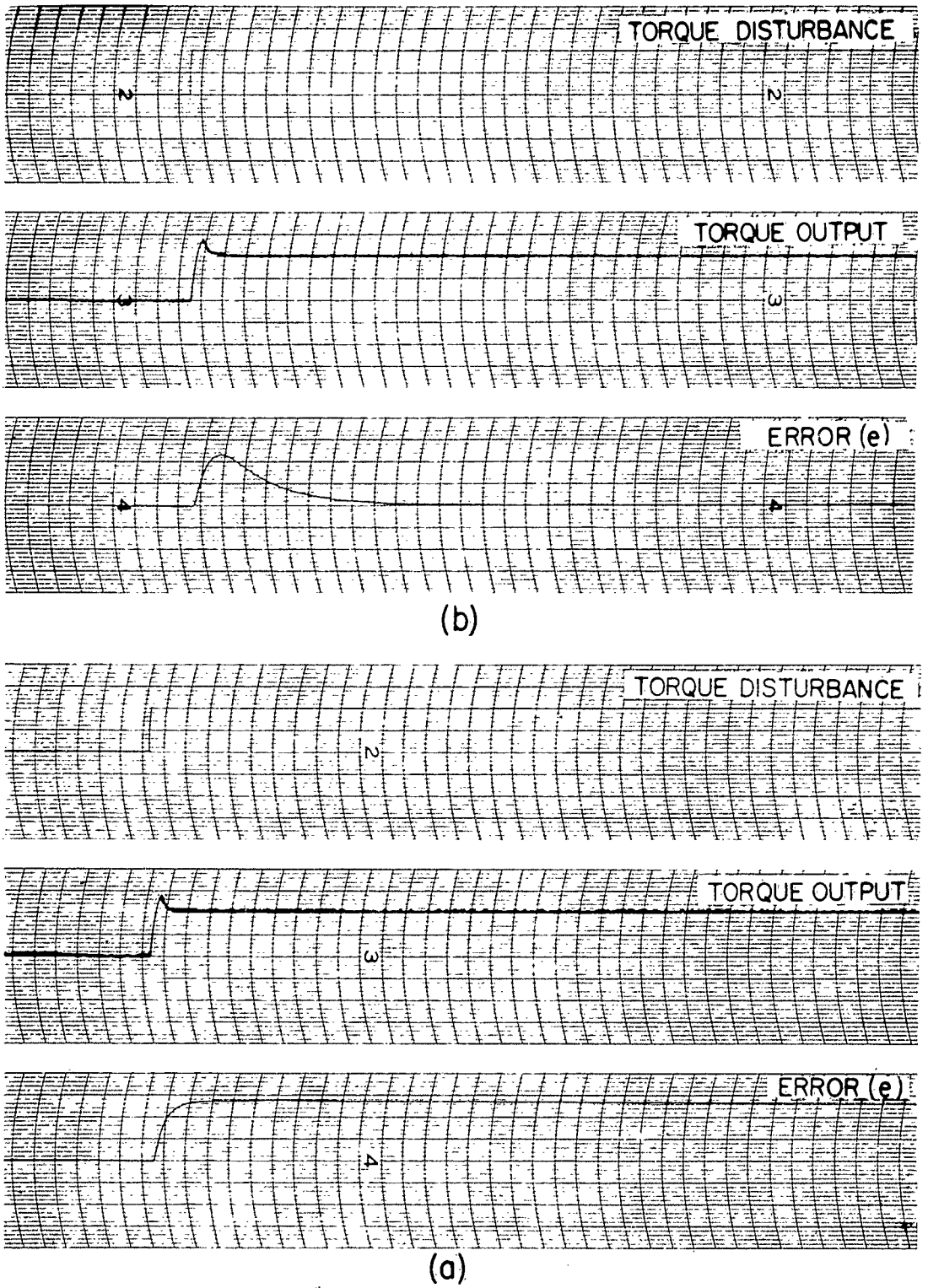


Figure 16. Phase Plane Plot for System With Non-Linear Compensation and Integration Not Limited



a) Integration Present
b) Integration Removed

Figure 17. Response to Step Input of Torque

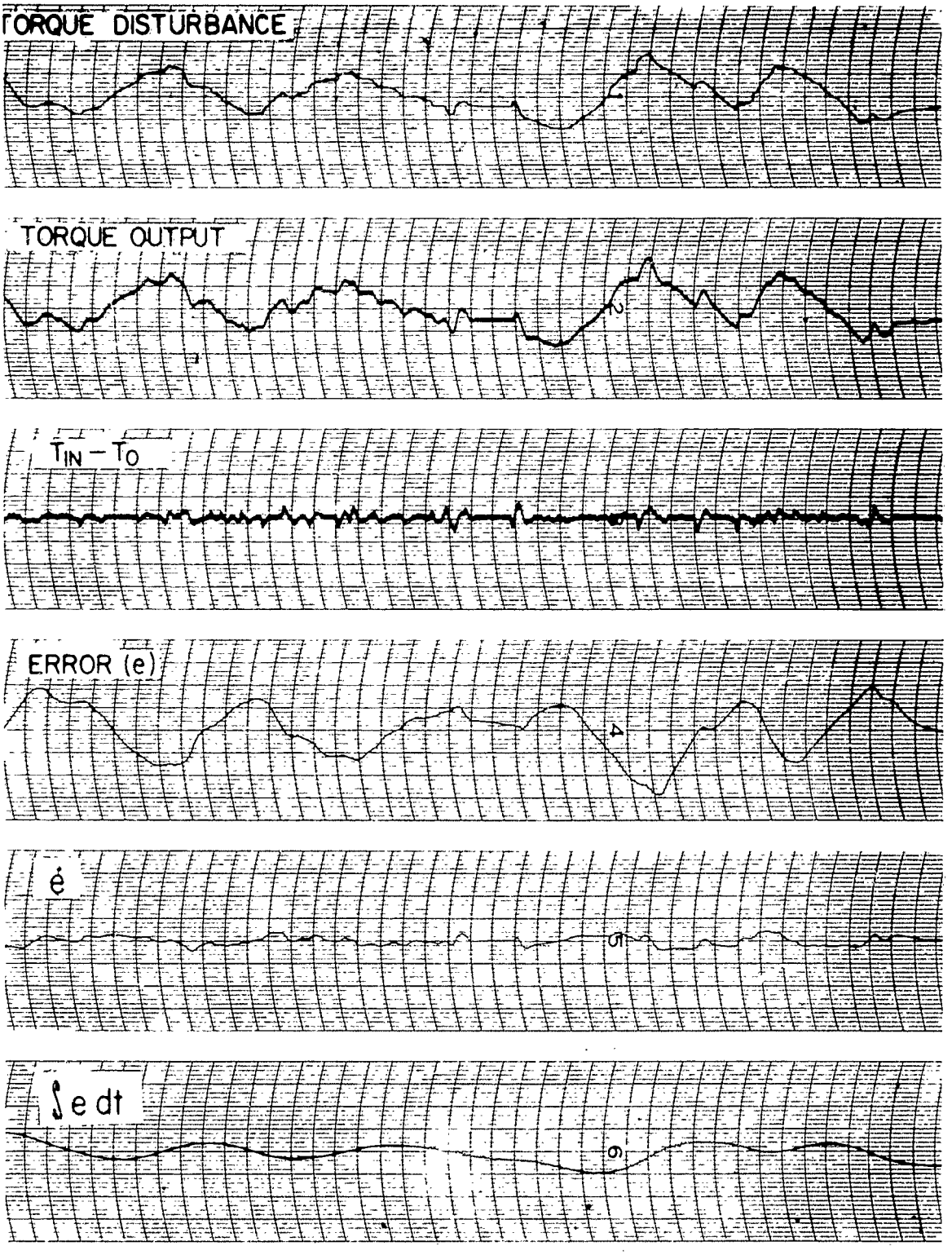


Figure 18. Response to Random Torque Disturbance

ASD-TDR-63-119

Applications and Applicability of Optimum Nonlinear Control

By

R. Oldenburger

Director, Automatic Control Center
Purdue University
Lafayette, Indiana

APPLICATIONS AND APPLICABILITY OF OPTIMUM NONLINEAR CONTROL

Rufus Oldenburger

Director, Automatic Control Center
Purdue University

ABSTRACT

Much attention has been devoted in recent years to the problem of the optimum control of a system where the manipulated variable is subject to a saturation limitation. This is termed optimum nonlinear control. Much of the effort in this area belongs to the minimal time school of Bellman, Pontryagin and others. In 1944 the author began a theoretical and experimental study of optimum nonlinear control where it is desired to produce response that is best in every practical engineering sense. This work is discussed here, namely the theory as far as appears practical for engineering applications, limitations on applicability, and examples of controllers patented and produced on the basis of the theory.

INTRODUCTION

When a hydraulic governor is connected to the throttle, or equivalent member, of a prime mover this throttle is stroked by the servomotor of the governor. The governor output is the piston of the servomotor. This is connected to the throttle. The speed m' of this piston depends on the opening of the control valve metering oil, or other liquid, to the servomotor cylinder. Normally this valve is of the spool type where the position of the valve plunger determines the valve opening. Full opening of this control valve corresponds to the maximum speed of the servomotor piston, in fact one can, for practical purposes, make this piston travel at maximum speed in either direction by moving the valve plunger through full stroke to one end or the other. Servomotors are usually built so that the maximum servomotor piston speed is the same in each direction. For convenience this is assumed here. The problem of optimum nonlinear control is understood to be that of bringing the prime mover to equilibrium speed in the best manner, assuming that it is initially in a nonequilibrium state. Although the prime mover is in the initial nonequilibrium state because of disturbances on the system of which it is a part, it is assumed that the system is undisturbed after it leaves the initial nonequilibrium state until equilibrium is attained. Thus if the load is dropped on an engine under isochronous control, although the engine is at equilibrium speed, it experiences an initial acceleration proportional to the amount of load dropped. See Fig. 1. It is now desired to return the system to equilibrium with the maximum error ϵ, the duration θ of the transient, the area between the error curve and the time axis, and other numerical quantities minimized simultaneously. In Fig. 1 the quantity c is the error in the controlled variable, i.e. engine rpm (revolutions per minute) error, and t is time. In 1944

the writer analyzed the Woodward governors on Hamilton Standard propellers. Here the aircraft engine speed is regulated by varying the propeller pitch. This pitch corresponds to the position of the piston of the hydraulic servomotor built into the hub of the propeller. The maximum piston speed, and hence rate of change of propeller pitch, is limited to a value depending on pilot comfort and other factors. The writer solved the problem of optimum nonlinear control of such engine-propeller units, and subsequently carried out the theory for the control of one variable as far as he felt practical considerations would permit.

The following questions arise.

1. Is there always a unique transient, optimum in every reasonable sense?

2. If there is a best transient can it be obtained by operating the rate m' only at its maximum numerical value or zero?

3. If there is an optimum transient does there exist a function, or set of functions, of the controlled variable c and quantities obtained by mathematical operations on c so that making the rate m' depend only on this function, or these functions, the optimum transient will be obtained automatically?

THEORY

Consider the equations

$$c' + \alpha c = K_1 q \tag{1}$$

$$T q' + q = m \tag{2}$$

where K_1, α and T are positive numbers, while c' and q' are the derivatives dc/dt and dq/dt of c and q respectively with respect to time t. The quantity α need not be a constant but may be a function of time t. Physically q may be taken to be engine torque or proportional to it, c' as rpm/sec, c as rpm, αc as damping torque, and m as throttle-torque time constant. The author [1] found early in his studies that the damping term αc may be neglected for the magnitudes of α that occur in engineering practise, as far as optimum nonlinear control is concerned. This results in a great reduction in the mathematical complexity of the problem, without adversely affecting the results.

The saturation limit on the derivative m' of the manipulated variable m is equivalent to

$$|m'| \leq K_2 \tag{3}$$

for a positive number K_2 and the absolute value $|m'|$ of m'.

The answer to the three questions above is "yes" for the case where $T = 0$. This is also true of the initial conditions likely to occur in practise in the case where $T \neq 0$. Some initial conditions exist in this last case where optimum transients (minimum overswing and underswing, duration, etc.) can be attained only by operating $|m'|$ below saturation during at least part of the transient, i.e.

$$|m'| < K_2 \tag{4}$$

It is particularly important that the answer to the third question be "yes". For the automatic control of the variable c the controller usually must operate on c only and from this information vary m to bring the system to equilibrium. It is thus not convenient to have the variable m depend on time t explicitly.

We define the <u>absquare</u> $\{x\}$ of a variable x by

$$\{x\} = |x|x \tag{5}$$

Thus $\{x\}$ is the <u>signed square</u> of x. This function comes up automatically in optimum nonlinear control where it plays a significant role.

In as yet unpublished work the writer proved that optimum nonlinear control of the system described by relations (1), (2) and (3) with $\alpha = 0$ is obtained by letting

$$m' = -K\Sigma \tag{6}$$

where K is arbitrarily large ($+\infty$) and Σ is the control function given by

$$\Sigma = \psi + \frac{\{\psi'\}}{2K_1K_2} - (\operatorname{sgn}\psi')K_1K_2T^2\ln^2\left(1 + \sqrt{1 - (1 + [\operatorname{sgn}\psi']\frac{c''}{K_1K_2})e^{-|\psi'|/TK_1K_2}}\right) \tag{7}$$

where

$$\psi = c + Tc' \tag{8}$$

and

$$\operatorname{sgn}\psi' = 1, 0, -1 \tag{9}$$

according as

$$\psi' > 0, \qquad \psi' = 0, \qquad \text{or} \qquad \psi' < 0 \tag{10}$$

When the log term in Σ is imaginary it is to be dropped.

In view of the limit (3)

$$\begin{aligned} m' &= -K_2 \quad \text{when} \quad \Sigma > 0 \\ &= K_2 \quad " \quad \Sigma < 0 \\ &= 0 \quad " \quad \Sigma = 0 \end{aligned} \tag{11}$$

For a load rejection the resulting transient is as shown in Fig. 2. Similar curves apply for other initial conditions. The quantities t_1, t_2 and t_3 denote switching instants at which in practise Σ changes sign. For $t > t_3$ we have $\Sigma = 0$ and the system is in equilibrium with $m' = 0$. From $t = 0$ to t_1 we have $\Sigma > 0$ whence $m' = -K_2$. From $t = t_1$ to $t = t_2$ we have $\Sigma < 0$ whence $m' = K_2$. Similarly from $t = t_2$ to $t = t_3$ we have $\Sigma > 0$ and $m' = -K_2$. After $t = t_3$ the relation $\Sigma \equiv 0$ holds.

If the lag T is zero, the times t_2 and t_3 coincide. In any case every other solution lies above the curve of Fig. 2, and is thus worse in every sense, or this solution crosses the curve of Fig. 2 after the point where $t = t_1$, or leaves this curve plunging below it. In the latter case the curve reaches equilibrium after $t = t_3$ so that the duration of the transient is greater and this curve is worse than that of Fig. 2 in every reasonable engineering sense.

If the system to be controlled is of higher order than that of equations (1) and (2) the control functions involve the third and higher order derivatives of c. Since it is not possible to obtain such derivatives in the physical world the treatment of such high order systems is largely academic.

If a system has two or more dominant lags, as when T splits up into two time constants T_1 and T_2 with $T_1 = T_2$ the improvement that can be obtained by going to optimum nonlinear control is negligible, and in the opinion of the writer not economically justifiable. In any case theory and experiment indicate that it is advisable to lump the lags of a system (the lag due to the damping term α being neglected).

To obtain stability it is convenient to have a linear band of operation near equilibrium. It is advisable to use the control function

$$m' = -K\left[c + (\gamma + \beta|c'|)\, c'\right] \tag{12}$$

for positive constants β and K. The value of γ is chosen to give good system stability in the absence of large disturbances. Then β should be adjusted so that $\beta|c'|_{max}$ dominates γ for the maximum value $|c'|_{max}$ of $|c'|$ to be encountered in practise. Practical experience shows that it is desirable to have

$$10\gamma \leq \beta|c'|_{max} \leq 25\gamma \tag{13}$$

A curve as in Fig. 1 was obtained in the laboratory for a case where the system lags have time constants of 0.4 sec., 0.1 sec., 0.1 sec., and 0.05 sec. respectively. A very great variety of physical systems and initial conditions were tested in the laboratory by the author and his associates over a period of several years, while hardware was developed and patents obtained to exploit the results.

A reasonable compromise of the precise theory may be attained by letting

$$\Sigma = \psi + \frac{|\psi|_{av}\,\psi'}{2K_1K_2} \tag{14}$$

for an average value $|\psi|_{av}$ of $|\psi'|$ for the responses to larger disturbances, so that the nonlinear control function Σ of formula (7) is replaced by the linear function Σ of relation (14).

Tests showed that in practise one can always adjust K and γ in Σ of formula (12) with $\beta = 0$ so that the maximum error for the worst sudden disturbance encountered, such as a full load change on an engine, is the same as when optimum nonlinear control is used. Also, the curves are practically the same in their entirety. If the magnitude of the disturbance, such as load, is reduced, we can reach a point where the error is half as large when a nonlinear control function Σ is employed. In any case we can halve the area between the c curve and the t axis. See Fig. 3 where the two peaks are for identical partial load changes on the same diesel engine but the control is linear or nonlinear as noted. Here best linear and nonlinear control are compared where for larger disturbances these controls yield practically the same results. If we go to smaller disturbances the responses for the linear and nonlinear Σ become identical for practical purposes.

The extravagant claims of improvement by going from linear to nonlinear control by others is based on comparing optimum nonlinear results with certain linear controls, but not the best that can be obtained by the use of linear control functions. How then can the great amount of research on optimum nonlinear control be justified? One reason is that with the optimum nonlinear control theory on hand the engineer can tell how far an actual design deviates from the best, and he can modify the design accordingly. Also the disturbances that occur most frequently may be the intermediate ones, where a two to one improvement with nonlinear control is attained. A customer will often pay for this magnitude of improvement, especially in military and space applications. Finally, the nonlinear control equipment is sometimes less expensive than the corresponding linear device.

TWO OR MORE CONTROLLED VARIABLES

The work of the author and his associates [1], [2] has been primarily restricted to the control of one variable. Early work on bang-bang control by D. McDonald [3], A. M. Hopkin [4] and others was also restricted to one controlled variable. The approach of the writer is being extended to two and more variables. Recently there has arisen the "minimal time" school of Bellman [5], Kalman [6], Pontryagin [7] and others involving a very large number of workers in the United States, the Soviet Union and elsewhere. This school is concerned with systems in one or more controlled variables with manipulated variables subject to saturation limits, where it is desired to bring the system from an initial state to another state in such a way as to minimize the duration of the transient, or minimize (maximize) some other functional of time and the controlled and manipulated variables. The theory does not apply to minimizing the maximum error. This is the performance index one is normally most concerned with in practise. Thus the user wishes to know whether the

control will be accurate to 1/10%, 1% or some other value. He usually does not care about minimizing response time. Actually, the user looks at the entire response curves in determining the goodness of a control. The situation is analogous to the determination of beauty queens. It is not enough to make a finite number of numerical measurements and base the selection on these only, such as on the dimensions 35-25-35. A response curve has infinitely many numerical properties. The approach of the writer differs from the minimal time school in that he looks at the complete system responses rather than some numerical properties of these responses, selected for mathematical or other convenience. On the other hand, the great value of the minimal time school approach is that by minimizing time one also often minimizes every other reasonable measure of goodness simultaneously. On the other hand there are sometimes infinitely many distinct means for bringing the given system from the initial state to the final state in minimal time, and additional theory must be employed to select from these and perhaps other solutions one that is best in every reasonable engineering sense.

Consider the system described by the equations

$$\begin{aligned} c_1' &= m_1 + m_2 \\ c_2' &= m_1 - m_2 \end{aligned} \tag{15}$$

for controlled variables c_1 and c_2 with time derivatives c_1' and c_2' respectively, and manipulated variables m_1 and m_2 where the magnitudes $|m_1'|$, $|m_2'|$ of the derivatives m_1', m_2' satisfy

$$\begin{aligned} |m_1'| &\leq 1 \\ |m_2'| &\leq 1 \end{aligned} \tag{16}$$

Let

$$\begin{aligned} C_1 &= c_1 + c_2 \\ C_2 &= c_1 - c_2 \end{aligned} \tag{17}$$

whence equations (15) are equivalent to

$$\begin{aligned} C_1' &= 2 m_1 \\ C_2' &= 2 m_2 \end{aligned} \tag{18}$$

We wish to bring the system from an initial state

$$(c_1, c_2, c_1', c_2') = (c_{10}, c_{20}, c_{10}', c_{20}') \tag{19}$$

for real numbers c_{10}, c_{20}, c'_{10}, c'_{20} to the equilibrium state

$$(c_1, c_2, c'_1, c'_2) = (0, 0, 0, 0) \tag{20}$$

in the best manner with m_1 and m_2 subject to the saturation condition (16), it being assumed that the initial state is not equilibrium. If the variables c_1 and c_2 are brought to equilibrium ($c_1 \equiv 0$, $c_2 \equiv 0$) in a minimal time M it follows that C_1 and C_2 are brought to the equilibrium ($C_1 \equiv 0$, $C_2 \equiv 0$) in this time M. To simplify the argument suppose that $c_{10} = c_{20} = 0$, whence $C_{10} = C_{20} = 0$.

The first equation (18) with the first condition (16) describes a single controlled variable system. For this the optimum transient is as shown in Fig. 4a where the duration M_1 of the transient is given by

$$M_1 = \frac{(1 + \sqrt{2})\, c'_{10}}{2}, \qquad C'_{10} = c'_{10} + c'_{20} \tag{21}$$

For the variable C_2 there is a second optimum transient shown in Fig. 6b where the transient duration M_2 is given by

$$M_2 = \frac{(1 + \sqrt{2})\, C'_{20}}{2}, \qquad C'_{20} = c'_{10} - c'_{20} \tag{22}$$

Suppose that

$$c'_1 > c'_2 > 0 \tag{23}$$

whence

$$C'_{10} > C'_{20} > 0 \tag{24}$$

The curves in Figs. 4a and b are plotted to the same time scale. Clearly

$$M_1 > M_2 \tag{25}$$

whence

$$M = M_1 \tag{26}$$

In view of the inequality (25) there are infinitely many solutions of C_2 versus time with $|m_2|$ not at its maximum value, which will bring C_2 to equilibrium in the time M_1. It follows that there are infinitely many solutions that bring c_1 and c_2 to equilibrium in the minimal time M. Thus the optimum nonlinear control problem is not solved by minimizing the duration of the transient. We are currently investigating this problem.

The work of the minimal time school must be assessed on the basis of

engineering as well as mathematical considerations. It is one of many possible approaches to optimum control, particularly attractive because of its mathematical elegance.

Recently N. P. Smith and the writer solved the problem of the optimum nonlinear control of a system where a single control function optimizes the response to all or almost all disturbances that may occur. Thus this function applies to ramps, sinusoids, etc. as well as step changes.

APPLICATIONS

A practical scheme for achieving optimum nonlinear control is shown in Fig. 5. This involves a nonlinear device composed of a sharp edged orifice in series with a linear (cylindrical) pipe section. We let p_1, p_2 and p_3 be the pressures of the liquid ahead of the linear section, before the sharp edged orifice, and after it respectively. For a constant a

$$p_1 - p_2 = a\,Q \tag{27}$$

where Q is the volumetric rate of flow through the unit. The pressure drop $p_2 - p_3$ across the sharp edged orifice satisfies

$$p_2 - p_3 = b\,\{Q\} \tag{28}$$

for the absquare $\{Q\}$ and a constant b. It follows that

$$p_1 - p_3 = a\,Q + b\,\{Q\} \tag{29}$$

When Q is small $b\{Q\}$ becomes K_3Q for a constant K_3 whence relation (29) is the linear expression

$$p_1 - p_3 = (a + K_3)\,Q \tag{30}$$

This principle is used in Fig. 5 in the case of a hydraulic speed governor for prime movers. The motion of the bellows output (B-bellows) is assumed to be small. The output m is proportional to the pressure p_1. Let A_1 and y be the area and coordinate of the free end of the A-bellows. The land area A_3 is small so that

$$A_1\,y' = Q \tag{31}$$

for practical purposes. Let K_3 be the scale of the A-bellows spring. It follows that

$$A_1\,p_3 = K_3\,y \tag{32}$$

The pilot valve is constructed so that

$$Q = -K_4 x \tag{33}$$

for the coordinate x of the valve plunger and flow coefficient K_4. By relations (31) and (32)

$$Q = \frac{A_1^2}{K_3} p_3' \tag{34}$$

whence by relation (29)

$$p_1 = p_3 + \frac{a A_1^2}{K_3} p_3' + \frac{b A_1^4}{K_3^2} \{p_3'\} \tag{35}$$

The force equation for the valve plunger is now

$$K_5 x = K_6 c + A_3 p_3 \tag{36}$$

for constants K_5 and K_6 and rpm c of the ballhead. By equations (33), (34) and (36)

$$\frac{K_5 A_1^2}{A_3 K_4 K_3} p_3' + p_3 = -\frac{K_6 c}{A_3} \tag{37}$$

The time constant $(K_5 A_1^2 / A_3 K_4 K_3)$ is taken small so that equation (37) reduces to

$$p_3 = -\frac{K_6 c}{A_3} \tag{38}$$

The stroke of the B-bellows is proportional to the stroke z of a piston which controls the rate of flow of liquid to the servomotor with piston coordinate m. For a constant K_7

$$m' = K_7 z \tag{39}$$

Also for a constant K_8

$$z = K_8 p_1 \tag{40}$$

By relations (35), (38) - (40) it follows that equation (12) is satisfied where

$$K = \frac{K_6 K_7 K_8}{A_3}$$

$$\gamma = \frac{a A_1^2}{K_3} \tag{41}$$

$$\beta = -\frac{b A_1^4 K_6}{K_3^2 A_3}$$

The results of running this governor [8] on a General Motors GM-671 engine are shown in Fig. 3.

An electrical version is pictured in Fig. 6. The triangles represent operational amplifiers. The absquare is obtained by the use of a thyrite resistor which has the property

$$i = C \{V\} \tag{42}$$

for the current i through the thyrite, voltage drop V, and a constant C. To satisfy relation (42) the voltage range is one for which the maximum value is about 10 times the minimum. For small V relation (42) becomes linear. In the controller of Fig. 6 the variable c is differentiated to obtain c'. A voltage V proportional to c' is applied to the thyrite to yield a current proportional to $\{c'\}$. Signals proportional to c, c' and $\{c'\}$ are added to obtain a voltage proportional to the control function Σ. This voltage is employed to stroke a valve in the usual manner [9].

A magnitude amplifier waterwheel governor version of Fig. 6 was developed for production [10].

The nonlinear governors discussed here are illustrative of the few optimum nonlinear controllers built by industry. Although there are areas where the improvement of performance justifies the use of such controllers, for most control problems linear control functions are still indicated, whether or not saturation is involved.

CONCLUSIONS

For physical systems in one controlled variable with one dominant lag and the input subject to saturation, a two to one improvement in the maximum error for intermediate disturbances may be obtained by going from the best linear control function to a nonlinear control function involving the absquare. Sharp edged orifices and nonlinear electrical resistors make the practical production of such control functions possible. Due to lags and other factors always present in physical systems more improvement cannot be obtained. The writer's nonlinear approach has the advantage that all reasonable criteria for optimum response are satisfied simultaneously. This includes minimizing the maximum error for a transient, which the user considers a most important practical criterion of goodness. He is usually not concerned with minimizing time.

Unfortunately minimizing the maximum error is excluded from the class of functions and functionals treated by the well-known optimization school of Pontryagin, Bellman and others. For the case of two or more controlled variables the minimal time approach fails in selecting the optimum nonlinear response. N. P. Smith and the writer recently found a single control function to give optimum nonlinear response to a large class of disturbances, if not all. Thus optimum response is obtained regardless of whether or not the disturbance is a step, ramp, sinusoid, or something else.

ACKNOWLEDGEMENT

The writer thanks the National Aeronautics and Space Administration for support in connection with this paper.

REFERENCES

1. Oldenburger, R., "Optimum Nonlinear Control." Trans. ASME, Vol. 79, 1957 pp. 527-546.

2. Oldenburger, R., Nicklas, J. C., Gamble, E. H., Optimum Non-Linear Control of a Second Order Nonlinear System, NASA Technical Note D-825, April 1961.

3. McDonald, D., "Nonlinear Techniques for Improving Servo Performance," National Electronics Conference, Vol. 6, 1950, pp. 400-421.

4. Hopkin, A. M., "A Phase-Plane Approach to the Compensation of Saturating Servomechanisms." Trans. AIEE, Vol. 70, Part I, 1951, pp. 631-639.

5. Bellman, R., Glicksberg, I., Gross, O., "On the 'Bang-Bang' Control Problem," Quarterly of Applied Mathematics, Vol. 14, 1965, pp. 11-18.

6. Kalman, R. E., "Analysis and Design Principles of Second Higher Order Saturating Servomechanisms," Trans. AIEE, Vol. 74, Part II, 1955, pp. 294-310.

7. Rozonoer, R. I., "L. S. Pontryagin's Maximum Principle in Theory of Optimum Systems." Automation and Remote Control, Parts I-III, Vol. 20, No. 10, Oct.-Dec., 1959, pp. 1320-1334, 1405-1421, 1517-1532.

8. Oldenburger, R., "Method and Apparatus for Hydraulic Control Systems," U.S. Patent No. 2,931,342, April 5, 1960.

9. Oldenburger, R., "Method and Apparatus for Controlling a Condition," U.S. Patent No. 2,960,629, November 15, 1960.

10. Oldenburger, R., "Nonlinear Speed and Load Governor for Alternators," U.S. Patent No. 2,908,826, October 13, 1959.

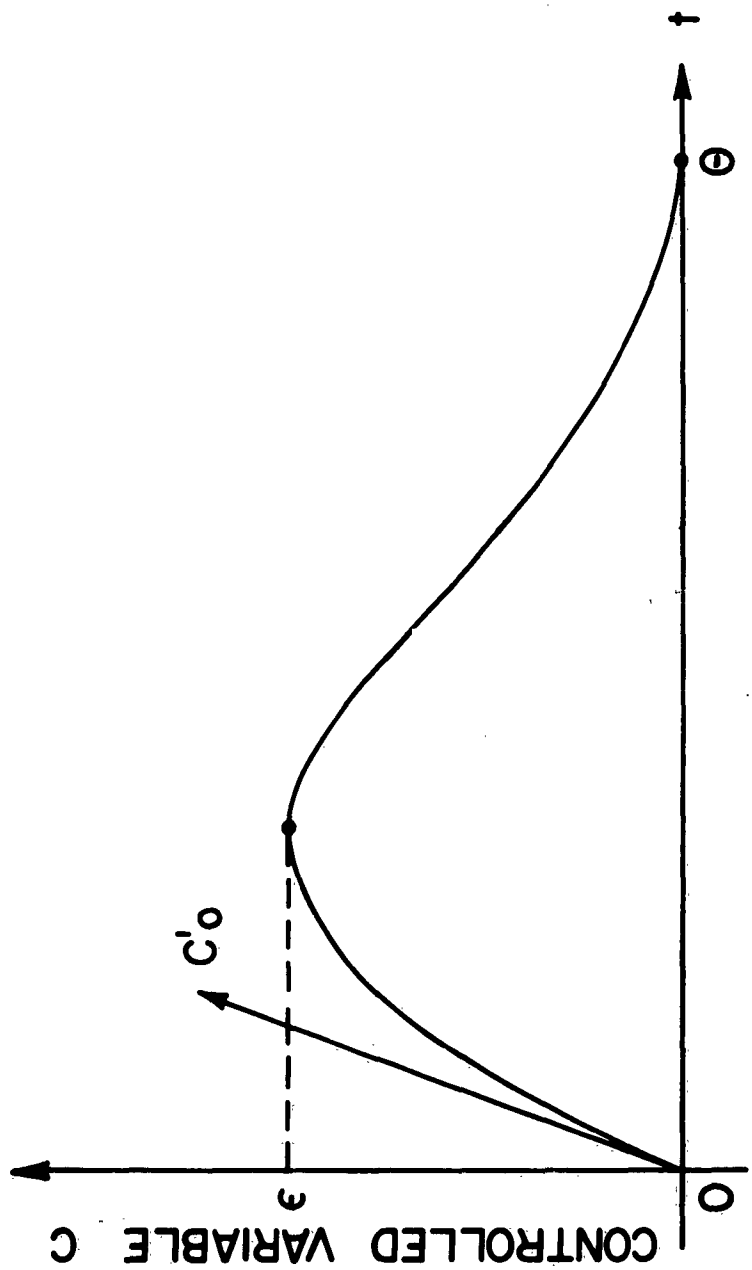


FIGURE 1. OPTIMUM RESPONSE

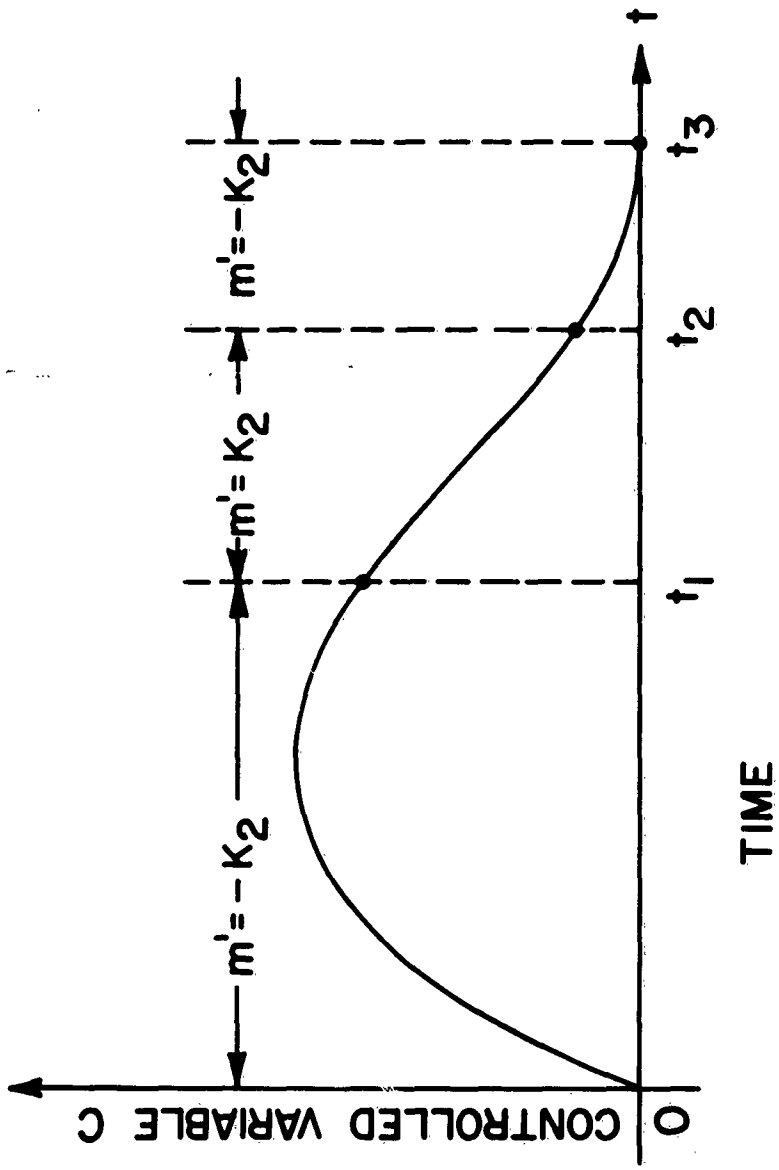


FIGURE 2. OPTIMUM RESPONSE FOR SYSTEM WITH TWO LAGS

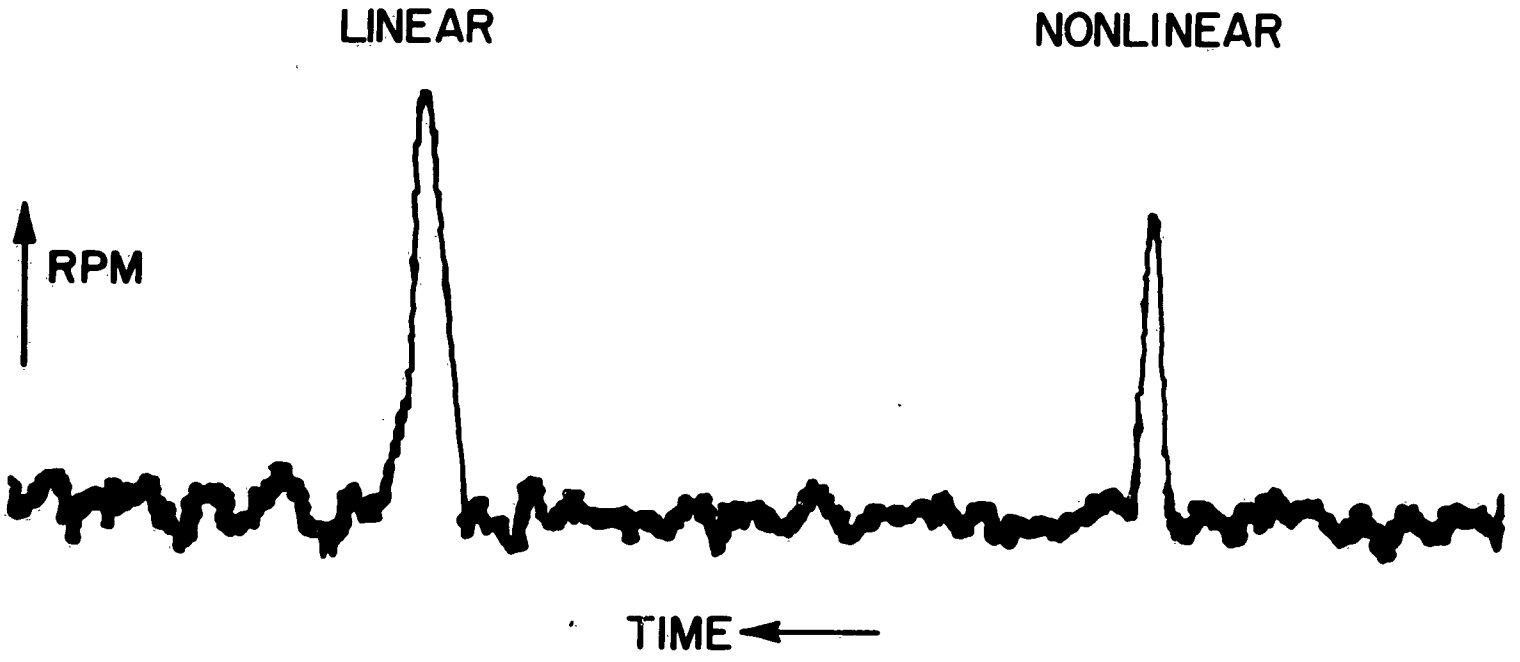


FIGURE 3. EXPERIMENTAL COMPARISON OF LINEAR AND NONLINEAR CONTROL FOR SAME DISTURBANCE

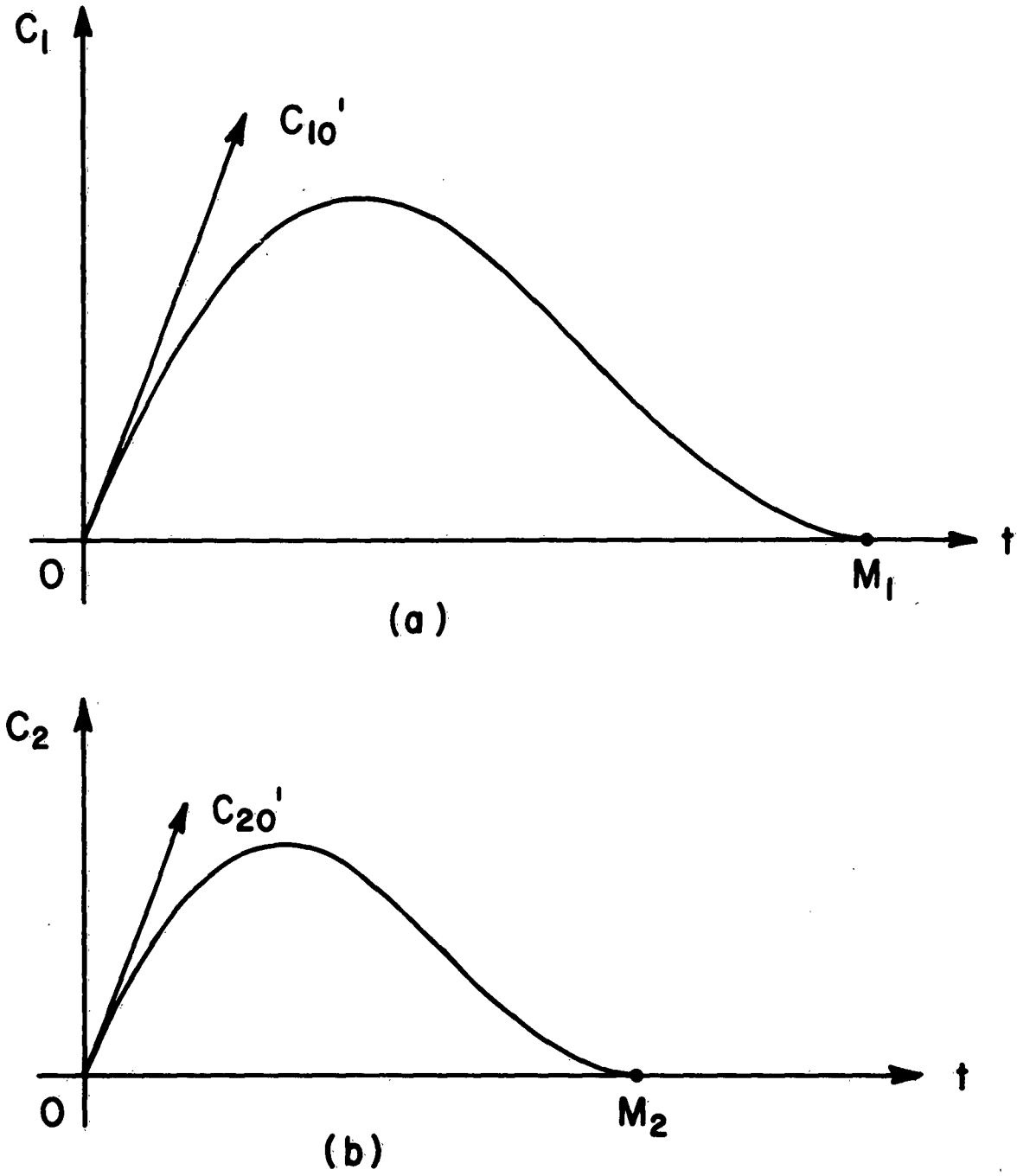


FIGURE 4. CASE OF TWO CONTROLLED VARIABLES

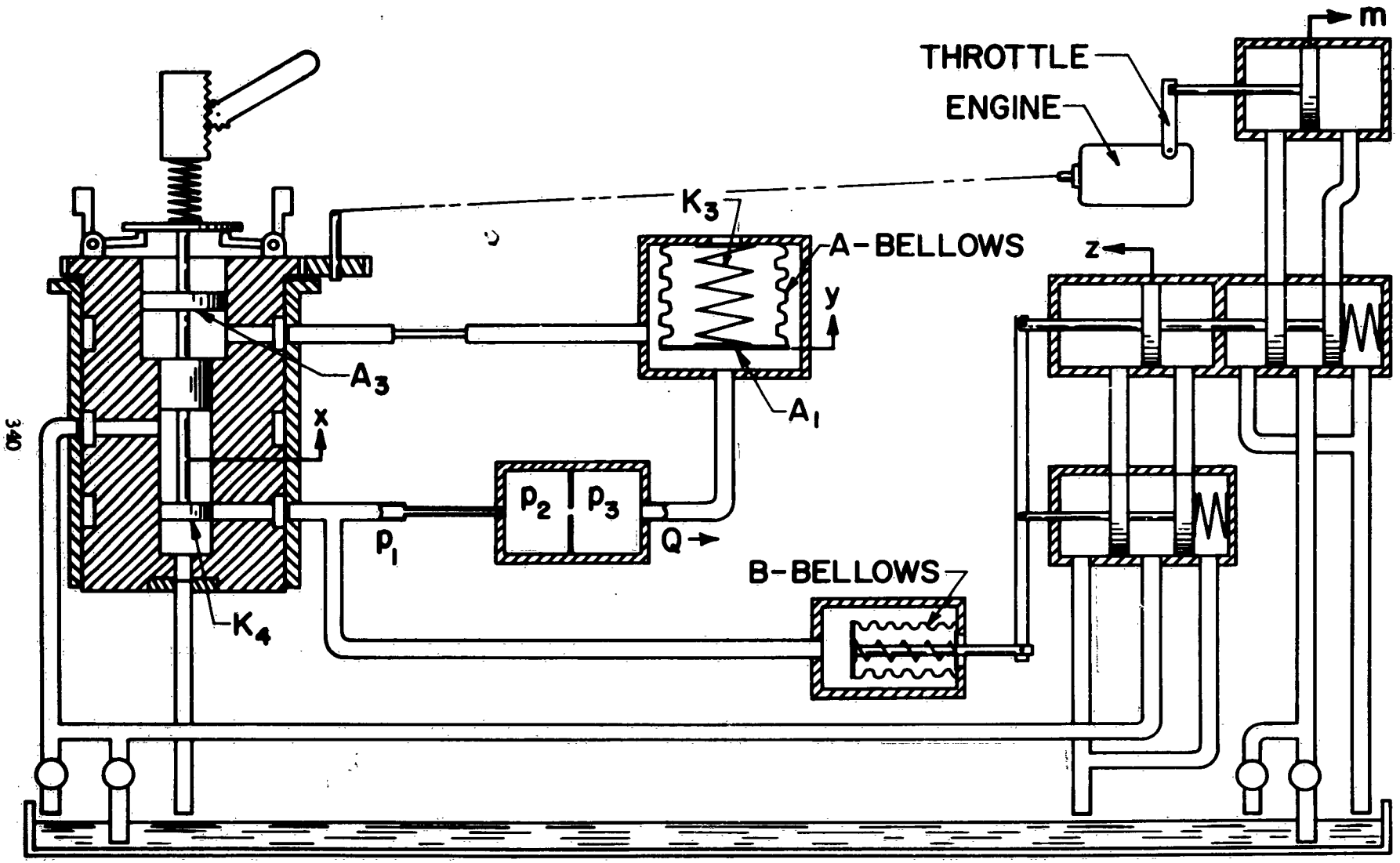


FIGURE 5. NONLINEAR HYDRAULIC GOVERNOR

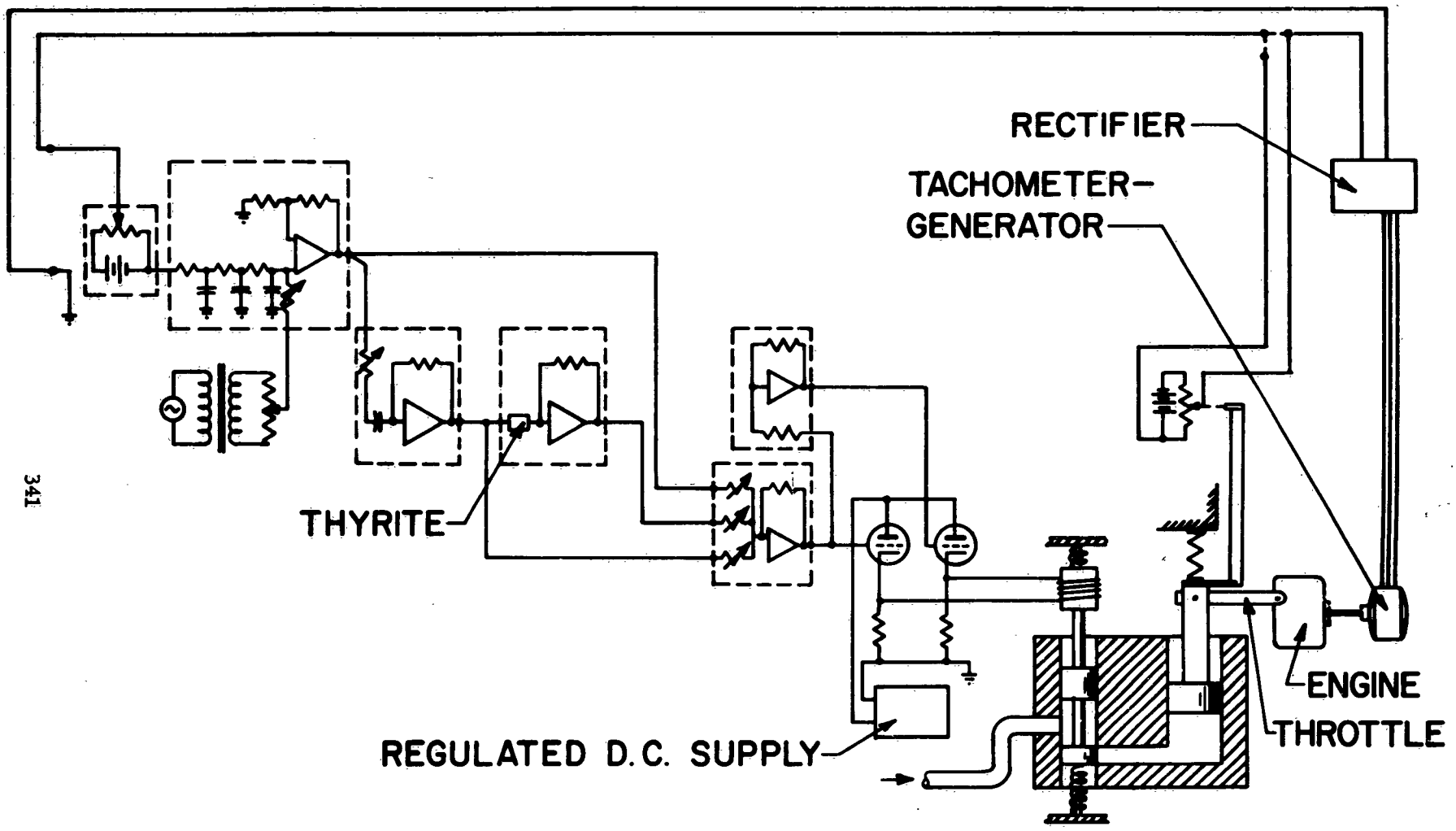


FIGURE 6. NONLINEAR ELECTRONIC GOVERNOR

Optimal Programming of Multivariable Control Systems in the Presence of Noise

By

A. E. Bryson, Jr.

Harvard University
Cambridge, Massachusetts

OPTIMAL PROGRAMMING OF MULTIVARIABLE CONTROL SYSTEMS IN THE PRESENCE OF NOISE

by

Arthur E. Bryson, Jr.

Professor of Mechanical Engineering, Harvard University, and

Consultant, Raytheon Co.

Abstract

This paper is a brief review of the current status of optimal programming theory as applied to multivariable control systems in the presence of noise. Open-loop programming of control variables for a noise-free non-linear system is discussed first, followed by a description of closed-loop gain programming for a linearized version of the same type of system. Optimal filtering of noisy measurements is then reviewed for linear time-varying systems. A combination of these techniques promises to provide optimal control for a wide class of systems and performance indices.

OPTIMAL PROGRAMMING FOR NONLINEAR SYSTEMS (NOISE-FREE)

The solution of optimal programming problems on computers may now be considered practical and straightforward through the use of gradient (steepest-ascent) techniques developed within the last few years (Refs. 1-7). The type of problem that can be solved is as follows:

Given: A physical system governed by:

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

where $x(t)$ = n by 1 matrix of state variables

$u(t)$ = q by 1 matrix of control variables

$x(t_o)$ Given

$f(x, u, t)$ = n by 1 matrix of known functions (non-linear)

Find: $u(t)$ to maximize

$$\phi [x(t_1) , t_1] ; \quad t_1 > t_o$$

with terminal constraints

$$\psi [x(t_1) , t_1] = 0$$

where

$\phi (x, t)$ is pay-off function (given)

$\psi (x, t)$ is an r by 1 matrix of given functions, $r \leq n - 1$.

Solution: Calculus of variations using steepest-ascent numerical calculations.

This is an open-loop control system with programmed control variables, $u(t)$. This is not a feasible type of control to use in many systems due to errors in the determination of initial conditions and parameters of the system. However, if this "path" is used as a nominal path and small perturbations are considered around it, a linear feedback control system with programmed gains can be synthesized that will bring the system very close to the desired terminal conditions.

NEIGHBORING-OPTIMUM FEEDBACK CONTROL (NOISE-FREE)

A simple approach to this problem is what might be called "minimum incremental effort" control. It is a straightforward application of a quadratic performance index to the system linearized around the optimum path of the previous section. This is discussed, with several examples, in Refs. 8 and 9. A slightly more sophisticated approach is used in Refs. 10 and 11 where the performance index is the same one used in the original nonlinear problem; here a true neighboring-optimum is found using the second variation. An outline of the type of problem solved by this latter technique is as follows:

Given: $u(t)$ that minimizes $\phi[x(t_1) , t_1]$ with $\psi[x(t_1) , t_1] = 0$

and given $x(t_o)$, for $\dot{x} = f(x, u, t)$

Find: $\delta u(t)$ such that $u(t) + \delta u(t)$ minimizes $\phi[x(t_1) , t_1]$ with

$\psi[x(t_1) , t_1] = 0$ and given $x(t_o) + \delta x(t_o)$ for $\dot{x} = f(x, u, t)$

Solution: $\delta u(t) = - \Lambda(t, t_o)\ \delta x(t_o)$, where $\Lambda(t, t_o)$ is pre-calculated from second variation.

Note 1: The concept is readily extended to continuous measurement of $\delta x(t)$, $\Rightarrow$ $\delta u(t) = - \Lambda(t)\ \delta x(t)$.

Note 2: $\delta x(t)$ due to $\delta u(t)$ governed by linear equations:

$$\frac{d}{dt}(\delta x) = F(t)\ \delta x + G(t)\ \delta u \quad ; \quad F_{ij} = \left(\frac{\partial f_i}{\partial x_j}\right)_{nom.}, \quad G_{ij} = \left(\frac{\partial f_i}{\partial u_j}\right)_{nom.}$$

OPTIMAL ESTIMATION FOR NOISY TIME-VARYING LINEAR SYSTEMS

In Refs. 12-14, a practical filter was developed for time-varying linear systems with Gaussian random noise in the system and in the measurements. The procedure may be regarded as keeping track of the mean values and covariances of a multivariable Gaussian distribution in a Markov process. An easy way to understand the optimal estimation procedure is to consider only one step of a discrete process where either (1) a set of measurements is made or (2), a set of discrete forcing functions is applied. First, consider the improvement in

the estimate as the result of a set of measurements:

Given: (1) A priori estimate of state variables and covariances:

$$\mathcal{E}[x] = \bar{x} \qquad \text{(an n vector)}$$

$$\mathcal{E}[(x - \bar{x})(x - \bar{x})^T] = \bar{P} \qquad \text{(an n x n symmetric matrix)}$$

(2) A set of measurements and covariances:

$$\mathcal{E}[z] = \bar{z} \qquad \text{(a p vector)}$$

$$\mathcal{E}[(z - \bar{z})(z - \bar{z})^T] = \bar{R} \qquad \text{(a p x p symmetric matrix)}$$

where

$$z = Hx$$

Find: Maximum likelihood estimate $\hat{x}$, that minimizes

$$\phi = \frac{1}{2} \tilde{z}^T \bar{R}^{-1} \tilde{z} + \frac{1}{2} \tilde{x}^T \bar{P}^{-1} \tilde{x} \quad ,$$

where

$$\tilde{(\ )} = \hat{(\ )} - \bar{(\ )} \quad .$$

Answer:

$$\hat{x} = \bar{x} + PH^T \bar{R}^{-1} (\bar{z} - H \bar{x})$$

$$P = \bar{P} - \bar{P}H^T (\bar{R} + H\bar{P}H^T)^{-1} H\bar{P}$$

$$= (\bar{P}^{-1} + H^T\bar{R}^{-1}H)^{-1}$$

where

$$P = \mathcal{E}[(x - \hat{x})(x - \hat{x})^T]$$

Next, consider the degradation in the estimate as the result of a set of forcing functions:

Given: (1) A priori estimate of state variables and covariances

$$\mathcal{E}[x] = \bar{x}$$

$$\mathcal{E}[(x - \bar{x})(x - \bar{x})^T] = \bar{P}$$

(2) A set of forcing functions and covariances:

$$\mathcal{E}[u] = \bar{u}$$

$$\mathcal{E}[(u - \bar{u})(u - \bar{u})^T] = \bar{Q}$$

where

$$\Delta x = Gu$$

Find: Revised estimate of state variables and covariances after application of forcing function.

Answer: $\hat{x} = \bar{x} + G\bar{u}$

$P = \bar{P} + G\bar{Q}G^T$

where

$P = \mathcal{E}[(x - \hat{x})(x - \hat{x})^T]$

If we now apply the above to a discrete process with a transition matrix $\Phi_{n,n-1}$ that describes the transition from the state at the $(n-1)^{st}$ step to the n^{th} step, we have the following:

STOCHASTIC CONTROLLER FOR DISCRETE MEASUREMENTS AND DISCRETE FORCING FUNCTIONS

$$\hat{x}_n = \begin{cases} \Phi_{n,n-1}\,\hat{x}_{n-1} & \text{if ordinary point} \\ \Phi_{n,n-1}\,\hat{x}_{n-1} + K_n\,(\bar{z}_n - H_n\,\Phi_{n,n-1}\,\hat{x}_{n-1}) & \text{if measurement made} \\ \Phi_{n,n-1}\,\hat{x}_{n-1} + G_n\,\bar{u}_n & \text{if forcing function applied} \end{cases}$$

$$P_n = \begin{cases} \Phi_{n,n-1}\,P_{n-1}\,\Phi^T_{n,n-1} & \text{if ordinary point} \\ [(\Phi_{n,n-1}\,P_{n-1}\,\Phi^T_{n,n-1})^{-1} + H_n^T\,\bar{R}_n^{-1}\,H_n]^{-1} & \text{if measurement made} \\ \Phi_{n,n-1}\,P_{n-1}\,\Phi^T_{n,n-1} + G_n\bar{Q}_nG_n^T & \text{if forcing function applied} \end{cases}$$

$K_n = P_nH_n^T\bar{R}_n^{-1}$, gain in optimal estimator

In the limit, as the process and the measurements become continuous, the results of Ref. 14 are obtained. Instead of a transition matrix we then have differential equations:

$$\dot{x} = F(t)\,x + G(t)\,u \quad .$$

In place of covariance matrices we have the correlation matrices $\bar{R}(t)$ and $\bar{Q}(t)$ that serve to describe the "white noise" in the measurements and in the forcing functions respectively:

$$\mathcal{E}[z(t)] = \bar{z}(t) \quad ,$$

$$\mathcal{E}\left\{[z(t) - \bar{z}(t)]\,[z(\tau) - \bar{z}(\tau)]^T\right\} = \bar{R}(t)\,\delta(t - \tau) \quad ,$$

where $z = Hx$, and

$$\epsilon[u(t)] = \bar{u}(t) \quad ,$$

$$\epsilon\left\{[u(t) - \bar{u}(t)]\,[u(\tau) - \bar{u}(\tau)]^T\right\} = \bar{Q}(t)\ \delta(t - \tau) \quad .$$

The optimal filter then becomes:

$$\dot{\hat{x}} = F\,\hat{x} + G\,\bar{u} + PH^T\,\bar{R}^{-1}\,(\bar{z} - H\hat{x}) \quad ; \quad x(t_o) = \bar{x}_o$$

$$\dot{P} = FP + PF^T - PH^T\bar{R}^{-1}\,HP + G\bar{Q}G^T \quad ; \quad P(t_o) = \bar{P}_o$$

where

$$\epsilon[x(t_o)] = \bar{x}_o$$

$$\epsilon\left\{[x(t_o) - \bar{x}_o]\,[x(t_o) - \bar{x}_o]^T\right\} = \bar{P}_o \quad .$$

Ref. 15 discusses the estimation problem from the point of view of maximum likelihood estimation which, in turn, is a type of "least square" fitting.

COMBINATION OF OPTIMAL FILTER AND CLOSED-LOOP OPTIMAL GAIN PROGRAMMING

In Ref.16 it is shown that, for quadratic performance indices, the combination of the optimal filter and the closed-loop optimal gain programming technique (noise-free) produce optimal control for linear systems in the presence of noise. There is reason to believe this combination will be optimal for a more general class of performance indices. Using the nomenclature of the previous sections the control system is given by

$$u(t) = -\Lambda(t)\,\hat{x}(t)$$

$$\dot{x} = F\,x + G(u + w)$$

$$\dot{\hat{x}} = F\,\hat{x} + Gu + PH^T\,\bar{R}^{-1}\,(\bar{z} - H\,\hat{x}) \quad ; \quad \hat{x}(t_o) = \bar{x}_o \quad ; \quad \bar{z} = Hx + v$$

$$\dot{P} = FP + PF^T - PH^T\,\bar{R}^{-1}\,HP + G\bar{Q}G^T \quad ; \quad P(t_o) = \bar{P}_o$$

where

$$\epsilon(w) = \epsilon(v) = 0$$

$$\epsilon[w(t)\,w^T(\tau)] = \bar{Q}(t)\ \delta(t - \tau)$$

$$\epsilon[v(t)\,v^T(\tau)] = \bar{R}(t)\ \delta(t - \tau)$$

where $\Lambda(t)$ is a set of pre-calculated time-varying gains determined from noise-free control considerations regarding $\hat{x}(t)$ as though it was the exact value of $x(t)$. Note $P(t)$ may be pre-calculated and the time-varying gains $K(t) = PH^T\,\bar{R}^{-1}$ stored in a memory for use during operation of the control system (see Fig.1). Obviously this system can also be combined with the neighboring-optimum feedback scheme described above.

REFERENCES

1. Kelley, H. J., "Gradient Theory of Optimal Flight Paths," Jour. Amer. Rocket Soc., Vol. 30, pp. 947-853, Oct. 1960.

2. Bryson, A. E., Denham, W. F., Carroll, F. J. and Mikami, K., "Lift or Drag Programs that Minimize Re-entry Heating," Inst. Aerospace Sciences, Annual Meeting, Jan. 1961; Jour. Aerospace Sci., Vol. 29, No. 4, April 1962, pp. 420-430.

3. Bryson, A. E., "A Gradient Method for Optimizing Multi-Stage Allocation Processes," Proc. Harvard University Symposium on Digital Computers and their Applications, April 1961 (to be Published).

4. Bryson, A. E. and Denham, W. F., A Steepest-Ascent Method for Solving Optimum Programming Problems, Raytheon Co. Rpt. BR-1303, 10 August 1961. Also: Journal Appl. Mech. (Trans. ASME, Series E) June 1962, pp. 247-257.

5. Mikami, K., Battle, C. T., Goodell, R. S., and Bryson, A. E., Three Dimensional Trajectory Optimization Study, ASD-TDR-62-295 Parts I, II, III, May 1962. (Also Raytheon Co. Rpt. BR-1759).

6. Bryson, A. E., Mikami, K. and Battle, C. T., Optimum Lateral Turns for a Re-entry Glider, Aerospace Engr., April 1962.

7. Leitmann, G. (Editor), Optimization Techniques, Academic Press, August 1962. (See particularly Ch. 6 by H. J. Kelley, "Method of Gradients").

8. Bryson, A. E. and Denham, W. F., "Multivariable Terminal Control to Minimize Mean Square Deviation from a Nominal Path," Proceedings of Inst. Aero. Sci. Symp. on Vehicle Systems Optimization, November, 1961, Garden City, N. Y. (Also Raytheon Co. Rpt. BR-1333, Sept. 20-1961).

9. Bryson, A. E. and Denham, W. F., "A Guidance Scheme for Supercircular Re-entry of a Lifting Vehicle," Jour. Amer. Rocket Soc., Vol. 32, No. 6, June 1962, pp. 894-898.

REFERENCES (Cont'd)

10. Kelley, H. J., "Guidance Theory and Extremal Fields," Presented Nat. Aerospace Electronics Convention, May 14-16, 1962, Dayton, Ohio.

11. Breakwell, J. V., and Bryson, A. E., Neighboring -Optimum Terminal Control for Multivariable Non-Linear Systems," to be presented at Soc. Indus. and Appl. Math. Conf. on Control, Boston, Nov. 1962.

12. Swerling, P. "A Proposed Stagewise Differential Correction Procedure for Satellite Tracking and Prediction," Jour. Astro. Science, Autumn 1959.

13. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Jour. Basic. Engr. (Trans. ASME Series D) Vol. 82, 1960, pp. 34-45.

14. Kalman, R. E. and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," Jour. Basic Engr. (Trans. ASME, Series D) Vol. 83, 1961, pp. 95-108. (see also RIAS Tech. Rpt. 61-1 for elaboration).

15. Bryson, A. E. and Frazier, M., "Smoothing for Linear and Nonlinear Dynamic Systems," Proc. of Conf. on Optimum Systems Synthesis, Wright-Patterson Air Force Base, Ohio, Sept. 11-13, 1962.

16. Gunckel, T. F. and Franklin, G. F., "A General Solution for Linear Sampled-Data Control Systems," Trans. ASME, Jour. Basic Engr. (to appear).

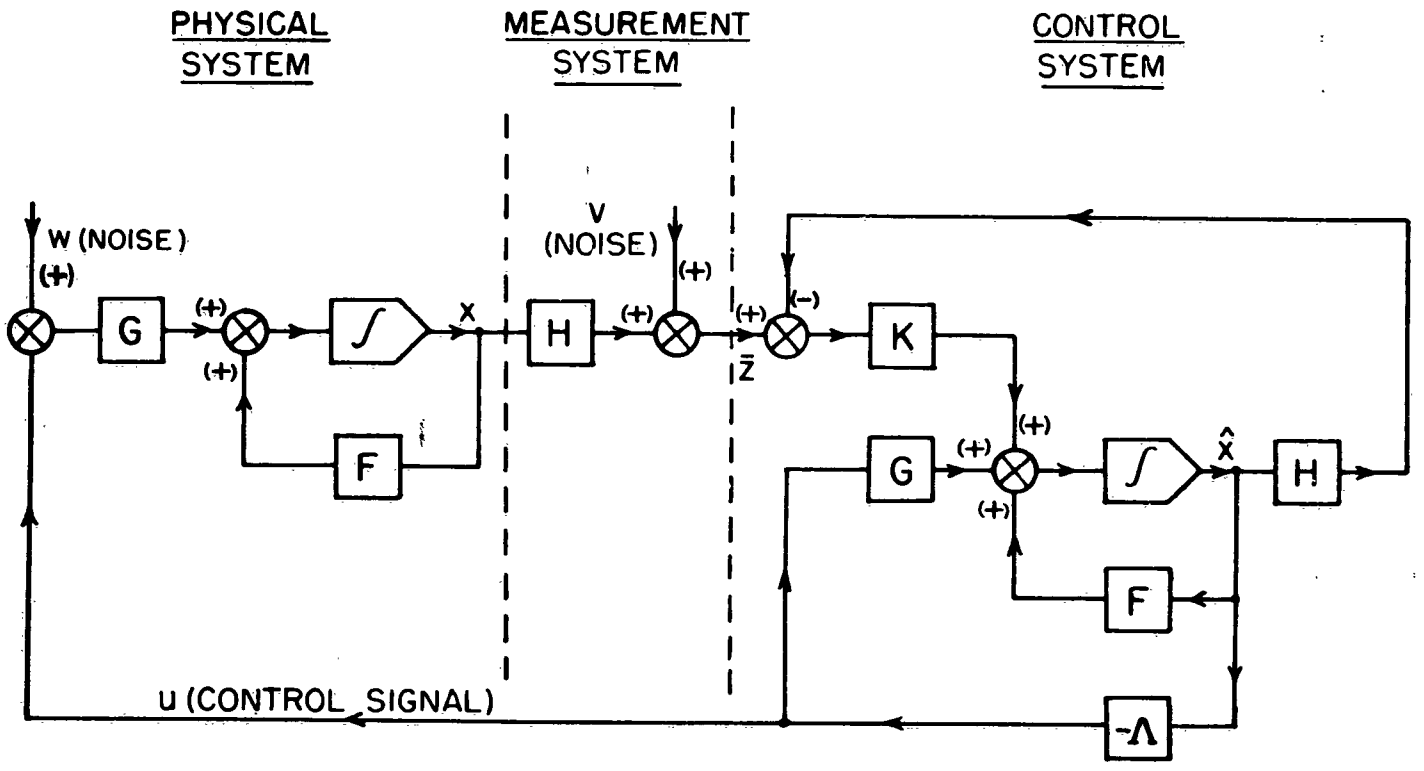


Figure 1. Combination of Optimal Filter and Closed-Loop Optimal Gain Programming

ASD-TDR-63-119

Smoothing for Linear and Nonlinear Dynamic Systems

By

A. E. Bryson
M. Frazier

Harvard University
Cambridge, Massachusetts

and

Raytheon Company
Bedford, Massachusetts

SMOOTHING FOR LINEAR AND NONLINEAR DYNAMIC SYSTEMS

Arthur E. Bryson* and Malcolm Frazier**

ABSTRACT

A method is presented for making maximum likelihood estimates of the state variables of linear and nonlinear dynamic systems over a finite time interval, utilizing measurements made during this time interval and, if available, a priori estimates of the initial conditions and the forcing functions. For a linear system, the filtering results are identical to those of Kalman and Bucy; the smoothing result is an extension of their work.

ACKNOWLEDGEMENT

The first author wishes to thank the Boeing Company and the Raytheon Company for support of this work.

STATEMENT OF THE SMOOTHING PROBLEM

Given: (1) A dynamic system governed by nonlinear differential equations,

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

where

$x(t)$ = n x 1 matrix of state variables,
$u(t)$ = m x 1 matrix of forcing functions,
f = n x 1 matrix of known functions of x, u, t
t = independent variable (often time)
$\dot{(\ )} = \frac{d}{dt}(\ )$

(2) An a priori estimate of initial conditions,

$$E[x(t_o)] = \bar{x}_o \text{, and}$$

an a priori estimate of errors in $\bar{x}_o$,

$$E\left\{[x(t_o) - \bar{x}_o][x(t_o) - \bar{x}_o]^T\right\} = \bar{P}_o \text{, an n x n matrix}$$

(symmetric).

---

*Professor of Mechanical Engineering, Harvard University, Cambridge, Mass.

**Senior Engineer, Raytheon Company, Bedford, Mass.

(3) An a priori estimate of the forcing functions,

$$E[u(t)] = \bar{u}(t), \text{ and}$$

an a priori estimate of errors in $\bar{u}(t)$,

$$E\left\{[u(t) - \bar{u}(t)]\,[u(\tau) - \bar{u}(\tau)]^T\right\} = \bar{Q}(t)\,\delta(t-\tau)$$

(4) A set of measurements for $t_o \leq t \leq t_f$,

$$E[z(t)] = \bar{z}(t), \text{ a } p \times 1 \text{ matrix of functions,}$$

where the theoretical relations between "exact" measurements, $z(t)$, and state variables $x(t)$, is $h[z(t), x(t), t] = 0$, a $p \times 1$ matrix of known functions, and an a priori estimate of errors in $\bar{z}(t)$,

$$E\left\{[z(t) - \bar{z}(t)]\,[z(\tau) - \bar{z}(\tau)]^T\right\} = \bar{R}(t)\,\delta(t-\tau)$$

In the above equations $E[\ ]$ means "expected value of $[\ ]$", i.e. an ensemble average over many identical systems performing simultaneously. The sumbol $\delta(t-\tau)$ is the Dirac delta function and, appearing in the correlation equations above, it indicates "white noise" in the forcing functions and the measurements. In effect what this means is that we assume the correlation times of the noise are short compared to times of interest in the problem. The sumbol $(\ )^T$ indicates the transpose of the matrix $(\ )$.

Find: The maximum likelihood estimate of $u(t) = \hat{u}(t)$ and the initial conditions $x(t_o) = \hat{x}_o$. Note this gives the maximum likelihood estimate of $x(t) = \hat{x}(t)$ from

$$\dot{\hat{x}} = f(\hat{x}, \hat{u}, t)$$

$$x(t_o) = x_o$$

SOLUTION TO THE SMOOTHING PROBLEM

The maximum likelihood estimate will be obtained by minimizing

$$(1) \qquad J = \frac{1}{2}\tilde{x}_o^T \bar{P}_o^{-1} \tilde{x}_o + \frac{1}{2}\int_{t_o}^{t_f} (\tilde{z}^T \bar{R}^{-1} \tilde{z} + \tilde{u}^T \bar{Q}^{-1} \tilde{u})\, dt,$$

where

$$(\tilde{\ }) = (\hat{\ }) - (\bar{\ })$$

with the constraints:

(2) $$h(\hat{z}, \hat{x}, t) = 0$$

(3) $$\dot{\hat{x}} - f(\hat{x}, \hat{u}, t) = 0.$$

This is a standard problem in the calculus of variations.* It may be solved by adjoining the constraints (2) and (3) to the expression for J in (1):

(4) $$J = \frac{1}{2} \tilde{x}_o^T \bar{P}_o^{-1} \tilde{x}_o + \int_{t_o}^{t_f} \left\{ \frac{1}{2} \tilde{z}^T \bar{R}^{-1} \tilde{z} + \frac{1}{2} \tilde{u}^T \bar{Q}^{-1} \tilde{u} + \lambda^T [\dot{\hat{x}} - f(\hat{x}, \hat{u}, t)] + \mu^T h(\hat{z}, \hat{x}, t) \right\} dt$$

Now consider variations of $\hat{u}$ and $\hat{x}_o$ from their optimum values:

(5) $$\delta J = \tilde{x}_o^T \bar{P}_o \delta \hat{x}_o + \int_{t_o}^{t_f} \left\{ \tilde{z}^T \bar{R}^{-1} \delta \hat{z} + \tilde{u}^T \bar{Q}^{-1} \delta \hat{u} + \lambda^T [\delta \dot{\hat{x}} - \frac{\partial f}{\partial x} \delta \hat{x} - \frac{\partial f}{\partial u} \delta \hat{u}] + \mu^T [\frac{\partial h}{\partial z} \delta \hat{z} + \frac{\partial h}{\partial x} \delta \hat{x}] \right\} dt$$

or,

(6) $$\delta J = \tilde{x}_o^T \bar{P}_o^{-1} \delta \hat{x}_o + [\lambda^T \delta x]_{t_o}^{t_f} + \int_{t_o}^{t_f} \left\{ (\tilde{z}^T \bar{R}^{-1} + \mu^T \frac{\partial h}{\partial z}) \delta \hat{z} + (-\dot{\lambda}^T - \lambda^T \frac{\partial f}{\partial x} + \mu^T \frac{\partial h}{\partial x}) \delta \hat{x} + (\tilde{u}^T \bar{Q}^{-1} - \lambda^T \frac{\partial f}{\partial u}) \delta \hat{u} \right\} dt$$

Let us now choose $\lambda(t)$ and $\mu(t)$ so that

(7) $$\dot{\lambda}^T + \lambda^T \frac{\partial f}{\partial x} - \mu^T \frac{\partial h}{\partial x} = 0; \quad \lambda(t_f) = 0, \quad \tilde{x}_o^T \bar{P}_o^{-1} - \lambda^T(t_o) = 0,$$

(8) $$\tilde{z}^T \bar{R}^{-1} + \mu^T \frac{\partial h}{\partial z} = 0.$$

Then Eqn. (6) becomes,

(9) $$\delta J = \int_{t_o}^{t_f} (\tilde{u}^T \bar{Q}^{-1} - \lambda^T \frac{\partial f}{\partial u}) \delta \hat{u} \, dt.$$

*See for example Bryson, A. E., and Denham, W. F., "A Steepest Ascent Method for Solving Optimum Programming Problems," Jour. Appl. Mech., June 1962, pp. 247-257.

If $\hat{u}(t)$ is the optimal estimate of the forcing functions then $\delta J = 0$ for arbitrary $\delta\hat{u}$. For this to be true, it is necessary that

$$\tilde{u}^T \bar{Q}^{-1} - \lambda^T \frac{\partial f}{\partial u} = 0 \tag{10}$$

Summarizing, the optimal estimate of u(t) and x(t) will be obtained by solving the two point boundary value problem,

Diff. Eqns.

$$\dot{\hat{x}} = f(\hat{x}, \hat{u}, t) \quad \text{with} \quad \hat{u} = \bar{u} + \bar{Q}\left(\frac{\partial f}{\partial u}\right)^T \lambda , \tag{11}$$

$$\dot{\lambda} = -\left(\frac{\partial f}{\partial x}\right)^T \lambda - \left(\frac{\partial h}{\partial x}\right)^T \left(\frac{\partial h}{\partial z}\right)^{-T} \bar{R}^{-1}(\hat{z} - \bar{z}), \quad \text{with } h(\hat{z}, \hat{x}, t) = 0. \tag{12}$$

Boun. Cond.

$$\hat{x}(t_o) = \bar{x}(t_o) + \bar{P}_o \lambda(t_o) \tag{13}$$

$$\lambda(t_f) = 0 \tag{14}$$

Note 1: If no a priori estimates of $x(t_o)$ and u(t) are given, this essentially means $\bar{P}_o \to \infty$, $\bar{Q} \to \infty$. The conditions involving $\bar{x}_o$, $\bar{u}$, $\bar{P}_o$ and $\bar{Q}$ above are simply replaced by,

$$\lambda^T \frac{\partial f}{\partial u} = 0 \tag{15}$$

$$\lambda(t_o) = 0 \tag{16}$$

Note 2: A simple extension of this analysis would be to make $\bar{R}$ and $\bar{Q}$ functions of x and u.

Note 3: Another simple extension would be to include estimation of some parameters of the system.

SMOOTHING FOR LINEAR DYNAMIC SYSTEMS

If the system is linear, it can be put into the form,

$$\dot{x} = F(t)\,x + G(t)\,u \tag{17}$$

$$z = H(t)\,x + w(t) \tag{18}$$

where w(t) is a known set of bias functions. The smoothing problem then simplifies to the following:

$$\frac{d}{dt}\begin{bmatrix}\hat{x}\\ \lambda\end{bmatrix} = \begin{bmatrix}F & G\bar{Q}G^T\\ H^T\bar{R}^{-1}H, & -F^T\end{bmatrix}\begin{bmatrix}\hat{x}\\ \lambda\end{bmatrix} + \begin{bmatrix}G\bar{u}\\ -H^T\bar{R}^{-1}(\bar{z}-w)\end{bmatrix} \tag{19, 20}$$

with boundary conditions,

$$\hat{x}(t_o) - \bar{P}_o\,\lambda(t_o) = \bar{x}_o \tag{21}$$

$$\lambda(t_f) = 0 \tag{22}$$

This two-point boundary value problem can be solved in a straightforward manner due to the linearity of eqns. (19) and (20) as follows:

Let $\hat{x}_p(t)$ and $\lambda_p(t)$ be solutions of Eqns. (19) and (20) with initial conditions

$$\hat{x}_p(t_o) = \bar{x}(t_o), \tag{23}$$

$$\lambda_p(t_o) = 0. \tag{24}$$

Let $\Phi_x(t)$ and $\Phi_\lambda(t)$ be nxn matrices which are solutions to the homogeneous version of Eqns. (19) and (20), i.e., with $\bar{u} = \bar{z} - w = 0$, with initial conditions,

$$\Phi_x(t_o) = \bar{P}_o\,, \tag{25}$$

$$\Phi_\lambda(t_o) = \text{unit matrix} = I. \tag{26}$$

It follows that

$$\hat{x}(t) = \hat{x}_p(t) + \Phi_x(t)\,\lambda(t_o) \tag{27}$$

$$\lambda(t) = \lambda_p(t) + \Phi_\lambda(t)\,\lambda(t_o) \tag{28}$$

In order to satisfy the boundary condition (22), we must have

$$\lambda(t_o) = -[\Phi_\lambda(t_f)]^{-1}\lambda_p(t_f) \tag{29}$$

Combining (29) with (27), (28) and the relation,

$$\hat{u}(t) = \bar{u}(t) + \bar{Q}G^T\lambda(t) \tag{30}$$

we have finally,

$$\hat{x}(t) = \hat{x}_p(t) - \Phi_x(t)\,[\Phi_\lambda(t_f)]^{-1}\,\lambda_p(t_f) \tag{31}$$

$$\hat{u}(t) = \bar{u}(t) + \bar{Q}G^T\left\{\lambda_p(t) - \Phi_\lambda(t)\,[\Phi_\lambda(t_f)]^{-1}\,\lambda_p(t_f)\right\} \tag{32}$$

$$\hat{z}(t) = H\hat{x} + w \tag{33}$$

This is the solution to the smoothing problem, i.e., the problem of finding the maximum likelihood estimates of x(t), u(t), and z(t) over an entire interval $t_o \leq t \leq t_f$.

The filtering problem is concerned with finding the maximum likelihood estimate of $x(t_f)$ only, and the prediction problem is concerned with finding the maximum likelihood estimate of x(t) for $t > t_f$. Elegant solutions to the filtering and prediction problems have been given by Kalman and Bucy*. Solutions to these latter two problems by the present method are given in the next section.

It is also possible to find the covariance matrix of errors in the smoothing estimate $\hat{x}(t)$ by considering small variations in $\hat{x}$, $\bar{x}$, $\hat{u}$, $\bar{u}$, $\hat{z}$, $\bar{z}$ in Eqns. (19) through (22):

$$\frac{d}{dt}\begin{bmatrix}\delta\hat{x}\\ \delta\lambda\end{bmatrix} = \begin{bmatrix}F & G\bar{Q}G^T\\ H^T\bar{R}^{-1}H, & -F^T\end{bmatrix}\begin{bmatrix}\delta\hat{x}\\ \delta\lambda\end{bmatrix} + \begin{bmatrix}G\delta\bar{u}\\ -H^T\bar{R}^{-1}\delta\bar{z}\end{bmatrix} \tag{34, 35}$$

$$\delta\lambda(t_f) = 0 \tag{36}$$

$$\delta\hat{x}(t_o) - \bar{P}_o\,\delta\lambda(t_o) = \delta\bar{x}_o \tag{37}$$

Let

$$\begin{cases} P_{xx}(t) = E(\delta\hat{x}\,\delta\hat{x}^T) \\ P_{x\lambda}(t) = E(\delta\hat{x}\,\delta\lambda^T) \\ P_{\lambda\lambda}(t) = E(\delta\lambda\;\delta\lambda^T) \end{cases} \tag{38}$$

---

*Kalman, R. E., and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," Jour. Basic Engr., March 1961.

It then follows immediately from Eqns. (34) to (37) and the relations

$$E[\delta\bar{u}(t)\,\delta\bar{u}(\tau)^T] = \bar{Q}(t)\,\delta(t-\tau), \tag{39}$$

$$E[\delta\bar{z}(t)\,\delta\bar{z}(\tau)^T] = \bar{R}(t)\,\delta(t-\tau), \tag{40}$$

$$E[\delta\bar{u}(t)\,\delta\bar{z}(\tau)^T] = 0, \tag{41}$$

that the covariance matrices (38) are solutions to

$$\dot{P} = \mathcal{F}P + P\mathcal{F}^T + \mathcal{G} \tag{42}$$

where

$$P = \begin{bmatrix} P_{xx} & P_{x\lambda} \\ P_{x\lambda}^T & P_{\lambda\lambda} \end{bmatrix}, \quad \mathcal{F} = \begin{bmatrix} F & G\bar{Q}G^T \\ H^T\bar{R}^{-1}H, & -F^T \end{bmatrix}, \tag{43, 44}$$

$$\mathcal{G} = \begin{bmatrix} G\bar{Q}G^T, & 0 \\ 0 & H^T\bar{R}^{-1}H \end{bmatrix}.$$

The boundary conditions for (42) are

$$\begin{cases} P_{\lambda\lambda}(t_f) = 0 \\ P_{x\lambda}(t_f) = 0 \\ P_{xx}(t_o) - \bar{P}_o P_{x\lambda}^T(t_o) - P_{x\lambda}(t_o)\bar{P}_o + \bar{P}_o P_{\lambda\lambda}(t_o)\bar{P}_o = \bar{P}_o \end{cases} \tag{45}$$

Thus, finding the covariance matrices is also a two-point boundary value problem.

It can be shown that the correlation matrix of errors in $\hat{u}(t)$ is given by,

$$GQG^T = G\bar{Q}G^T + G\bar{Q}G^T P_{x\lambda}^T + P_{x\lambda} G\bar{Q}G^T, \tag{46}$$

where

$$E[\delta\hat{u}(t)\,\delta\hat{u}(\tau)^T] = Q(t)\,\delta(t-\tau). \tag{47}$$

Similarly,

$$H^T\bar{R}^{-1}R\bar{R}^{-1}H = -H^T\bar{R}^{-1}HP_{x\lambda} - P_{x\lambda}^T H^T\bar{R}^{-1}H, \tag{48}$$

where

$$E[\delta\hat{z}(t)\,\delta\hat{z}(\tau)^T] = R(t)\,\delta(t-\tau). \tag{49}$$

FILTERING FOR LINEAR SYSTEMS

If only $\hat{x}(t_f)$ is wanted, the two point boundary value problem of Eqs. (19) - (22) may be converted to an initial value problem. Referring to Eq. (31), we define a new quantity

$$y(t) = \hat{x}_p(t) - \Phi_x(t)[\Phi_\lambda(t)]^{-1}\lambda_p(t) \tag{50}$$

Note that

$$y(t_f) = \hat{x}(t_f) \tag{51}$$

Let us also define another quantity

$$P_y(t) = \Phi_x(t)[\Phi_\lambda(t)]^{-1} \tag{52}$$

Differentiating (52) with respect to t and using the relations

$$\frac{d}{dt}\begin{bmatrix}\Phi_x \\ \Phi_\lambda\end{bmatrix} = \begin{bmatrix}F & G\bar{Q}G^T \\ H^T\bar{R}^{-1}H, & -F^T\end{bmatrix}\begin{bmatrix}\Phi_x \\ \Phi_\lambda\end{bmatrix} \tag{53}$$

$$\begin{cases}\Phi_x(t_o) = \bar{P}_o \\ \Phi_\lambda(t_o) = I\end{cases} \tag{54}$$

It is straightforward to show that

$$\dot{P}_y = FP_y + P_yF^T - P_yH^T\bar{R}^{-1}HP_y + G\bar{Q}G^T \tag{55}$$

$$P_y(t_o) = \bar{P}_o \tag{56}$$

Differentiating (50) with respect to t and using (19), (20), (23), (24), (52), and (55), it is also straightforward to show that

$$\dot{y} = Fy + G\bar{u} + P_yH^T\bar{R}^{-1}(\bar{z} - Hy - w) \tag{57}$$

$$y(t_o) = \bar{x}_o \tag{58}$$

Eqs. (55) through (58) are precisely those obtained earlier by Kalman and Bucy (Loc. Cit.). Fig. 1 shows a block diagram of this filter. Note that $P_y(t)$ may be precalculated or calculated simultaneously with $y(t)$.

The solution to the smoothing problem (linear) of Eqs. (19) through (22) can be obtained by "filtering" to obtain $y(t_f) = x(t_f)$, then integrating Eqs. (19) and (20) backward with boundary conditions

(59, 60) $$\hat{x}(t_f) = y(t_f) \; ; \quad \lambda(t_f) = 0 \; .$$

A block diagram of this "smoother" is shown in Fig. 2.

Similarly it can be shown that

(61) $$P_y(t_f) = P_{xx}(t_f) \; .$$

Hence, the solution to the covariance Eqs. (42) - (45) for the smoothing problem can be obtained by "filtering" to obtain $P_y(t_f) = P_{xx}(t_f)$, then integrating Eqs. (42) - (44) backward with boundary conditions

(62) $$P_{xx}(t_f) = P_{yy}(t_f) \; ,$$

(63) $$P_{x\lambda}(t_f) = 0 \; ,$$

(64) $$P_{\lambda\lambda}(t_f) = 0 \; .$$

The impulse response functions (kernels) for the linear filter are easily found using the equations adjoint to equations (19) and (20),

(65)
(66) $$\frac{d}{dt} \begin{bmatrix} \Psi_x \\ \Psi_\lambda \end{bmatrix} = \begin{bmatrix} -F^T & , & -H^T \bar{R}^{-1} H \\ -G \bar{Q} G^T , & & F \end{bmatrix} \begin{bmatrix} \Psi_x \\ \Psi_\lambda \end{bmatrix}$$

with boundary conditions,

(67) $$\Psi_x(t_f) = I = \text{unit matrix}$$

(68) $$\Psi_x(t_o) \bar{P}_o + \Psi_\lambda(t_o) = 0 \; .$$

It follows immediately that

(69) $$\hat{x}(t_f) = \int_{t_o}^{t_f} [\Psi_x^T G \bar{u} - \Psi_\lambda^T H^T \bar{R}^{-1} (\bar{z} - w)] \, dt + \Psi_x^T(t_o) \bar{x}_o \; .$$

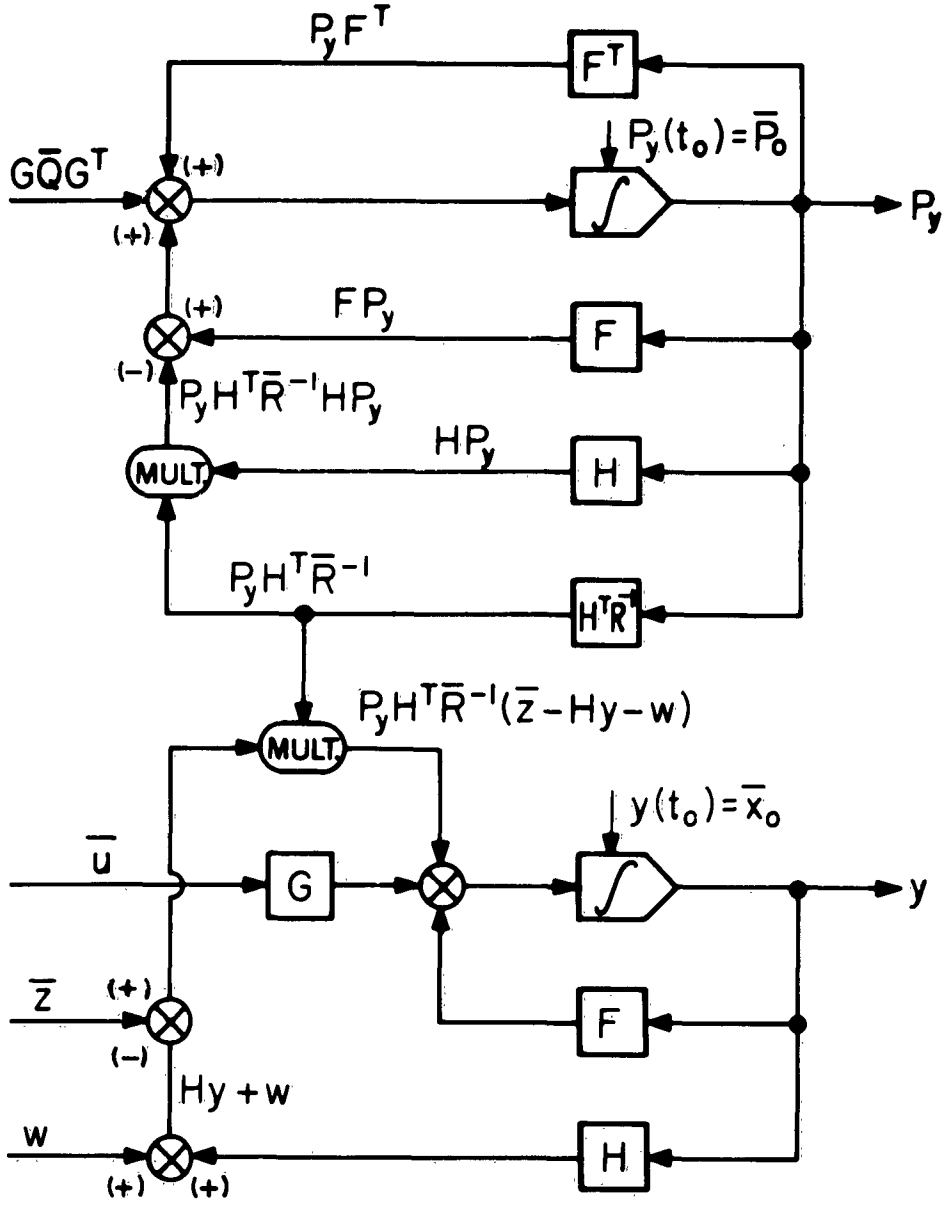


FIGURE 1. OPTIMAL FILTER FOR LINEAR MULTIVARIABLE SYSTEM

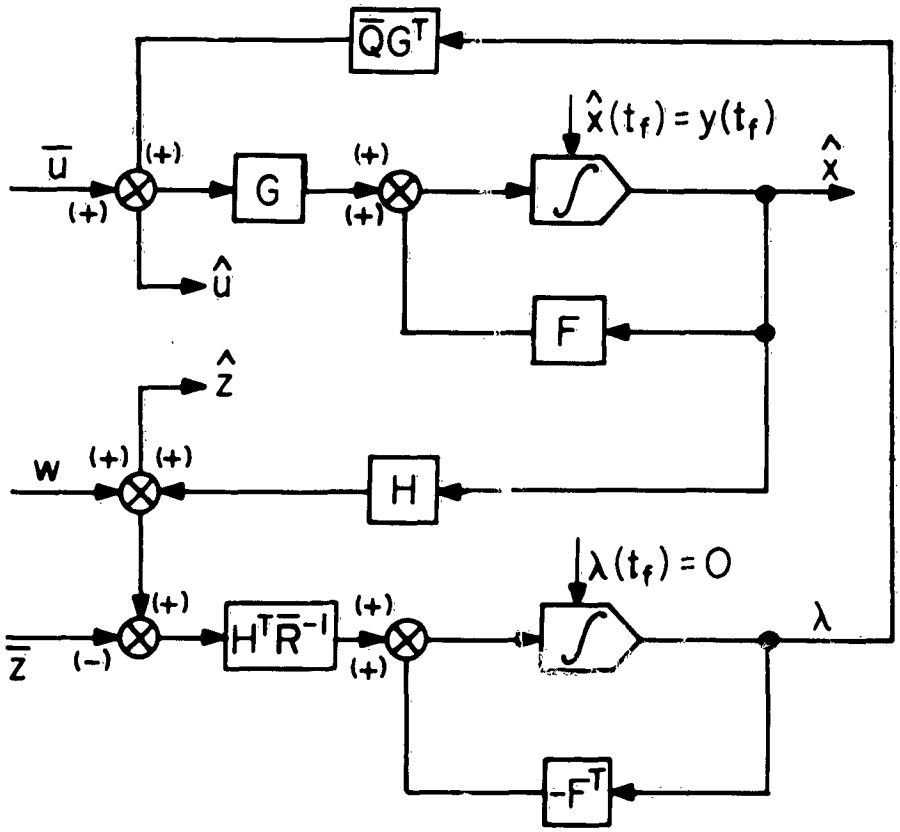


FIGURE 2. OPTIMAL SMOOTHER FOR LINEAR MULTIVARIABLE SYSTEM. (TO BE RUN BACKWARD AFTER $\hat{X}(t_f)$ OBTAINED FROM OPTIMAL FILTER)

ASD-TDR-63-119

First Order Implications of the Calculus of Variations in Guidance and Control

By

R. E. Kalman

Research Institute for Advanced Studies (RIAS)
Baltimore, Md.

FIRST ORDER IMPLICATIONS OF THE CALCULUS OF VARIATIONS IN GUIDANCE AND CONTROL

by

R. E. Kalman

Research Institute for Advanced Studies (RIAS)
Baltimore, Md.

Due to evitable conflicts of space and time, this paper appears here only in digest form. A detailed exposition will appear in the near future as a RIAS report [5].

The principal contribution of this paper is the recognition that Carathéodory's classification of regular extremals and the theory of controllability are closely related topics.

1. All notations being the same as in [1], consider a standard variational problem of the following type: (i) The initial point $P_0 = \{(t_0, x_0)\}$ of the motion to be optimized is fixed; (ii) the final point must lie on the surface S_1 in $(1+n)$- dimensional (t, x)-space; (iii) the motions are governed by the equation $dx/dt = f(t, x, u(t))$, (iv) values of $u(t)$ lie in the set $U(t)$ for each t; (v) the quantity to be minimized is the integral of function $L(t, x(t), u(t))$ along the motion from P_0 to S_1.

Under these circumstances, the problem is analyzed [1] by defining first a pre-Hamiltonian function

$$H(t, x, p, u) = L(t, x, u) + \langle p, f(t, x, u)\rangle. \tag{1.1}$$

The n-vector p is called the costate.

Pontryagin's theorem states that every minimizing motion is a pseudo-extremal. This means that if $x^o(t)$ is an optimal motion, generated by the control function $u^o(t)$, then there exists also a comotion $p^o(t)$ such that

(1.2) $H(t, x^o(t), p^o(t), u) \geq H(t, x^o(t), p^o(t))$ for all $u \in U(t)$

along any optimal motion, and $(x^o(t), p^o(t))$ are solutions of the equations

$$\begin{aligned} dx/dt &= H_p(t, x, p, u^o(t)) \\ dp/dt &= -H_x(t, x, p, u^o(t)). \end{aligned} \tag{1.3}$$

Any pair $(x^o(t), p^o(t))$ which satisfies (2-3) is a pseudo-extremal, rather than an extremal, because in general one cannot prove that such motions are optimal over any interval of time.

A pseudo-extremal is regular if (1.2) holds with a strict inequality sign for all $u \neq u^o(t)$.* In that case, subject only to mild smoothness conditions on the set $U(t)$, equations (1.3) may be replaced by an equivalent set, which however does not involve $u^o(t)$,

$$\begin{aligned} dx/dt &= H^o_p(t, x, p) \\ dp/dt &= -H^o_x(t, x, p), \end{aligned} \tag{1.4}$$

where

*Although $p^o(t)$ is not uniquely defined by $x^o(t)$, it can be shown that the choice of $p^o(t)$ is immaterial as far as the definition of regularity is concerned.

$$(1.5) \qquad H^{o}(t, x, p) = \min_{u \in U(t)} H(t, x, p, u),$$

is the true (classical) Hamiltonian function of the problem. (1.4) are the well-known Euler-Lagrange-Hamilton equations.

With Carathéodory, an extremal is defined as a motion $\gamma^{o} = (x^{o}(t), p^{o}(t))$ which minimizes the integral $\int L\,dt$ with respect to any comparison curve that connects two points P_1 and P_2 on γ^{o}, provided P_1 and P_2 are sufficiently close. This is optimality over small intervals of time.

The principal theorem of the classical calculus of variations then asserts: Every regular pseudo-extremal is an extremal.

In applying the Hamiltonian theory to the problem of second variation, several authors (Bryson, Kelley, etc.) proceed by using equations (1.2-3), subject to a side condition which guarantees that the first variation vanishes. In the regular case, this is unnecessary if the formalism of (1.4-5) is used. In these equations, it is not even necessary to know what the variation of u has to be, since the values of u are implicitly determined once and for all by (1.5).

2. Concentrating attention on the regular case, that is, (1.4-5), it is well known that the problem of second variation may be studied advantageously by considering the so-called accessory (or Jacobi) variational problem, which has a quadratic Hamiltonian function. For this type of Hamiltonian, the variational problem has linear character, and there is an extensive and useful mathematical literature (see [1] for further references.)

To be more specific, let $\gamma^{o} = (x^{o}(t), p^{o}(t))$ be a regular extremal. If we consider small variations about γ^{o}, that is, the behavior of $\xi = x - x^{o}$ and $\pi = p - p^{o}$, we are led to

$$(2.1) \qquad \begin{aligned} d\xi/dt &= H^{o}_{px}\xi + H^{o}_{pp}\pi \\ d\pi/dt &= H^{o}_{xp}\xi - H^{o}_{xp}\pi \end{aligned}$$

where the matrices H^o_{px}, etc., are evaluated along the path γ^o and are in general functions of time. The regularity condition is equivalent to the fact that the matrix condition H^o_{pp} is non-positive definite. The Hamiltonian function associated with (2.1) is

$$\|\xi\|^2_{H^o_{xx}} + 2\langle \pi, H^o_{px}\xi\rangle + \|\pi\|^2_{H^o_{pp}}. \tag{2.2}$$

If we define

$$H^o_{px}(t) = F(t),\ H^o_{pp}(t) = -G(t)G'(t),\ H^o_{xx}(t) = Q(t),$$

then the Hamiltonian (2.2) and the Euler equations (2.1) are precisely those which result from the following variational problem: (i-ii) It is required to transfer the point (t_o, ξ_o) to the surface Σ_1, which is the tangent plane to S_1 at the point where γ^o meets S_1; (iii) the equations of motion are

$$d\xi/dt = F(t)\xi + G(t)\mu(t), \tag{2.3}$$

(iv) there are no restrictions on values of $\mu(t)$; (v) the Lagrangian is defined as

$$2L(t, \xi, \mu) = \|\xi\|^2_Q + \|\mu\|^2. \tag{2.4}$$

It is then easy to verify that the Hamiltonian H^o corresponding to (2.4) is precisely (2.2); moreover

$$u^o = -G'p. \tag{2.5}$$

To study the variational problem defined by (2.3) and (2.4), it is important to investigate the notion of controllability [2] of the system (2.3). This question, which concerns the accessory problem, also has a natural interpretation for the original variational problem.

3. In a celebrated paper [3] published in 1933, Caratheodory introduced a classification of regular extremals: Take a given (regular) extremal γ^0. Find all nearby extremals which terminate at P_1. These extremals will fill out a "manifold" of points which includes γ^0. Let the dimension of this manifold by $n + 1 - q$ $(q \geq 0)$. Then the integer q is the Caratheodory class of the extremal.

Variational problems of class 0 behave as the conventional variational problems without side conditions. On the other hand, all regular extremals of a Mayer problem are of class $q \geq 1$. Time optimal problems are often of class $q = n - 1$. Thus the Caratheodory class indicates the difficulty of a variational problem of the Lagrange type, i.e., of problems of interest in optimal control theory.

Carathéodory [3] has shown also that the class number of a regular extremal can be calculated by forming certain determinants related to the variational equations (2.1). It turns out that his technique is identical with methods used (independently) by the writer for the study of controllability and observability questions in linear systems [2]. The situations to which this technique is applied are different, but it is possible to connect the two notions. The following result is typical:

A (regular) extremal is of class $q = 0$ if and only if its variational equations are completely controllable.

The proof is essentially the same as was given in [4] for a slightly different problem.

In general, $n - q$ = dimension of the space of controllable states of (2.3).

A trivial, but practically necessary, extension of the Caratheodory classification is the following: Given a regular extremal γ^0 terminating on the surface S_1, let $n + 1 - q(s_1)$ be the dimension of the space filled out by nearby extremals which also

terminate on S_1. This defines the class $q(S_1)$ of γ^0 relative to S_1. Since it is easier to "hit" a surface than a point, in general we will have that $q(S_1) \geq q$. This concept has not been investigated so far.

4. The technique outlined above can be applied profitably to many problems of trajectory optimization. In fact, the study of the "class" often coincides with the study of the Jacobi determinant, which is necessary to locate conjugate points. By concentrating attention on the geometric properties of nearby extremals, one obtains a precise and intuitively appealing characterization of optimal trajectories.

Full details will be found in [5].

References

1. R. E. Kalman, The theory of optimal control and the calculus of variations, RIAS Report TR 61-3; to appear in Proc. of RAND-UC Conference on "Variational Problems in System Theory."

2. R. E. Kalman, Y. C. Ho, and K. S. Narendra, Controllability of Linear Dynamical Systems, Contr. to Diff. Eqs., Vol 1, 1963 (Interscience).

3. C. Caratheodory, "Uber die Einteilung der Variationsprobleme von Lagrange nach Klassen," Comm. Math. Helv., 5(1933) 1-19.

4. R. E. Kalman, discussion of paper by L. Markus and E. B. Lee, J. Basic Engr. (Trans. ASME), 84 D (1962) 21-22.

5. R. E. Kalman, "Caratheodory classes and the controllability properties of accessory variational problems," RIAS Report, to appear.

ASD-TDR-63-119

A Synthesis Method for Optimal Controls

By

L. W. Neustadt

Aerospace Corporation
El Segundo, California

A SYNTHESIS METHOD FOR OPTIMAL CONTROLS

Lucien W. Neustadt

Aerospace Corporation
El Segundo, California

ABSTRACT

A method is suggested for computing optimal controls. This method is based on the Pontryagin maximum principle, and consists in periodically computing the appropriate initial conditions of the linearized adjoint system as a function of the state of the system. The adjoint system is then used to determine the optimal control. Detailed computational procedures - using a steepest ascent method of successive approximations - are developed for some specific optimal problems including the "bang-bang" time-optimal and minimum control energy problems.

INTRODUCTION

One of the difficulties which arises in the implementation of optimal control systems is the complex control law that is generally required. In conventional control systems the controllers are made to be simple functions of the output (or state) of the object being controlled (the so-called plant). However, if the plant is to be controlled optimally, the relation existing at any time between the state of the plant and the value of the optimal control is both difficult to precompute and difficult to implement in an actual system.

For certain types of control systems, and certain optimality criteria, methods have been described for precomputing an optimal control (References 1 through 4). In each of these methods, the entire optimal control function can be computed once the initial and desired terminal states are specified. In principle, a system using these methods would operate as follows: the initial state of the system is measured and fed into a computer which calculates the time history of the optimal control, stores the latter, and drives the controller to follow the stored program. Such systems, once they begin to operate, are, of course, "open-loop"; i.e., no corrections are made to the control program based on measurements of the actual state of the plant. However, "feedback" can be introduced by periodically updating the control program by recomputing the optimal control at any time, considering the state of the plant at that time to be an "initial condition."

By suitably choosing our computational method, we shall describe one such procedure which appears to be particularly suitable to such "updating." Our method, which is based on the Pontryagin maximum principle, can in principle be applied to a wide class of optimal control systems. However, computational algorithms have so far been developed for only certain particular types of systems, which do, however, include many important special cases. These algorithms are described in Section III.

GENERAL DESCRIPTION OF THE SYNTHESIS PROCEDURE

We confine ourselves to control systems whose behavior can be described by a system of ordinary differential equations which have the form

$$\dot{x}_i = f_i (x_1, \ldots, x_n, u_1, \ldots, u_r, t), \quad i = 1, \ldots, n . \qquad (1)$$

The x_i describe the state of the system, and the u_j describe the state of the controllers. We assume the f_i to be continuous in all their arguments, and to be continuously differentiable with respect to the x_i and t. If we denote by x the vector $(x_1, \ldots, x_n)$, by u the vector $(u_1, \ldots, u_r)$, and by f the vector $(f_1, \ldots, f_n)$, Equations (1) can be written in vector form as

$$\dot{x} = f(x, u, t) . \qquad (2)$$

The control vector u may, of course, depend on the time, but we shall require $u(t)$ to be piecewise continuous for reasons of physical realizability. Finally, we shall suppose that the controllers are constrained and that this constraint can be described by the requirement that $u(t) \in \Omega$ for all t, where Ω is a set in r-space which is independent of t and x. We assume that there are no explicit constraints on the values of x.

We suppose that the aim of the control process is to transfer the system from one given state to another. That is, we are given two values x^0 and x^1 of the vector x, and an initial time, t_0; we wish to find an admissible control (i.e., a piecewise continuous function whose range is in Ω) $u(t)$ such that the solution of Equation (2) with $u = u(t)$, $x(t_0) = x^0$ takes on the value x^1 at some time $t_F > t_0$. The terminal time t_F may be fixed or free. The case where the target position depends on the time: $x^1 = x^1(t)$, can be put into the same form by defining a new variable $\bar{x} = x(t) - x^1(t)$ (assuming that $x^1(t)$ is twice continuously differentiable). Also, the aim of the process may be to attain a set rather than a point. In many cases, the treatment of the problem is then only slightly more complicated than if a point is to be obtained (e.g., see Reference 4, Section IV), and we shall confine ourselves to point targets.

Let us suppose that the optimality criterion is given by an integral of the form:

$$J = \int_{t_0}^{t_F} f_0 (x(t), u(t), t)\, dt \qquad (3)$$

where the function f_0 satisfies the same conditions as $f_1, \ldots, f_n$. Thus, our aim is to find an admissible control $u(t)$ which "transfers" the control system given by Equation (2) from a given state x^0 at the time $t = t_0$ to another state x^1 at the time t_F (which may be either fixed or free) in such

a way as to minimize the integral (3). Let us assume that such an optimal control exists. (The existence problem is itself of interest. For discussions thereof see Reference 5, § 19 and References 6 through 9.) Then, according to the Pontryagin maximum principle (Reference 5, Chapters I and II), there exists a nontrivial solution $\psi(t) = (\psi_0(t), \psi_1(t), \ldots, \psi_n(t))$ of the system of equations:

$$\begin{cases} \dot{\psi}_i = -\sum_{j=0}^{n} \dfrac{\partial f_j(x^*(t), u^*(t), t)}{\partial x_i} \psi_j, \quad i = 1, \ldots, n, \\ \psi_0 = \text{const.} \leq 0, \end{cases} \tag{4}$$

such that

$$\sum_{i=0}^{n} \psi_i(t) f_i(x^*(t), u^*(t), t) = \max_{u \in \Omega} \sum_{i=0}^{n} \psi_i(t) f_i(x^*(t), u, t) \tag{5}$$

at all points of continuity of $u^*(t)$ - where we have denoted the optimal control and optimal trajectory by $u^*(t)$ and $x^*(t)$, respectively.

Thus, if the initial conditions for the system (4) are known, it is possible to simultaneously solve Equations (1), (4), and (5) (provided that relation (5) determines $u^*(t)$ uniquely for t in a dense subset of (t_0, t_F)), and thereby compute the optimal control $u^*(t)$.

Let the point x^1 be fixed. Suppose that an optimal control exists for (x^0, t_0) in some region R in (x,t) space, and that for each initial value in R this optimal control is unique (if we disregard values at points of discontinuity). Then, for each such initial value, there is a solution $\psi(t)$ of Equations (4), with initial value ψ^0, which determines the optimal control $u^*(t; x^0, t_0)$ through relation (5). The corresponding solution $\psi(t)$ is not unique (it certainly can be multiplied by an arbitrary positive constant without affecting the maximum in Equation (5)). Let us denote by $H(x^0; t_0)$ the set of all $(n+1)$-vectors ψ^0 which have the following property: the control $u^*(t)$, determined by Equation (5) with $\psi(t)$ the solution of Equation (4) satisfying the initial condition $\psi(t_0) = \psi^0$, optimally transfers the control system from x^0 at the time t_0 to the state x^1.

If (x^0, t_0) is a fixed point in R, $x^*(t)$ and $u^*(t)$ are the corresponding optimal trajectory and control, and $u^*(t)$ is determined by a particular solution $\psi^*(t)$ of Equations (4) (so that $\psi^*(t_0) \in H(x^0; t_0)$), it is clear that $(x^*(\bar{t}), \bar{t}) \in R$ and

$$\psi^*(\bar{t}) \in H(x^*(\bar{t}); \bar{t}) \tag{6}$$

for any $\bar{t} \in (t_0, t_F)$ (since the portion of the control process for $t \geq \bar{t}$ must also be optimal).

We are now in position to describe the suggested synthesis procedure. At the initial time t_o, the initial state of the system x^o is measured. Depending on the target state x^1 and the integral functional J, one member ψ^o of the set $H(x^o; t_o)$ is found. Then, for a short time interval, Equations (1), (4), (5) are solved simultaneously - in real time - and the optimal control $u^*(t)$ is computed. At the end of this interval, say at the time t_1, the state of the system $x(t_1)$ is measured, and a member ψ^1 of the set $H(x(t_1); t_1)$ is found. The procedure is then repeated until the end of the process. The system thus operates similarly to a sampled-data feedback control system.

The principal difficulty in applying the above method lies in finding a member of the set $H(x; t)$. For certain particular optimal processes, as we shall describe below in Section III, a successive approximation method can be employed. Such a method is particularly attractive for the following reason. If $\psi^i \in H(x(t_i); t_i)$ is the initial value for system (4) in the interval (t_i, t_{i+1}), and if the corresponding solution of Equations (4) is $\psi^i(t)$, then, under ideal operation, $\psi^i(t_{i+1}) \in H(x(t_{i+1}); t_{i+1})$ because of relation (6). Thus, if $H(x; t)$ is continuous at $(x(t_{i+1}), t_{i+1})$ in an obvious sense, $\psi^i(t_{i+1})$ should be an excellent approximation to a member of $H(x(t_{i+1}); t_{i+1})$ even in the presence of disturbances.

During an interval (t_i, t_{i+1}) Equations (1), (4), (5) should be relatively easy to compute in "real time." For example, Equations (1) and (4) may be integrated on an analog computer. We assume, of course, that relation (5) uniquely determines $u^*(t)$. In many problems, it turns out that the maximum in this relation is a relatively simple function of x^*, ψ, and t. For example, this is true in the cases described in Section III below. In other problems, the maximum may have to be computed by more sophisticated techniques such as nonlinear programming.

SPECIFIC COMPUTATIONAL METHODS

In this section we shall describe successive approximation procedures for finding members of the sets $H(x; t)$ in some specific optimization problems. We shall omit the derivations of these methods since they have been presented elsewhere (see References 3 and 10).

First, consider the well-known "bang-bang" regulator problem. Let Equations (1) have the form

$$\dot{x} = A(t)\, x + B(t)\, u \qquad (7)$$

where $A(t)$ is an $n \times n$ matrix, $B(t)$ is an $n \times r$ matrix, and both A and B depend continuously on the time, t. Let Ω be the unit cube defined by $|u_i| \leq 1$, $i = 1, \ldots, r$; let the target x^1 be the origin: $x^1 = 0$; and let $f_o \equiv 1$. Thus, $J = t_F - t_o$, and the minimization is with respect to time (clearly, t_F is free).

Let $X(t)$ be the $n \times n$ matrix solution of the system

$$\dot{X}(t) = A(t)\, X(t) \;, \qquad X(t_o) = I, \text{ the identity} \;.$$

Relations (4) and (5) are now equivalent to

$$\dot{\psi}_i = - \sum_{j=1}^{n} a_{ji}(t)\, \psi_j(t) \ , \quad i = 1 \ , \ \ldots \ , \ n \ , \tag{8}$$

$$u_j^*(t) = \operatorname{sign}\left(\sum_{i=1}^{n} \psi_i(t)\, b_{ij}(t) \right) \ , \quad j = 1 \ , \ \ldots \ , \ r \ , \tag{9}$$

where $A = (a_{ij})$ and $B = (b_{ij})$. Let us assume that the zeros of $\Sigma_{i=1}^{n} \psi_i(t)\, b_{ij}(t)$ are isolated for every nontrivial solution of Equations (8) (and each $j = 1, \ldots, r$), so that Equation (9) uniquely determines a piecewise continuous control.

For a given x^o and t_o we wish to find an n-vector ψ^o such that relation (9) - with $(\psi_1(t), \ldots, \psi_n(t)) = \psi(t)$ the solution of Equations (8) with $\psi(t_o) = \psi^o$ - defines the time-optimal control which transfers the system given by Equation (7) from x^o at the time t_o to the origin. The set of all such vectors ψ^o will be denoted by $h(x^o; t_o)$.

We define a function $F(\eta; x^o, t_o)$, where η is an n-vector, such that F takes on its maximum value when, and only when, $\eta \in h(x^o; t_o)$. Namely, let

$$f(t, \eta; x^o, t_o) = \int_{t_o}^{t} \eta \cdot X^{-1}(s)\, B(s)\, u(s; \eta)\, ds + \eta \cdot x^o \ ,$$

where $u(s; \eta)$ is defined by

$$u_j(t; \eta) = \operatorname{sign}\left(\eta \cdot X^{-1}(t)\, b^j(t)\right) \ ,$$

b^j being the j-th column of B. Then, it is easy to see that f is a strictly increasing function of t (for fixed $\eta \neq 0$). We shall restrict ourselves to those η for which $\eta \cdot x^o < 0$, so that $f(0, \eta; x^o, t_o) < 0$. It can be shown that if t^* is the minimum time in which the origin can be attained, then $f(t^*, \eta; x^o, t_o) \geq 0$, where equality holds if, and only if, $\eta \in h(x^o; t_o)$. Thus, if we define $F(\eta; x^o, t_o)$ as the solution of the equation

$$f(F(\eta; x^o, t_o), \eta; x^o) = 0 \ ,$$

F has the desired properties, the maximum of F being the minimum transfer time t^*.

Finally, the maximum of F can be computed directly by the method of steepest ascent, since the gradient of $F(\eta)$ is proportional to

$$\int_{t_o}^{F(\eta)} X^{-1}(s)\, B(s)\, u(s;\eta)\, ds + x^o \quad .$$

Also, the gradient is different from zero if $\eta \neq h$, so that F has no local maxima away from h.

Generalizations of this problem, wherein Ω is replaced by an arbitrary convex, compact set, and the origin is replaced by an arbitrary target point x^1, are described in References 3 and 10.

Computer solutions for the problem described above, using both analog and digital computers, have been described by Paiewonsky and Williamson (see Reference 11).

As a second example, consider a control system whose behavior is again described by Equation (7), where Ω is as before, the target x^1 is arbitrary, the terminal time t_F is fixed, and f_o is given by

$$f_o(u) = \sum_{j=1}^{r} \lambda_j \left| u_j \right|^p \quad , \tag{10}$$

$p \geq 1$ and $\lambda_j \geq 0$ (but not all zero) being constants. Then,

$$J = \int_{t_o}^{t_F} \sum_{j=1}^{r} \lambda_j \left| u_j(t) \right|^p dt \quad ,$$

and the physical meaning of J, particularly if $p = 1$ or 2, is clear. Relations (4) are again equivalent to Equations (8). With little loss of generality, assume that ψ_o in Equation (4) is equal to -1. Then, relation (5) takes the form

$$u_j^*(t) = \begin{cases} \left| \zeta_j(t)/p\lambda_j \right|^{1/(p-1)} \cdot \operatorname{sign} \zeta_j(t) & \text{if } \left| \zeta_j(t) \right| \leq p\lambda_j \text{ and } \lambda_j \neq 0 \quad , \\ \operatorname{sign} \zeta_j(t) & \text{if } \left| \zeta_j(t) \right| > p\lambda_j \text{ and } \lambda_j \neq 0 \quad , \\ \operatorname{sign} \zeta_j(t) & \text{if } \lambda_j = 0 \quad , \end{cases} \tag{11}$$

for $p > 1$, and

$$u_j^*(t) = \begin{cases} 0 & \text{if } \left| \zeta_j(t) \right| < \lambda_j \\ \operatorname{sign} \zeta_j & \text{if } \left| \zeta_j(t) \right| > \lambda_j \end{cases} \tag{12}$$

if $p = 1$. Here, $\zeta_j(t) = \Sigma_{i=1}^{n} \psi_i(t)\, b_{ij}(t)$. We assume that the zeros of $\zeta_j(t)$ are isolated if $p > 1$ and $\lambda_j = 0$, and that the zeros of $\zeta_j(t) \pm \lambda_j$ are isolated if $p = 1$, for every $j = 1, \ldots, r$, for any nonzero solution of Equations (8).

We define the set of n-vectors $h(x^o; t_o)$ in a manner analogous to before, and we define a function $g(\eta; x^o, t_o)$ which takes on its maximum value when, and only when, $\eta \epsilon h(x^o; t_o)$. Namely, let

$$g(\eta; x^o, t_o) = \eta \cdot X^{-1}(t_F)\, x^1 - \int_{t_o}^{t_F} \left[\eta \cdot X^{-1}(t)\, B(t)\, \hat{u}(t;\eta) + f_o(\hat{u}(t;\eta))\right] dt - \eta \cdot x^o ,$$

where $\hat{u}(t;\eta)$ is given by Equation (5) (or Equations (11) or (12)) with $\psi_1(t), \ldots, \psi_n(t)$ equal to the components of $\eta \cdot X^{-1}(t)$. (It is well-known that the latter vector is the solution of Equations (8) satisfying the initial values $\psi_i(t_o) = \eta_i$.) Furthermore, the maximum of g can be computed by the method of steepest ascent using the fact that

$$\text{grad}\, g(\eta) = - x^o + X^{-1}(t_F)\, x^1 - \int_{t_o}^{t_F} X^{-1}(t)\, B(t)\, \hat{u}(t;\eta)\, dt ,$$

and noting that g has no local maxima away from h.

The relations above are derived in Reference 10, wherein more general functions f_o are also treated.

The case where f_o is defined by

$$f_o(u) = \left\| u(t) \right\| ,$$

which arises in minimum fuel guidance problems, is described extensively in Reference 12.

As a final example, consider again the case where f_o is defined by Equation (10), where now $p > 1$, every λ_j is positive, and Ω is the entire r-dimensional space (so that there are no restrictions on the values of u). Then, the treatment differs from the one above only in that Equation (11) is replaced by

$$u_j^*(t) = \left| \zeta_j(t) / p\lambda_j \right|^{1/(p-1)} \cdot \text{sign}\, \zeta_j(t) ,$$

and the definition of $\hat{u}(t;\eta)$ is changed accordingly.

If variations of the above examples are considered, it is often easy to construct functions having properties similar to those of F and g which lead to a successive approximation method of finding the appropriate initial conditions to Equations (4).

REFERENCES

1. Balakrishnan, A. V., An Operator Theoretic Formulation of a Class of Control Problems and a Steepest Descent Method of Solution, Report, A-62-1730-202, Aerospace Corporation, El Segundo, California.

2. Ho, Y. C., "A Successive Approximation Technique for Optimal Control Systems Subject to Input Saturation," J. of Basic Eng., Vol 84 Ser. D., No. 1, pp 33-40 (1962).

3. Neustadt, L. W., "Synthesizing Time Optimal Control Systems," J. of Math. Anal. and Applic., Vol 1, No. 4, pp 484-493 (1960).

4. Neustadt, L. W., "Minimum Effort Control Systems," J. of SIAM Series A on Control, Vol 1, No. 1 (1962).

5. Pontryagin, L. S., V. G. Boltyanskii, R. V. Gamkrelidze, E. F. Mishchenko, The Mathematical Theory of Optimal Processes, Interscience Division of John Wiley and Sons, New York, N. Y., (an English translation) 1962.

6. Filippov, A. F., "On Certain Questions in the Theory of Optimal Control," Vestnik Moskovsk. Univers., Ser. Mat., Mekh., Astr., Fiz., Khim., No. 2 (1959), pp 25-32. English translation in J. of SIAM Series A on Control, Vol 1, No. 1 (1962).

7. Lee, E. B. and L. Markus, "Optimal Control for Nonlinear Processes," Archive for Rat. Mech. and Anal., Vol 8, No. 1, pp 36-58 (1961).

8. Roxin, E., "The Existence of Optimal Controls," Mich. Math. J., Vol 9, pp 109-119 (1962).

9. Warga, J., Relaxed Variational Problems, J. of Math. Anal. and Applic., Vol 4, No. 1, pp 111-128 (1962).

10. Neustadt, L. W., On Synthesizing Optimal Controls, to be presented to IFAC Congress, Basle, Switzerland in 1963. Also available as Report A-62-1732. 1-2, Aerospace Corporation, El Segundo, California.

11. Paiewonsky, B. H., and G. G. Williamson, Synthesis of Time Optimal Control for Linear Systems, Report of Aeronautical Research Associates of Princeton, Princeton, N. J.

12. Meditch, J., "Optimal Thrust Programming for Minimal Fuel Midcourse Guidance, presented to Optimum System Synthesis Conference, Wright-Patterson AFB, Dayton, Ohio, 1962.

ASD-TDR-63-119

On Optimal Control for Systems with Numerator Dynamics

By

C. A. Harvey
E. B. Lee
L. Markus

Minneapolis-Honeywell Regulator Company

and

University of Minnesota, Minneapolis, Minn.

ON OPTIMAL CONTROL FOR SYSTEMS WITH NUMERATOR DYNAMICS+

C. A. Harvey* E. B. Lee* L. Markus**

*Minneapolis-Honeywell Regulator Company, Minneapolis, Minn.
**University of Minnesota, Minneapolis, Minnesota

INTRODUCTION

Let us consider an example which illustrates the concepts we wish to discuss.

The small amplitude pitch motions of many aircraft can be summarized approximately by solutions of the equations

(a) $\dot{\alpha} - k_1 \alpha - \dot{\theta} = k_2 \delta$ (Flight Path equation)

(b) $\ddot{\theta} - k_3 \dot{\theta} - k_4 \alpha - k_5 \dot{\alpha} = k_6 \delta$ (Pitch equation)

(c) $\dot{\delta} = u$ (Servo equation)

Here

$\alpha \sim$ aircraft fuselage reference angle of attack
$\theta \sim$ pitch attitude change
$\delta \sim$ elevator deflection
$u \sim$ control variable, and k_i = constants.

We shall assume, as is natural, $|u| \leq 1$.

Consider the relationship between u and $\dot{\theta}$. Letting $x=\dot{\theta}$ and eliminating α, and $\dot{\alpha}$ from (a), (b), and the derivative of (b) gives

$$(d)\quad \ddot{x} + (-k_3-k_5-k_1)\dot{x} + (-k_4-k_1-k_3)x = -k_6\left(\frac{k_2k_5}{k_6}+1\right)\dot{\delta} + \left(-k_1+\frac{k_2k_4}{k_6}\right)\delta$$

Differentiating (d) and using (c) to eliminate δ gives an equation of the form

$$(e)\quad \dddot{x} + a\ddot{x} + b\dot{x} = c\dot{u} + du$$

+The results reported here were achieved at the Minneapolis-Honeywell Regulator Company, under sponsorship of NASA Contract NASr 27.

The transfer function for $\frac{x}{u}$ is $\frac{cs + d}{s(s^2+as+b)}$, from which the term numerator dynamics stems.

If we consider (e) to be the equation of motion we run into certain technical difficulties. For example, how one defines a solution if the control $u = u(t)$ was composed of a series of step functions. There are now standard procedures for transforming equations of the form of (e) into a system of equations in which the derivatives of the control variable do not appear. (See for example, Laning and Battin, "Random Processes in Automatic Control", p. 191, or just reverse the above procedure.) We prefer to work with the system (a) - (c).

A typical control problem is to reduce the pitch rate, $\dot{\theta} = x$, to zero in finite time and in such a manner that it will remain zero if no additional disturbances appear. Actually we should impose certain conditions on the other variables, for example, require $|\delta| \leq \epsilon$ (the bounded phase coordinate problem).

In section 2 we consider the various modes of control. It is shown that most optimal control problems are included in the problem of controlling only one state vector component after choosing a suitable coordinate system. We have shown that such problems arise in a natural way when considering system transfer functions with numerator dynamics. Section 3 is devoted to necessary and sufficient conditions for time optimal control. Here we show that control of one state variable is equivalent to controlling to a closed convex set of possible states.

Section 4 is devoted to considerations of constructing the time optimal feedback controller.

MODES OF CONTROL

Consider a physical system as represented by the equations‡

$$\dot{x}^i = a^i_j x^j + b^i u; \; i, j = 1,2,\ldots,n; \tag{1}$$

or in vector form

$$\dot{x} = Ax + bu$$

here x is the system state and u the control variable.

We wish to determine a bounded scalar function $u = u(x)$ such that some or all of the solution components $x^i(t)$, $i = 1, 2,\ldots,n$ of (1) behave in a desired manner on the finite time interval $[0,\tau]$.

‡ Repeated index indicates summation with respect to that index.

We shall adhere to the following definition:
Multiple component control is defined as control of less than n of the state variables by bringing them to zero in finite time and holding them at zero thereafter.
Single component control is defined as control of a single state variable by bringing this variable to zero in finite time and holding it zero thereafter.

It is seen that single component control is a special case of multiple component control. Time-optimal multiple component control is defined in the obvious way when $u(t)$ is bounded; namely, bring the state variables to be controlled to zero in the minimum time such that they may be held at zero thereafter with $u(t)$ satisfying the bounding constraint. Time-optimal single component control is discussed in section 3.

It is now our purpose to show that multiple component control can be accomplished by single component control.

As in [1] suppose that for a system described by (1) it is desired to control the variables $x^1, x^2, \ldots, x^m$, $1 \leq m \leq n$. If $m = 1$, the problem is the single component problem of controlling x^1. Thus, assume $m > 1$ and for convenience introduce the following notation:

$$\xi^1 = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^m \end{bmatrix} \qquad \xi^2 = \begin{bmatrix} x^{m+1} \\ x^{m+2} \\ \vdots \\ x^n \end{bmatrix} \qquad \beta^1 = \begin{bmatrix} b^1 \\ b^2 \\ \vdots \\ b^m \end{bmatrix} \qquad \beta^2 = \begin{bmatrix} b^{m+1} \\ b^{m+2} \\ \vdots \\ b^n \end{bmatrix}$$

$$A^1 = \begin{bmatrix} a_1^1 & \cdots & a_m^1 \\ \vdots & & \\ a_1^m & \cdots & a_m^m \end{bmatrix} \qquad A^2 = \begin{bmatrix} a_{m+1}^1 & \cdots & a_n^1 \\ \vdots & & \vdots \\ a_{m+1}^m & \cdots & a_n^m \end{bmatrix}$$

$$A^3 = \begin{bmatrix} a_1^{m+1} & \cdots & a_m^{m+1} \\ \vdots & & \\ a_1^n & \cdots & a_m^n \end{bmatrix} \qquad A^4 = \begin{bmatrix} a_{m+1}^{m+1} & \cdots & a_n^{m+1} \\ \vdots & & \\ a_{m+1}^n & \cdots & a_n^n \end{bmatrix}$$

Then (1) becomes

$$\dot{\xi}^1 = A^1 \xi^1 + A^2 \xi^2 + \beta^1 u \qquad (2)$$
$$\dot{\xi}^2 = A^3 \xi^1 + A^4 \xi^2 + \beta^2 u$$

and it is desired that ξ^1 be controlled.

Holding ξ^1 zero requires that $\xi^1 = \dot{\xi}^1 = 0$, in which case (2) becomes

$$0 = A^2 \xi^2 + \beta^1 u \qquad (3)$$
$$\dot{\xi}^2 = A^4 \xi^2 + \beta^2 u$$

If $\beta^1 = 0$ and $A^2 = 0$, it is impossible to control ξ^1 because the first equation of (2) becomes $\dot{\xi}^1 = A^1 \xi^1$ and the proposed problem is of no interest. If $\beta^1 = 0$ and $A^2 \neq 0$, it may be assumed without loss of generality that the first row of A^2 contains a non-zero element. Then the state variable

$$y^{m+1} = \sum_{j=m+1}^{n} a_j^1 x^j$$

is necessarily controlled when ξ^1 is controlled. Thus, the original problem can be reformulated so that it becomes a problem of controlling m+1 state variables, i.e. $x^1, x^2, \ldots, x^m$ and y^{m+1}.

Hence, consider the case when $\beta^1 \neq 0$ and assume without loss of generality that $b^m \neq 0$. Making a transformation

$$y = Sx, \text{ where } S = \begin{bmatrix} S^1 & 0 \\ 0 & I^{n-m} \end{bmatrix}$$

$$S_1 = \begin{bmatrix} b^m & 0 & \cdots & 0 & -b^1 \\ & b^m & \cdots & 0 & -b^2 \\ & & \cdots & & \cdots \\ & & & b^m & -b^{m-1} \\ & & & & 1 \end{bmatrix}$$

and setting

$$\zeta^1 = \begin{bmatrix} y^1 \\ \vdots \\ y^m \end{bmatrix} = S^1 \xi^1 \quad \text{and} \quad \zeta^2 = \begin{bmatrix} y^{m+1} \\ \vdots \\ y^m \end{bmatrix} = \begin{bmatrix} x^{m+1} \\ \vdots \\ x^n \end{bmatrix}$$

equations (2) become

$$\begin{aligned}\dot{\zeta}^1 &= S^1 A^1 (S^1)^{-1} \zeta^1 + S^1 A^2 \zeta^2 + S^1 \beta^1 u \\ \dot{\zeta}^2 &= A^3 (S^1)^{-1} \zeta^1 + A^4 \zeta^2 + \beta^2 u\end{aligned} \tag{4}$$

S^1 was chosen so that

$$S^1\beta^1 = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ b^m \end{bmatrix}$$

If the kth row, $k < m$, of $S^1 A^2$ has non-zero elements, the control of ζ^1 implies the control of the kth component of $S^1 A^2 \zeta^2$ and the original problem may be reformulated so that it becomes a problem of controlling $m + 1$ state variables.

From the foregoing considerations, it is clear that if it is possible to control ξ^1 only, then (4) may be obtained with

$$S^1\beta^1 = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ b^m \end{bmatrix} \quad S^1A^2 = \begin{bmatrix} 0 & \cdots & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & & 0 \\ a_{m+1}^{m(1)} & \cdots & a_n^{m(1)} \end{bmatrix} \quad S^1A^1(S^1)^{-1} = \begin{bmatrix} a_1^{1(1)} & \cdots & a_m^{1(1)} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_1^{m(1)} & \cdots & a_m^{m(1)} \end{bmatrix}$$

Now if each $a_m^{i(1)} = 0$ with $i < m$, the components $y^1, y^2, \ldots y^{m-1}$ of ζ^1 are not controllable; hence ζ^1 and consequently ξ^1, are not controllable. Thus it may be assumed without loss of generality that $a_m^{m-1(1)} \neq 0$. Define

$$S^2 = \begin{bmatrix} a_m^{m-1(1)} & 0 & \cdots & 0 & -a_m^{1(1)} & 0 \\ & a_m^{m-1(1)} & \cdots & 0 & -a_m^{2(1)} & 0 \\ & & & \vdots & \vdots & \vdots \\ & & & a_m^{m-1(1)} & -a_m^{m-2(1)} & 0 \\ & & & & 1 & 0 \\ & & & & & 1 \end{bmatrix}$$

and denote the ijth element of $S^2\ S^1\ A^1 [S^1]^{-1}\ [S^2]^{-1}$ by $a_j^{i(2)}$. Then $a_m^{i(2)} = 0$, $i < m-1$ and $a_m^{m-1(2)} = a_m^{m-1(1)} \neq 0$. Now if each $a_{m-1}^{i(2)} = 0$ with $i < m-1$, the first $m-2$ components of $S^2\ \zeta^1$ are not controllable and hence ξ^1 is not controllable. Thus it may be assumed without loss of generality that $a_{m-1}^{m-2(2)} \neq 0$. Define

$$S^3 = \begin{bmatrix} a_{m-1}^{m-2(2)} & 0 & \cdots & 0 & a_{m-1}^{1(2)} & 0 & 0 \\ & a_{m-1}^{m-2(2)} & \cdots & 0 & -a_{m-1}^{2(2)} & 0 & 0 \\ & & \ddots & \vdots & \vdots & \vdots & \vdots \\ & & & a_{m-1}^{m-2(2)} & -a_{m-1}^{m-3(2)} & 0 & 0 \\ & & & & 1 & 0 & 0 \\ & & & & & 1 & 0 \\ & & & & & & 1 \end{bmatrix}$$

Continuing this process, $S^* = S^{m-1}\ S^{m-2} \cdots S^2\ S^1$ is determined so that

$S^* A^1 (S^*)^{-1}$ has the form $\begin{bmatrix} \alpha_1^1 & \alpha_2^1 & 0 & -- & -- & 0 \\ \alpha_1^2 & \alpha_2^2 & \alpha_3^2 & -- & -- & -- \\ -- & -- & -- & -- & -- & -- \\ -- & -- & -- & -- & -- & 0 \\ \alpha_1^{m-1} & -- & -- & -- & -- & \alpha_m^{m-1} \\ \alpha_1^m & -- & -- & -- & -- & \alpha_m^m \end{bmatrix}$

with $\alpha_{i+1}^i \neq 0$; $i = 1,2,\ldots,m-1$. $S^* A^2 = S^1 A^2$ and $S^*\ \beta^1 = S^1 \beta^1$.

Letting $z = \begin{bmatrix} S^* & 0 \\ 0 & I^{n-m} \end{bmatrix} x$ and $\zeta^3 = \begin{bmatrix} z^1 \\ \vdots \\ z^m \end{bmatrix}$ $\zeta^4 = \begin{bmatrix} z^{m+1} \\ \vdots \\ z^n \end{bmatrix}$

(2) becomes

$$\begin{aligned} \dot{\zeta}^3 &= S^*\ A^1 (S^*)^{-1}\ \zeta^3 + S^*\ A^2\ \zeta^4 + S^* \beta^1 u \\ \dot{\zeta}^4 &= A^3 (S^*)^{-1}\ \zeta^3 + A^4\ \zeta^4 + \beta^2 u \end{aligned} \qquad (5)$$

rom (5),

$$z^{k+1} = \frac{1}{\alpha^k_{k+1}}\left[\dot{z}^k - \sum_{j=1}^{k} \alpha^k_j z^j\right], \quad k = 1,2,\ldots,m-1$$

With the aid of this recursive relation, it is possible to obtain z^k as a linear combination of z^1 and its first k-1 derivatives for $k \leq m$. Thus, controlling z^1 accomplishes control of ξ^1.

NECESSARY AND SUFFICIENT CONDITIONS FOR TIME OPTIMAL CONTROL

In section 2 it was shown that any multiple-component time-optimal control problem for a linear system with a single control variable is equivalent to a single-component time optimal control problem. For this reason, only single-component control problems need be considered.

Suppose $x^1(t)$ is the variable which we wish to bring to zero and hold at zero in minimum time T with $|u| \leq 1$. If $x^1(t)$ is to be zero for $t > T$ then $\dot{x}^1(t) = 0$ for $t > T$. Thus $a^1_j x^j(t) + b^1 u(t) = 0$ for $t > T$.

Let G be the set of points $x_{10}, x_{20},\ldots,x_{n0}$ in Euclidean n-space such that the solution of 1) with $u^1(0) = x_{10} = 0$, $x^j(0) = x_{j0}$ satisfies $a^1_j x^j(t) + b^1 u(t) = 0$ for all $t > 0$ for some $u(t)$ with $|u(t)| \leq 1$.

Assume* $b^1 \neq 0$; thus we can solve for $u(t) = \dfrac{-a^1_j x^j(t)}{b^1}$ and substitute this result into equation (1) to obtain the homogeneous equation

$$\dot{x}^1 = 0$$
$$\dot{x}^i = a^i_j x^j - \frac{b_i a^1_j}{b^1} x^j, \quad i = 2,3,\ldots,n. \tag{h}$$

A point $p = (x_{10}, x_{20},\ldots,x_{n0})$ is in G if the solution $x_h(t)$ of the homogeneous equation satisfies

$$\left|\frac{a^1_j x^j_h(t)}{b^1}\right| = \left|\frac{a^1_j \varphi^j_k(t) x_{k0}}{b^1}\right| \leq 1 \text{ for all } t \geq 0,$$

---

* If $b^1 = 0$, which occurs frequently, we obtain the condition that $a^1_j x^j = 0$ all $t \geq 0$ which implies

$$\frac{d(a^1_j x^j)}{dt} = 0, \text{ all } t \geq 0.$$

In this manner we eventually obtain a condition for the control as a function of the state variable and reach the same conclusion about optimum control to the set G.

here the matrix $(\varphi_k^j(t))$ is the fundamental solution matrix of the homogeneous equation (h).

Let $J(t)$, $t > 0$, be the set of points $(x_{10} = 0, x_{20}, \ldots x_{n0})$ such that

$$\left| \frac{a_j^1 \varphi_k^j(t) x_{k0}}{b^1} \right| \leq 1$$

thus $G = \bigcap_{t \in [0,\infty)} J(t)$, which is thereby closed, convex and easily shown to be nonempty.

We conclude that control of one component of the state vector, is equivalent to control of the state vector to the closed, convex set $G \subset R^n$.

Consideration is now given the following problem: construct $u = u(t)$, with $|u| \leq 1$, such that $x(t) = (x^1(t), x^2(t), \ldots, x^n(t))$ moves from $x(0) = x_0$ to an intersection with the closed convex target set G in minimum time T.

For the remaining discussion assume the system 1) is normal, that is, the vectors $b, Ab, \ldots, A^{n-1}b$ are linearly independent.

If 1) is normal it is easily proved [2] that if $u(t)$ is an optimal control which steers $x(t)$ from x_0 to G it is the only optimum control (in the sense of being equal almost everywhere).

We call $u(t)$ an extremal control if

$$u(t) = \operatorname{sgn} \left\{ \eta_0 \, e^{-At} \, b \right\}$$

and $\eta_0 e^{-At}$ is an interior normal to G at the point $x(t)$, whenever the response $x(t)$ is a point of the boundary of G. Here η_0 is a row vector of unit length.

It is well known [2] that if $\hat{u}(t)$ is an extremal control for the normal system 1) which steers $x(t)$ from x_0 to G and G is such that a nonextremal control exists so that each solution of 1) which intersects G remains thereafter in G, then $\hat{u}(t)$ is the unique optimal control.

This is in effect a necessary and sufficient condition, because optimal control is necessarily extremal control.

SYNTHESIS OF TIME OPTIMUM FEEDBACK CONTROL

It is now our purpose to show that knowledge of the optimum control $u(t)$ for various initial states will enable one to construct the feedback controller $u = u(x(t))$. Before proceeding with this consider the question of existence of time optimal control.

It has been demonstrated [3] that if there exists even one allowable control which steers $x(t)$ from x_0 to G for the normal system 1) there will exist an optimum control. Obviously if the normal system 1) is stable (for example the roots of $|A - \lambda I| = 0$ have negative real parts) and G contains the origin, optimal control exists for all initial states. Since anything which can be done with allowable control can also be done with optimal control, the synthesis procedure to be outlined will provide the domain of controllability, that is, the set of initial states from which the set G can be reached using allowable control.

The necessary and sufficient condition discussed in section 3 is of interest in synthesizing the optimum feedback controller. The procedure used is the familiar one of running backwards in time from points of the boundary of the target G.

Starting at $x(0)$ in boundary of G with the extremal control $u(t) = \operatorname{sgn}\{\eta_0 e^{-At}b\}$, where η_0 is normal to a support plane of G at $x(0)$ and directed into G we consider the corresponding response $x(t)$ as $t \longrightarrow \tau$, $-\infty < \tau < 0$. All such states $x(t)$ which can be reached in the above manner determine the domain of controllability. Moreover along such optimal trajectories the control u can be specified as $u(x)$, the required feedback control function.

To illustrate this further consider an example:

$$\dot{x}^1 = x^2 + u$$

$$\dot{x}^2 = -x^1 + u$$

with $|u| \leq 1$. Suppose it is required that $x^1(t)$ go to zero and remain so in minimum time. We see easily that G is just the line segment: $|x^2| \leq 1$, $x^1 = 0$. For this example $\eta_0 e^{-At}b = \eta_0{}^1(\cos t - \sin t) + \eta_0{}^2(\sin t + \cos t)$ and the points of interest are those on trajectories with $x(0) \in G$ and $u(t) = \operatorname{sgn}(\eta_0 e^{-At}b)$ which corresponds to values of $t < 0$ where $u(t)$ changes sign, (η_0 satisfying the transversality condition at $x(0)$). For $|x^2(0)| < 1$, η_0^2 must be zero and the zeros of $\cos t - \sin t$ are at $\frac{\pi}{4} - n\pi$, $n = 1,2,\ldots$. Each of these values of t gives rise to a mapping of the line segment G (through the differential equations) onto another line segment. These line segments are joined by sets of arcs of

circles which correspond to values of t for which $\eta_o e^{-At_b}$ is zero for the sets of η_o allowed at the points of G where $|x^2(0)| = 1$. All of these point sets when taken with the set G divide the phase plane into two parts. In one part $u(x) = 1$ and in the other $u(x) = -1$ is the optimal feedback control. This is illustrated in figure 1.

REMARKS

The results presented here, after slight technical modification [4], apply equally well to systems described by recurrence equations (sampled data systems).

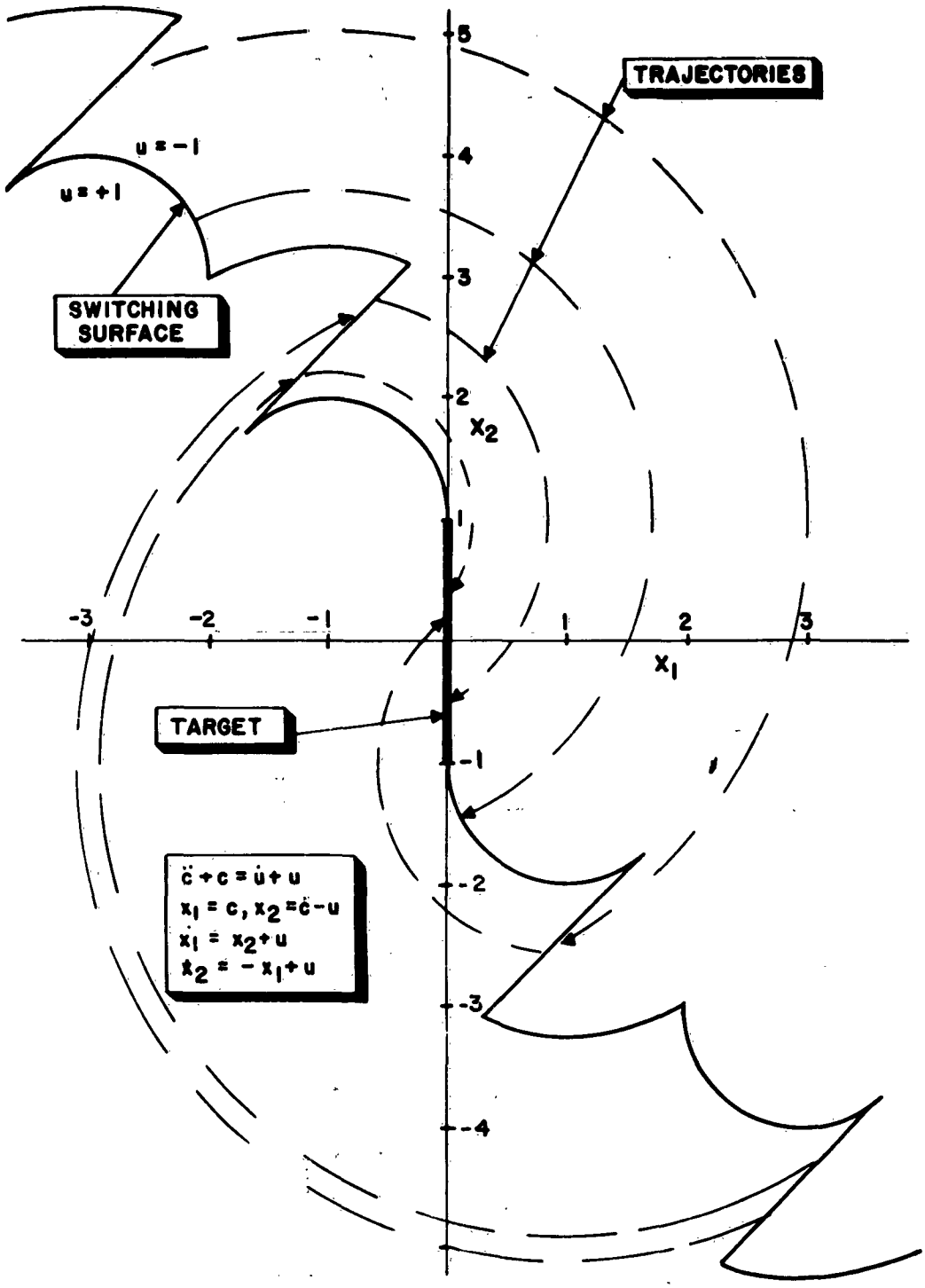


Figure 1. Switching Boundary for Time Optimal Control Example

REFERENCES

1. C. A. Harvey Modes of Control (Adaptive State Vector Control), Minneapolis-Honeywell Tech. Report MPG 1529-TR2, June 1962.

2. C. A. Harvey and E. B. Lee, "On Necessary and Sufficient Conditions for Time Optimal Control of Linear Systems," Jour. of Math Anal. and App. 1962.

3. E. B. Lee and L. Markus, "On the Existence of Optimal Control," ASME Jour. of Basic Engr. 1962.

4. E. B. Lee "Recurrence Equations and the Control of Their Evolution," Jour. of Math. Analysis and Application.

Aeronautical Systems Division, Dir/Aeromechanics, Flight Control Laboratory, Wright-Patterson AFB, Ohio.
Rpt Nr ASD-TDR-63-119, PROCEEDINGS OF THE OPTIMUM SYSTEM SYNTHESIS CONFERENCE 11 - 13 SEPTEMBER 1962, Feb 63, 390 p. incl illus., tables, and refs.

Unclassified Report

The proceedings contain a collection of 16 papers presented at the Optimum System Synthesis Conference held at the Aeronautical Systems Division 11 - 13 September 1962. The meeting was directed toward defining the present position of optimum

(over)

system synthesis and determining guides for future research in both applications and theory. Most papers are concerned with various aspects of recent applications and theoretical developments in optimal control such as steepest descent techniques, suboptimal controllers, optimum filtering, and functional analysis techniques. Some earlier results are also discussed.

1. Control theory, systems & optimal
2. Synthesis techniques
3. Optimization techniques and methods

I. AFSC Project 8219, Task 821904

II. L. Schwartz, Conference Monitor

III. Aval fr OTS

IV. In ASTIA collection

Aeronautical Systems Division, Dir/Aeromechanics, Flight Control Laboratory, Wright-Patterson AFB, Ohio.
Rpt Nr ASD-TDR-63-119, PROCEEDINGS OF THE OPTIMUM SYSTEM SYNTHESIS CONFERENCE 11 - 13 SEPTEMBER 1962, Feb 63, 390 p. incl illus., tables, and refs.

Unclassified Report

The proceedings contain a collection of 16 papers presented at the Optimum System Synthesis Conference held at the Aeronautical Systems Division 11 - 13 September 1962. The meeting was directed toward defining the present position of optimum

(over)

system synthesis and determining guides for future research in both applications and theory. Most papers are concerned with various aspects of recent applications and theoretical developments in optimal control such as steepest descent techniques, suboptimal controllers, optimum filtering, and functional analysis techniques. Some earlier results are also discussed.

1. Control theory, systems & optimal
2. Synthesis techniques
3. Optimization techniques and methods

I. AFSC Project 8219, Task 821904

II. L. Schwartz, Conference Monitor

III. Aval fr OTS

IV. In ASTIA collection