J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

AN ALGORITHM FOR THE ITERATIVE SOLUTION OF A CLASS OF TWO-POINT BOUNDARY VALUE PROBLEMS*

C. W. MERRIAM III†

Abstract. The algorithm, which is based on second variations, is intended for a class of two-point boundary value problems arising in control optimization. These optimization problems are characterized by positive definite second variations, the absence of point constraints on control and state variables, and free-point terminal boundary conditions. In a suitably small neighborhood of the optimal trajectory, the algorithm gives one-step convergence within the limits of the accuracy obtained with numerical integration. The relationships which are used here and arise in variational mathematics are stated in an appendix.

Introduction. The computational aspects of the two-point boundary value problem arising in control optimization and other variational problems recently have received considerable attention. In this paper, the variational problem of interest is the minimization of

$$e = \int_0^T f_0(\mathbf{x}, \mathbf{m}, t)\, dt \tag{1}$$

with respect to $\mathbf{m}$ where

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{m}, t) \tag{2}$$

and

$$\mathbf{x}(0) = \mathbf{a}. \tag{3}$$

The control and state vectors are taken to be $\mathbf{m} = \text{col}\,(m_1, m_2, \cdots, m_M)$ and $\mathbf{x} = \text{col}\,(x_1, x_2, \cdots, x_N)$ respectively. The vector function $\mathbf{f}(\mathbf{x}, \mathbf{m}, t)$, $f_0(\mathbf{x}, \mathbf{m}, t)$, and the required partial derivatives of these functions are assumed to be continuous and bounded for all finite values of their arguments. In addition, the assumption is made that $f_0(\mathbf{x}, \mathbf{m}, t)$ is formed properly such that the optimal $\mathbf{m}$ is unique, bounded, and a continuous function of $\mathbf{a}$ and t.

For this class of variational problems, Kelley [1] and Bryson [2] have provided a feasible and straightforward method,‡ variously called gradient or steepest-ascent, for obtaining numerical solutions to these variational problems. The method due to Kelley and Bryson offers many practical advantages. Specifically, the iterative procedure guarantees a monotone

* Received by the editors December 1, 1962 and in revised form January 1, 1963.

† General Electric Research Laboratory, Schenectady, New York.

‡ This method requires fewer restrictions than imposed here.

decreasing sequence* of values for e, and every suboptimal trajectory obtained satisfies (2) and (3) so that the iterative procedure can be terminated when an efficient trajectory has been obtained. Also, the computer memory requirements are linear in N and M whereas they would be exponential in N for the approach of discrete dynamic programming. Finally, the computations are performed with stable differential equations, unless (2) is unstable, whereas boundary condition iteration methods using the equations arising from the calculus of variations always involve unstable equations and the ensuing numerical difficulties.

More recently, refinements of the basic method due to Kelley and Bryson have been introduced [3] which primarily are intended to alleviate two remaining difficulties. First, experience with the basic method indicates that the convergence of e to the minimum value of e is considerably more rapid than the convergence of $\mathbf{m}$ to the optimal control vector [4]. Second, the rate of convergence tends to decrease in the neighborhood of the optimal trajectory. The algorithm presented here also is directed toward these two difficulties for the restricted class of variational problems discussed previously and is a direct extension of the basic method due to Kelley and Bryson.

Condition for a monotone decreasing sequence. The condition for a monotone decreasing sequence of values for e is related to the formalism used in the calculus of variations. Specifically, the integral

$$e = \int_0^T \left\{ f_0(\mathbf{x}, \mathbf{m}, t) + \sum_{n=1}^{N} p_n[f_n(\mathbf{x}, \mathbf{m}, t) - \dot{x}_n] \right\} dt \tag{4}$$

is minimized in the calculus of variations where p_n is treated as a Lagrange multiplier and is adjusted such that (2) is satisfied. This procedure yields the necessary conditions for a minimum given in (A9), (A10), and (A11) of the Appendix. The iterative procedure, however, is based on neglecting the condition given in (A11) by arbitrarily assuming a vector $\mathbf{m}^{(i)}$. Then the corresponding solution $\mathbf{x}^{(i)}$ is obtained from (A9), and also the corresponding solution $\mathbf{p}^{(i)}$ is obtained from (A10). The condition $e^{(i+1)} < e^{(i)}$ then is obtained in terms of

$$S_m^{(i)} = \frac{\partial f_0^{(i)}}{\partial m_m^{(i)}} + \sum_{n=1}^{N} \frac{\partial f_n^{(i)}}{\partial m_m^{(i)}} p_n^{(i)} \tag{5}$$

which results from neglecting (A11). The notational simplification $f_n^{(i)} = f_n(\mathbf{x}^{(i)}, \mathbf{m}^{(i)}, t)$, etc., is adopted throughout. Also the notations

$$M_n^{(i)} = m_n^{(i+1)} - m_n^{(i)}, \qquad X_n^{(i)} = x_n^{(i+1)} - x_n^{(i)}, \tag{6}$$

are used.

* The question of whether the limit point of this sequence is in general also the minimum value of e apparently is unresolved.

The condition $e^{(i+1)} < e^{(i)}$ is obtained directly from (4) with a Taylor series about the ith trajectory under the assumption $M_n^{(i)}$ is maintained sufficiently small. In this expansion, third and higher degree terms in the elements of $\mathbf{M}^{(i)}$ and $\mathbf{X}^{(i)}$ are neglected. Therefore, the series required are

$$(7)\qquad \begin{aligned} f_n^{(i+1)} \cong f_n^{(i)} + \sum_{m=1}^{M} \frac{\partial f_n^{(i)}}{\partial m_m^{(i)}} M_m^{(i)} + \cdots \\ + \sum_{m=1}^{N} \sum_{k=1}^{N} \frac{1}{2} \frac{\partial^2 f_n^{(i)}}{\partial x_m^{(i)} \partial x_k^{(i)}} X_m^{(i)} X_k^{(i)} \end{aligned}$$

and

$$(8)\qquad p_k^{(i+1)} \cong p_k^{(i)} + 2\sum_{l=1}^{N} p_{kl}^{(i)} X_l^{(i)} ,$$

where p_{kl} is defined in (A12). The expansion of (4) is completed by rearranging terms, by substituting the definitions given in (5), (A14), (A15), and (A16), and by substituting the equations given in (A10) and (A17). Also, integration by parts is performed, and the boundary conditions $X_n^{(i)}(0) = 0$, according to (3), and $p_k^{(i)}(T) = p_{kl}^{(i)}(T) = 0$, according to (A2), are imposed. These straightforward steps yield

$$(9)\qquad e^{(i+1)} \cong e^{(i)} + v^{(i)},$$

where the integral $v^{(i)}$ is given by

$$(10)\qquad \begin{aligned} v^{(i)} = \int_0^T \Bigg\{ \frac{1}{2} \sum_{n=1}^{M} \sum_{m=1}^{M} [T_{nm}^{(i)} M_n^{(i)} M_m^{(i)}] \\ + \sum_{n=1}^{M} \left[S_n^{(i)} + \sum_{m=1}^{N} R_{nm}^{(i)} X_m^{(i)} \right] M_n^{(i)} \\ + \frac{1}{2} \sum_{n=1}^{M} \sum_{m=1}^{N} \sum_{k=1}^{N} [R_{nm}^{(i)} K_{nk}^{(i)} X_m^{(i)} X_k^{(i)}] \Bigg\}\, dt. \end{aligned}$$

This integral involves both the first and second variations* of $e^{(i+1)}$ about $e^{(i)}$. The basic method due to Kelley and Bryson involves only first variations and is obtained by merely neglecting the second degree terms in the elements of $\mathbf{M}^{(i)}$ and $\mathbf{X}^{(i)}$ which appear in (10). The condition for a monotone decreasing sequence of values for $e^{(i)}$ is established in either case by choosing $\mathbf{M}^{(i)}$ such that $v^{(i)} < 0$ according to (10) but restricting the magnitudes of the elements of $\mathbf{M}^{(i)}$ to suitably small values such that $e^{(i+1)} < e^{(i)}$ according to (1). As opposed to the case of first variations only, however, the iteration algorithm is not established by selecting $\mathbf{M}^{(i)}$ such that the integrand of (10) is negative definite for all t.

Iteration algorithm. The iteration algorithm introduced here is based on

* This term is used here in the classical sense of the calculus of variations.

a second variational problem. Specifically, the perturbation vector $\mathbf{M}^{(i)}$ is selected so that (10) is minimized* subject to the linearized incremental state equation

$$(11) \qquad \dot{X}_n^{(i)} = \sum_{m=1}^{N} \frac{\partial f_n^{(i)}}{\partial x_m^{(i)}} X_m^{(i)} + \sum_{m=1}^{M} \frac{\partial f_n^{(i)}}{\partial m_m^{(i)}} M_m^{(i)}.$$

The solution to this variational problem is well-known in linear optimum controls [5], but unusual simplifications occur here. If $V^{(i)}$ is defined as the minimum value of $v^{(i)}$, then the methods used for linear optimum controls yield

$$(12) \qquad V^{(i)} = -\frac{1}{2}\int_0^T \left\{ \sum_{n=1}^{M} \sum_{m=1}^{M} T_{nm}^{(i)} G_n^{(i)} G_m^{(i)} \right\} dt.$$

The variable $G_m^{(i)}$ is defined by

$$(13) \qquad \sum_{m=1}^{M} T_{nm}^{(i)} G_m^{(i)} = -S_n^{(i)} - \sum_{k=1}^{N} \frac{\partial f_k^{(i)}}{\partial m_n^{(i)}} g_k^{(i)}$$

and

$$(14) \qquad -\dot{g}_k^{(i)} = \sum_{n=1}^{M} R_{nk}^{(i)} G_n^{(i)} + \sum_{n=1}^{N} \frac{\partial f_n^{(i)}}{\partial x_k^{(i)}} g_n^{(i)},$$

where $g_k^{(i)}(T) = 0$. The variable $g_k^{(i)}$ is analogous to $p_k^{(i)}$ which arises in the original minimization problem. However the variable $g_{kl}^{(i)}$, which would be analogous to $p_{kl}^{(i)}$, is zero. In addition, the elements of the optimal vector $\mathbf{M}^{(i)}$ are given by

$$(15) \qquad M_n^{(i)} = G_n^{(i)} - \sum_{k=1}^{N} K_{nk}^{(i)} X_k^{(i)}.$$

As discussed previously, the elements of $\mathbf{M}^{(i)}$ must be restricted in magnitude in order to validate the expansions which lead to (10) and (11). Here, step-size is restricted by introducing the parameter ϵ such that

$$(16) \qquad M_n^{(i)} = \epsilon G_n^{(i)} - \sum_{k=1}^{N} K_{nk}^{(i)} X_k^{(i)}.$$

When (11) and (16) are substituted into (10) and the appropriate manipulations are performed, the integral $v^{(i)}$ becomes

$$(17) \qquad v^{(i)} = (\epsilon^2 - 2\epsilon)[-V^{(i)}].$$

However $V^{(i)}$ is negative which is insured by the property that the matrix $[T_{nm}^{(i)}]$ is positive definite when $\mathbf{x}^{(i)}$ is in a suitably small neighborhood of

* A second two-point boundary value problem would be encountered here if the control and state variables were subject to point constraints.

the optimal trajectory for the class of variational problems discussed here. As a result, $v^{(i)}$ is a strictly convex function of the parameter ϵ and also is negative on the interval $0 < \epsilon < 2$. Therefore the iteration algorithm based on both first and second variations is taken to be (16) where stepsize is adjusted on the interval $0 < \epsilon \leqq 1$.

As expected, the algorithm given in (16) yields one-step convergence for $\epsilon = 1$ when $f_0(\mathbf{x}, \mathbf{m}, t)$ and $\mathbf{f}(\mathbf{x}, \mathbf{m}, t)$ are quadratic and linear, respectively, with respect to the elements of $\mathbf{x}$ and $\mathbf{m}$. This property results because no approximations are introduced by the expansion of $e^{(i+1)}$. Also, this property gives rise to extremely rapid convergence when $\mathbf{x}^{(i)}$ is in a suitably small neighborhood of the optimal trajectory.

Numerical considerations. In actual numerical solutions, the incidence of truncation errors, which are primarily due to numerical integration, makes the exact computation of the optimal trajectory and hence one-step convergence impossible. Therefore a suitable condition for terminating the computations is needed. A condition of this type is obtained with the introduction of the variables $U_m^{(i)}$ and $G_{mk}^{(i)}$ such that

$$(18) \qquad \sum_{m=1}^{M} T_{nm}^{(i)} U_m^{(i)} = -S_n^{(i)}; \qquad \sum_{m=1}^{M} T_{nm}^{(i)} G_{mk}^{(i)} = -\frac{\partial f_k^{(i)}}{\partial m_n^{(i)}}.$$

The variable $G_n^{(i)}$ then becomes

$$(19) \qquad G_n^{(i)} = U_n^{(i)} + \sum_{k=1}^{N} G_{nk}^{(i)} g_k^{(i)}.$$

The vector $\mathbf{U}^{(i)}$ depends on $\mathbf{S}^{(i)}$, and the condition $\mathbf{U}^{(i)} = \mathbf{0}$ occurs on the optimal trajectory where $\mathbf{S}^{(i)} = \mathbf{0}$ everywhere on the interval $0 \leqq t \leqq T$. The condition for terminating the computations is written in terms of a norm and is taken to be

$$(20) \qquad \| \mathbf{U}^{(i)} \| < \delta, \qquad 0 \leqq t \leqq T.$$

The number δ, although somewhat difficult to specify, is chosen in correspondence with the truncation errors.

The form of (16) suggested for actual numerical solutions is

$$(21) \qquad m_n^{(i+1)} = \epsilon G_n^{(i)} + K_n^{(i)} - \sum_{k=1}^{N} K_{nk}^{(i)} x_k^{(i+1)},$$

where

$$(22) \qquad K_n^{(i)} = m_n^{(i)} + \sum_{k=1}^{N} K_{nk}^{(i)} x_k^{(i)}$$

and $G_n^{(i)}$ is given by (19). This form of the algorithm is a linear control equation which can be used for control purposes [6] with $\epsilon = 0$ when $\mathbf{x}^{(i+1)}$

and $\mathbf{x}^{(i)}$ are in a suitably small neighborhood of the optimal trajectory. Presumably, the iterations would continue until the test given in (20), which can be performed during the backward-time integrations, is satisfied.

Another numerical aspect of the iteration algorithm based on first and second variations concerns the stability of the differential equations used in the computations. If $\mathbf{x}^{(i)}$ is in a suitably small neighborhood of the optimal trajectory, then the linearized equation given in (11) is a valid representation of (2). When (16) is substituted into (11), the incremental state equations associated with the forward-time computations from $t = 0$ to $t = T$ are found to be

$$(23)\quad \dot{X}_n^{(i)} = \sum_{k=1}^{N}\left[\frac{\partial f_n^{(i)}}{\partial x_k^{(i)}} - \sum_{m=1}^{M}\frac{\partial f_n^{(i)}}{\partial m_m^{(i)}}K_{mk}^{(i)}\right]X_k^{(i)} + \epsilon\sum_{m=1}^{M}\frac{\partial f_n^{(i)}}{\partial m_m^{(i)}}G_m^{(i)}.$$

In other words, the forward-time computations are performed with differential equations, namely (2), that possess the stability properties of linear optimum control systems [6], when $\mathbf{x}^{(i)}$ is in a suitably small neighborhood of the optimal trajectory, as opposed to the stability properties of (11). This property is particularly important from a numerical point of view when (11) is unstable. Also, suitable manipulations show that the backward-time computations from $t = T$ to $t = 0$ are performed with the adjoint equations of (23) as opposed to the adjoint equations of (11). Specifically, the substitution of (19) into (14) and the use of the definitions introduced in (18) and (A18) yield

$$(24)\quad -\dot{g}_k^{(i)} = \sum_{n=1}^{M} R_{nk}^{(i)}U_n^{(i)} + \sum_{n=1}^{N}\left[\frac{\partial f_n^{(i)}}{\partial x_k^{(i)}} - \sum_{m=1}^{M}\frac{\partial f_n^{(i)}}{\partial m_m^{(i)}}K_{mk}^{(i)}\right]g_n^{(i)}.$$

Similar manipulations of (A10) in conjunction with (5) yield

$$(25)\quad \begin{aligned} -\dot{p}_k^{(i)} &= -\sum_{n=1}^{M} R_{nk}^{(i)}U_n^{(i)} + \left[\frac{\partial f_0^{(i)}}{\partial x_k^{(i)}} - \sum_{m=1}^{M}\frac{\partial f_0^{(i)}}{\partial m_m^{(i)}}K_{mk}^{(i)}\right] \\ &\quad + \sum_{n=1}^{N}\left[\frac{\partial f_n^{(i)}}{\partial x_k^{(i)}} - \sum_{m=1}^{M}\frac{\partial f_n^{(i)}}{\partial m_m^{(i)}}K_{mk}^{(i)}\right]p_n^{(i)}. \end{aligned}$$

These two backward-time equations also possess the stability properties of linear optimum control systems when $\mathbf{x}^{(i)}$ is in a suitable small neighborhood of the optimal trajectory. Therefore the backward-time computations associated with this algorithm are performed with (24), (25), and (A17) which is of the stable Ricatti type when $\mathbf{x}^{(i)}$ is in a suitably small neighborhood of the optimal trajectory. When (24) and (25) are used in numerical integration, a procedure of setting the elements of $\mathbf{U}^{(i)}$ equal to zero wherever $\|\mathbf{U}^{(i)}\| < \delta$ is satisfied on the interval $0 \leqq t \leqq T$ usually is adopted. This procedure appears to inhibit the propagation of truncation errors in the neighborhood of the optimal trajectory where (5) may involve small differences of large numbers.

Conclusions. This algorithm possesses the property of an exceedingly rapid convergence rate when $\mathbf{x}^{(i)}$ is in a suitably small neighborhood of the optimal trajectory. The basic algorithm due to Kelley and Bryson many times is characterized as a step-by-step gradient ascent on a convex surface. However the algorithm based on only first variations does not account for the surface deformations caused by the step taken. The algorithm based on first and second variations includes a first order approximation to the deformations in this surface, thereby giving rise to rapid convergence under suitable conditions. The added elapsed computer time per iteration required to solve the added $N + N(N + 1)/2$ backward-time differential equations when second variations are used appears to be more than compensated for under these conditions. However this conclusion is based on limited experience with the algorithm and most likely does not apply for large N.

Another attribute of the algorithm developed here is that the computations always are performed with differential equations which tend to improve the stability of (2) and (A10) when $\mathbf{x}^{(i)}$ is within a suitably small neighborhood of the optimal trajectory. This property eases the difficulty of obtaining numerically accurate solutions which are required in order to converge to a nearly optimal trajectory.

On the other hand, the algorithm based on first and second variations suffers from the conceptual disadvantage that penalty functions must be used to approximate point constraints on the control and state variables. Also, penalty functions must be used for approximating fixed-point terminal boundary conditions. In this situation, (1) would have the alternate form

$$e = \int_0^T f_0(\mathbf{x}, \mathbf{m}, t)\, dt + F_0(\mathbf{x}, T),$$

where F_0 is a strictly convex, continuous function of the elements of $\mathbf{x}(T)$. The only modifications required in the previous results are the boundary conditions

$$(26) \qquad p_k(T) = \frac{\partial F_0(\mathbf{x}, T)}{\partial x_k}, \qquad p_{kl}(T) = \frac{1}{2}\frac{\partial^2 F_0(\mathbf{x}, T)}{\partial x_k\, \partial x_l}.$$

This algorithm also has the practical disadvantage that an added set of MN time functions, namely $K_{nk}^{(i)}$, must be stored in tabulated form, thereby increasing the amount of temporarily stored data.

Another disadvantage occurs in the selection of an initial trajectory when (2) is both nonlinear and unstable. Specifically, numerical examples support the conjecture that the backward-time differential equations are unstable when $\mathbf{x}^{(i)}$ gives rise to an unstable form of (23). The numerical solution of (A17) is particularly difficult when $[p_{kl}^{(i)}]$ is not positive definite, a condition which may occur if $\mathbf{x}^{(i)}$ is not in a suitably small neighborhood of the optimal trajectory. Similarly, the matrix $[T_{nm}^{(i)}]$ may not be positive definite for certain initial trajectories, thereby invalidating the required negative definite property of (12). As a result of these difficulties, the algorithm based on second variations is recommended only when $\mathbf{x}^{(i)}$ is in a suitably small neighborhood of the optimal trajectory.

Acknowledgments. The author is deeply indebted to Mr. R. J. Ringlee of the General Electric Company for many valuable discussions and suggestions which contributed to the development of this algorithm. Dr. I. Lee of the General Electric Research Laboratory made helpful suggestions concerning notation and the exposition of this work.

Appendix. This appendix is included in order to define notation and to state certain relationships arising in variational mathematics.

The minimum-error function is defined in accordance with dynamic programming [7] as

$$E(\mathbf{x}, t) = \min_{\mathbf{m}} \left[\int_t^T f_0(\mathbf{x}, \mathbf{m}, \sigma)\, d\sigma \right], \tag{A1}$$

subject to the boundary condition

$$E(\mathbf{x}, T) = 0. \tag{A2}$$

The Hamiltonian function is defined as

$$H(\mathbf{x}, t) = \min_{\mathbf{m}} \left[f_0(\mathbf{x}, \mathbf{m}, t) + \sum_{n=1}^{N} p_n f_n(\mathbf{x}, \mathbf{m}, t) \right], \tag{A3}$$

where the variable p_n is defined as

$$p_n = \frac{\partial E(\mathbf{x}, t)}{\partial x_n}. \tag{A4}$$

In accordance with these definitions, the minimum-error function satisfies the Hamilton-Jacobi equation [8] which is written as

$$\frac{\partial E(\mathbf{x}, t)}{\partial t} + H(\mathbf{x}, t) = 0. \tag{A5}$$

The characteristic equations [9] which are associated with (A5) and which are well known from the calculus of variations are derived readily. Specifically, the total time derivative of p_k is expanded in terms of partial derivatives as

$$\dot{p}_k = \frac{\partial p_k}{\partial t} + \sum_{n=1}^{N} \frac{\partial p_k}{\partial x_n} \dot{x}_n. \tag{A6}$$

In addition, the partial differentiation of (A5) with respect to x_k yields

$$\frac{\partial p_k}{\partial t} + \frac{\partial H(\mathbf{x}, t)}{\partial x_k} = 0, \tag{A7}$$

so that (A6) becomes

$$-\dot{p}_k = \frac{\partial H(\mathbf{x}, t)}{\partial x_k} - \sum_{n=1}^{N} \frac{\partial p_k}{\partial x_n} \dot{x}_n . \tag{A8}$$

If the partial differentiation of H indicated in (A8) is expanded and the state equation

$$\dot{x}_k = f_k(\mathbf{x}, \mathbf{m}, t) \tag{A9}$$

is used, then (A8) becomes

$$-\dot{p}_k = \frac{\partial f_0(\mathbf{x}, \mathbf{m}, t)}{\partial x_k} + \sum_{n=1}^{N} p_n \frac{\partial f_n(\mathbf{x}, \mathbf{m}, t)}{\partial x_k} . \tag{A10}$$

The relationship

$$\frac{\partial f_0(\mathbf{x}, \mathbf{m}, t)}{\partial m_m} + \sum_{n=1}^{N} p_n \frac{\partial f_n(\mathbf{x}, \mathbf{m}, t)}{\partial m_m} = 0 \tag{A11}$$

also is used in obtaining (A10) and is found from the minimization indicated in (A3). Equations (A9) and (A10) are the characteristic equations.

Additional relationships are required for the computational method based on second variations. Specifically, the variable

$$p_{kl} = \frac{1}{2} \frac{\partial p_k}{\partial x_l} \tag{A12}$$

is introduced [6, 10] subject to the requirement that $p_{kl} = p_{lk}$. When steps similar to those leading to (A8) are performed, the total time derivative of p_{kl} becomes

$$-\dot{p}_{kl} = \frac{1}{2} \frac{\partial^2 H(\mathbf{x}, t)}{\partial x_k \, \partial x_l} - \sum_{n=1}^{N} \frac{\partial p_{kl}}{\partial x_n} \dot{x}_n . \tag{A13}$$

The explicit form of the right-hand side of (A13) is found when the indicated partial differentiation of H is expanded. For the sake of simplicity, the definitions

$$K_{nk} = -\frac{\partial m_n}{\partial x_k}, \tag{A14}$$

$$R_{nk} = \frac{\partial^2 f_0(\mathbf{x}, \mathbf{m}, t)}{\partial m_n \, \partial x_k} + \sum_{m=1}^{N} \left[p_m \frac{\partial^2 f_m(\mathbf{x}, \mathbf{m}, t)}{\partial m_n \, \partial x_k} + 2 p_{mk} \frac{\partial f_m(\mathbf{x}, \mathbf{m}, t)}{\partial m_n} \right], \tag{A15}$$

and

$$T_{nm} = \frac{\partial^2 f_0(\mathbf{x}, \mathbf{m}, t)}{\partial m_n \, \partial m_m} + \sum_{k=1}^{N} p_k \frac{\partial^2 f_k(\mathbf{x}, \mathbf{m}, t)}{\partial m_n \, \partial m_m} \tag{A16}$$

are introduced for frequently occurring sets of terms. With these definitions, (A13) reduces to

$$\text{(A17)}\quad \begin{aligned} -\dot{p}_{kl} &= \frac{1}{2}\frac{\partial^2 f_0(\mathbf{x}, \mathbf{m}, t)}{\partial x_k \partial x_l} + \frac{1}{2}\sum_{n=1}^{N} p_n \frac{\partial^2 f_n(\mathbf{x}, \mathbf{m}, t)}{\partial x_k \partial x_l} \\ &+ \sum_{n=1}^{N}\left[p_{nl}\frac{\partial f_n(\mathbf{x}, \mathbf{m}, t)}{\partial x_k} + p_{nk}\frac{\partial f_n(\mathbf{x}, \mathbf{m}, t)}{\partial x_l}\right] - \frac{1}{2}\sum_{n=1}^{M} R_{nk}K_{nl}. \end{aligned}$$

The variable K_{nl} is found from the partial differentiation of (A11) with respect to x_l and is given by

$$\text{(A18)}\quad \sum_{m=1}^{M} T_{nm}K_{ml} = R_{nl}.$$

The matrices $[p_{kl}]$ and $[T_{nm}]$ are positive definite for the class of minimization problems treated here. The condition on $[p_{kl}]$, which is the matrix of second derivatives of $E(\mathbf{x}, t)$, results because the minimum-error function is a continuous, strictly convex function. The condition on $[T_{nm}]$ is the Legendre condition which is sufficient for a unique bounded minimum found from (A3).

REFERENCES

[1] H. J. KELLEY, *Gradient theory of optimal flight paths*, Journal of the American Rocket Society, 30 (1960), pp. 947–953.

[2] A. E. BRYSON, W. F. DENHEM, F. J. CARROLL, AND K. MIKAMI, *Lift or drag programs that minimize re-entry heating*, Journal of the Aerospace Sciences, 29 (1962), pp. 420–430.

[3] J. L. SPEYER, *Optimization and control using perturbation theory to find neighboring optimum paths*, presented at the SIAM Symposium on Multivariable System Theory, November, 1962.

[4] H. J. KELLEY, *Method of gradients*, Optimization Techniques (ed., G. Leitmann), Academic Press, New York, 1962, chap. 6.

[5] C. W. MERRIAM III, *A class of optimum control systems*, Journal of the Franklin Institute, 267 (1959), pp. 267–281.

[6] ———, *Synthesis of nonlinear optimum control systems for small regions of state-space*, Optimization Theory and the Design of Feedback Control Systems, McGraw-Hill, New York, 1964, chap. 9.

[7] R. BELLMAN, *A new formalism in the calculus of variations*, Dynamic Programming, Princeton University Press, Princeton, 1957, chap. 9.

[8] L. I. ROZONOER, *The variational method in quality analysis of automatic control systems*, Proceedings of the First International Congress of the IFAC, Butterworths, London, 1960.

[9] F. JOHN, *The general first order equation for a function of n independent variables*, Partial Differential Equations, New York University Institute of Mathematical Sciences, New York, 1952, chap. 3.

[10] W. KIPINIAK, *Control law computations*, Dynamic Optimization and Control, M.I.T. Press and John Wiley, New York, 1961, chap. 3.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

ASYMPTOTIC CONTROL THEORY*

RICHARD BELLMAN† AND RICHARD BUCY‡

1. Introduction. In recent years the mathematical theory of control has received an increasing amount of attention. New theories have been developed and older theories have been refined and extended [1, 2, 3, 4, 5, 6].

In this paper, we wish to initiate discussion of a problem in the calculus of variations which has not had the attention due it in the classical literature. The problem is concerned with the asymptotic behavior of the solution of a variational problem as the time interval becomes infinite. From the standpoint of control theory, and more generally from the standpoint of dynamic programming, this is a very natural type of behavior to study. In many significant cases, the "steady-state" policy is simpler conceptually, analytically and computationally.

We shall consider the minimization of the functional

$$J(u) = \frac{1}{2}\int_0^T (u^2 + L(x))\,dt, \tag{1.1}$$

over all functions u where

$$\dot{x} = f(x) + u, \qquad x(0) = c. \tag{1.2}$$

Let $V(c, T) = \min_u J(u)$. For finite and sufficiently small T the classical calculus of variations, or dynamic programming applies, under certain reasonable assumptions on L and f. We shall be interested, however, in the following questions.

(1) When does the problem for infinite T make sense?

(2) When it does, are the optimal motions and policies for infinite T the limits of the corresponding optimal motions and policies for finite T?

(3) What is the effect of using steady-state optimal policy for the finite problem?

This is an example of what we mean by *asymptotic control theory.*

For example, if $f = 0$ and $L = x^2 + \frac{1}{2}x^4$ the problem is that of minimizing the functional

$$J(u) = \int_0^T [\dot{x}^2 + x^2 + \tfrac{1}{2}\,x^4]\,dt \tag{1.3}$$

* Received by the editors September 23, 1963.

This research is sponsored by the United States Air Force under Project RAND—Contract No. AF 49(638)-700.

† RAND Corporation, 1700 Main Street, Santa Monica, California.

‡ Consultant, RAND Corporation. Currently at RIAS, 7212 Bellona Avenue, Baltimore, Maryland.

over all C^1 curves for which $x(0) = c$. The Euler equation is

$$\ddot{x} - x - x^3 = 0, \tag{1.4}$$

subject to the two-point boundary conditions

$$x(0) = c, \qquad \dot{x}(T) = 0. \tag{1.5}$$

Establishing the existence and uniqueness of solutions of (1.4) and determining the asymptotic behavior as $T \to \infty$ is analogous to the classical problem of Poincaré-Lyapunov [5], but materially more difficult because of the two-point boundary-value condition.

We shall first, using quite general arguments, show that $V(c, T)$ is monotone increasing as a function of T, and uniformly bounded under mild restrictions concerning $L(x)$. Taking advantage of the fact that the Euler equation posseses a first integral, we can analyze the behavior of the solution in detail as $T \to \infty$.

This analysis shows that the formal asymptotic series obtained from the partial differential equation

$$V_T = \min_u \left[\tfrac{1}{2}(u^2 + L(x)) + V_x(ax + u)\right]. \tag{1.6}$$

an equation derived from dynamic programming considerations which yields the Hamilton-Jacobi equation relevant to the variational problem when $f(x) = ax$ ([1] and [12]), is an actual asymptotic series for $V(c, T)$. This corresponds to the result easily derived in the case where the integrand in (1.1) is merely quadratic in x and u.

In the concluding section, we shall mention some open and apparently quite difficult questions in connection with asymptotic behavior and give some references to analogous results obtained for dynamic programming processes by Kalman and Bucy [6], Beckwith [7], Iglehart [8], Freimer [9], and Bellman [10].

2. Monotonicity and boundedness. Let us introduce the function

$$V(c, T) = \min_u J(u), \tag{2.1}$$

(with the assumption that $f(x) = ax$). Let $x(t, T)$, $u(t, T)$ represent the functions that furnish the minimum of $J(u)$ under the assumption that $L(x)$ is a nonnegative entire function of x. In most processes of interest $L(x)$ is a polynomial in x.

Since

$$\begin{aligned} V(c, T + \Delta) &= \int_0^T + \int_T^{T+\Delta} \\ &\geqq \min_u \int_0^T + \int_T^{T+\Delta} \\ &\geqq V(c, T) + \int_T^{T+\Delta}, \end{aligned} \tag{2.2}$$

we see that $V(c, T)$ is monotone increasing in T.

To show uniform boundedness in T, for fixed c, let us choose an appropriate control policy, say

$$\begin{aligned} u &= 0, && \text{when } a < 0, \\ u &= -2ax, && \text{when } a > 0, \\ u &= -x, && \text{when } a = 0. \end{aligned} \tag{2.3}$$

In each case, we see that $u = ce^{-bt}$ with b positive. Hence,

$$J(u) = \int_0^T [O(e^{-2bt}) + L(ce^{-bt})]\,dt. \tag{2.4}$$

Under the assumption that $L(x) = O(x)$ as $x \to 0$, the integral is uniformly bounded as $T \to \infty$.

Having established boundedness and monotonicity as $T \to \infty$, we can assert convergence,

$$V(c, T) \to V(c) \tag{2.5}$$

as $T \to \infty$.

It is not settled, however, whether or not the states $x(t, T)$ and the policies $u(t, T)$ converge as $T \to \infty$. The foregoing argument extends to quite general situations, but leaves unanswered the interesting and important questions concerning the convergence of policies.

3. Detailed analysis. We will be interested in an explicit solution to the partial differential equation

$$V_T = \tfrac{1}{2}L(c) + acV_c - \tfrac{1}{2}V_c^2, \tag{3.1}$$

subject to the boundary conditions

$$\begin{aligned} V(c, T)|_{T=0} &= 0, && a < 0, \\ V(c, T)|_{T=0} &= ac^2, && a > 0. \end{aligned} \tag{3.2}$$

As is well known [12], existence of a sufficiently smooth solution to (3.1) is a sufficient condition for the variational problem (1.1) to have a solution. The equation of (3.1) is (1.6) with the minimization carried out.

It will be assumed that L satisfies the following conditions.

(3.3)
(1) L is even, and positive.
(2) L and L_x are continuous and increasing for positive x.
(3) $L(x) = O(|x|)$ as $|x| \to 0$.
(4) L is analytic.

Now the Cauchy-Kowalewski theorem implies (3.1) has a unique local analytic solution [11].

With the aim of solving (3.1) we introduce the function $y(c, T)$, which corresponds physically to the final state of the controlled system along an optimal trajectory initiating at $(c, 0)$ and ending at (y, T). The following lemma shows that y is well defined for $c > 0$. The case $c < 0$ is similar. When $L(c) = c^2 + c^4$, y will be defined by an elliptic integral of the first kind.

LEMMA 1. *Suppose* $c > 0$, *and assume conditions* (3.3) *are fulfilled. Then for every* K, $\infty > K > 0$, *there exists a unique* $0 < y \leqq c$ *such that*

$$I(y) = \int_y^c \frac{dx}{\sqrt{a^2 x^2 + L(x) - L(y)}} = K. \tag{3.4}$$

Proof. Since it is clearly continuous, elementary bounding of $I(y)$ shows it takes on all finite positive values as y ranges over $(0, c]$. To show uniqueness assume for some finite positive K that there exist y_1 and y_2, $y_1 > y_2 > 0$, where both satisfy (3.4). Then

$$\int_{y_1}^c \frac{dx}{\sqrt{a^2x^2 + L(x) - L(y_1)}} = \int_{y_2}^c \frac{dx}{\sqrt{a^2x^2 + L(x) - L(y_2)}}. \tag{3.5}$$

But (3.5) implies

$$\int_{y_2}^{c-\Delta} \frac{dx}{\sqrt{a^2(x+\Delta)^2 + L(x+\Delta) - L(y_1)}} > \int_{y_2}^{c-\Delta} \frac{dx}{\sqrt{a^2x^2 + L(x) - L(y_2)}}, \tag{3.6}$$

where $\Delta = y_1 - y_2 > 0$. It suffices to show (3.7) to contradict (3.6):

$$2a^2x\Delta + a^2\Delta^2 > L(x) - L(x + \Delta) + L(y_1) - L(y_2). \tag{3.7}$$

Consider the right-hand side of (3.7) for $x \geqq y_1$. The mean value theorem gives

$$L(x + \Delta) - L(x) = L_x(\varphi)\Delta \geqq L_x(y_1)\Delta, \quad \varphi \in (x, x + \Delta),$$

$$L(y_1) - L(y_2) = L_x(\theta)\Delta \leqq L_x(y_1)\Delta, \quad \theta \in (y_2, y_1),$$

and for these x, (3.7) is satisfied. Then for $y_2 < x < y_1$,

$$L(x + \Delta) - L(y_1) = L_x(\delta)(x - y_2) \geqq L_x(y_1)(x - y_2),$$

$$L(x) - L(y_2) = L_x(\gamma)(x - y_2) \leqq L_x(x)(x - y_2) \leqq L_x(y_1)(x - y_2),$$

and (3.7) is satisfied for all $x \in (y_2, c - \Delta)$.

A closer examination of the previous lemma shows $I(y)$ decreases strictly as y increases through $(0, c]$ and hence $(\partial/\partial y)I(y) < 0$. Defining $y(c, T)$ as that value of y which satisfies

$$\int_y^c \frac{dx}{\sqrt{a^2x^2 + L(x) - L(y)}} = T, \tag{3.8}$$

it is easy to see by the implicit function theorem that y is C^1 in c and T on $(0, c] \times [0, \infty)$. Further, y is analytic! The following lemma characterizes the behavior of y as T tends to infinity.

LEMMA 2. *Under the assumptions* (3.3) *and* $a \neq 0$, *y defined by* (3.8) *tends exponentially to zero as* $T \to \infty$.

Proof.

$$0 < T = \int_y^c \frac{dx}{\sqrt{a^2x^2 + L(x) - L(y)}} \leqq \int_y^c \frac{dx}{|a|\,x} = \frac{1}{|a|} \ln \frac{c}{y},$$

or $0 \leqq y \leqq ce^{-|a|T}$.

Now we will characterize the solution of (3.1) subject to (3.2).

THEOREM[*] 1. *Equation* (3.1) *has the following analytic solutions in the regions* $c > 0$ *and* $c < 0$ *under assumptions* (3.3). *For* $a < 0$,

$$(3.9) \qquad V(c, T) = \begin{cases} \displaystyle\int_y^c a\xi + \sqrt{a^2\,\xi^2 + L(\xi) - L(y)}\, d\xi + T\frac{L(y)}{2}, c > 0, \\ 0, c = 0, \\ \displaystyle\int_y^c a\xi - \sqrt{a^2\,\xi^2 + L(\xi) - L(y)}\, d\xi + T\frac{L(y)}{2}, c < 0; \end{cases}$$

while for $a > 0$,

$$(3.10) \qquad V(c, T) = \begin{cases} \displaystyle\int_y^c a\xi + \sqrt{a^2\xi^2 + L(\xi) - L(y)}\, d\xi \\ \qquad\qquad\qquad\qquad + T\frac{L(y)}{2} + ay^2, c > 0, \\ 0, c = 0, \\ \displaystyle\int_y^c a\xi - \sqrt{a^2\xi^2 + L(\xi) - L(y)}\, d\xi \\ \qquad\qquad\qquad\qquad + T\frac{L(y)}{2} + ay^2, c < 0; \end{cases}$$

where y satisfies

$$\int_y^c \frac{d\xi}{\sqrt{a^2\xi^2 + L(\xi) - L(y)}} = T \textit{ for } c > 0,$$

$$-\int_y^c \frac{d\xi}{\sqrt{a^2\xi^2 + L(\xi) - L(y)}} = T \textit{ for } c < 0,$$

[*] Replacing $a\xi$ and $(a\xi)^2$ by $f(\xi)$ and $f(\xi)^2$ with f continuous and ay^2 by $2\int_0^y f(\xi)\, d\xi$, provides a solution to the equation $V_T = L + V_x f(x) - \frac{1}{2}V_x^2$ which is local unless y is defined for all T. Further Haar's uniqueness theorem [13] is applicable and implies (3.9) for $c > 0$ is the unique C^1 solution of (3.1).

and y has the same sign as c. Further,

$$(3.11)\quad \begin{aligned} V(c, \infty) &= \lim_{T\to\infty} V(c, T) = \int_0^c a\xi + \sqrt{a^2\xi^2 + L(\xi)}\, d\xi, \quad c > 0, \\ V_c(c, \infty) &= \lim_{T\to\infty} V_c(c, T) = ac + \sqrt{a^2c^2 + L(c)}\,, \quad c > 0, \end{aligned}$$

and the corresponding formulae for $c < 0$ and $c = 0$. Finally, the optimum control law is

$$(3.12)\quad u^0(t) = -ax(t) - \operatorname{sgn} x(t) \sqrt{a^2x(t)^2 + L(x(t)) - L(y)}.$$

Proof. Equations (3.9) and (3.10) follow from Lemma 1 and direct substitution. Equation (3.11) follows, since for fixed c, $V(c, T)$ and $V_c(c, T)$ are monotone in T and uniformly bounded, while the explicit form follows from Lemma 2, (3.3)–(3) and the dominated convergence theorem. Equation (3.12) is just the principle of optimality.

COROLLARY 1. *Under the previous assumptions, the value of the T-infinite case $V(c, \infty)$ satisfies $\frac{1}{2}L + V_c ac - \frac{1}{2}V_c^2 = 0$, just (3.1) with $V_T = 0$.*

4. Asymptotic behavior.

As mentioned above, the principle of optimality yields the partial differential equation

$$(4.1)\qquad V_T = \tfrac{1}{2}L(c) + acV_c - \tfrac{1}{2}V_c^2.$$

It does not seem possible to obtain the asymptotic behavior of V, even formally, by means of a series of the form

$$(4.2)\qquad V = V_0(c) + V_1(c, T) + \cdots,$$

without some additional information concerning the analytic structure of V_1, e.g.,

$$(4.3)\qquad V_1(c, T) = V_1(c)u_1(T).$$

Here $V_0(c) = \lim_{T\to\infty} V(c, T)$.

We can, however, obtain an interesting bound for $V(c, \infty) - V(c, T)$ in the following fashion. Consider the expression

$$(4.4)\qquad V(c, \infty) = \min_u \int_0^\infty [x^2 + u^2 + L(x)]\, dt.$$

Let $u(t, T)$, $x(t, T)$ denote the minimizing set of functions for the interval $[0, T]$. Then, it is clear that

$$(4.5)\quad V(c, \infty) \leqq \int_0^T [u(t, T)^2 + x(t, T)^2 + L(x(t, T))]\, dt + \int_T^\infty [\cdots]\, dt,$$

where in the second integral our choice of u and x are constrained only by the condition $x(T) = x(T, T)$. Write $x(T, T) = x(c, T)$, the state of the system at time T starting in state c at time 0 associated with the finite variational process over $[0, T]$. Then (4.5) yields the inequality

$$V(c, \infty) \leqq V(c, T) + V(x(c, T), \infty). \tag{4.6}$$

Hence, we can obtain an estimate of the difference between $V(c, \infty)$ and $V(c, T)$ if we obtain an estimate for $x(c, T)$ as $T \rightarrow \infty$.

Observe that the estimate for $V(c, \infty)$ is readily obtained by using a convenient approximate policy of the type described in §2.

The estimate for $x(c, T)$ is not readily obtained in general. Let us indicate how elementary arguments yield the result for the problem of minimizing

$$J(x) = \int_0^T [\dot{x}^2 + x^2 + x^4]\, dt, \tag{4.7}$$

where $x(0) = c$.

It is clear from the form of the integrand that if $c > 0$, then x is monotone decreasing. For, as indicated in Fig. 1, if x reached a turning point and started to increase, we could replace it by the dotted curve, obtaining obviously a smaller value of the integral. The Euler equation is

$$\ddot{x} - x - 2x^3 = 0, \qquad x(0) = c, \qquad \dot{x}(T) = 0. \tag{4.8}$$

If x decreases monotonically, the limit must be zero as $T \rightarrow \infty$. From the Poincaré-Lyapunov theorem, we know that all solutions of (4.8) which approach zero as $t \rightarrow \infty$ have an asymptotic expansion of the form $c_1e^{-t} + c_2e^{-2t} + \cdots$. Using this information in conjunction with the preceding results, we readily obtain an asymptotic series for $V(c, T)$ as $T \rightarrow \infty$.

5. Further problems. The technique we have used here to obtain the asymptotic behavior of the state variables and the control variable is quite special and does not extend to the multidimensional case, to control processes with constraints, to more general control processes involving

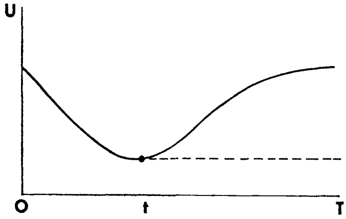


FIG. 1

distributed parameters, to general stochastic control processes, or to adaptive control processes. Some partial results can be derived, but on the whole there appears to be a need for a development of some new techniques.

We feel that it is worthwhile, in one case at least, to show that the expected results actually hold.

For asymptotic results in dynamic programming for processes of quite different nature, see [7, 8, 9, 10].

REFERENCES

[1] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, 1957.

[2] ———, *Adaptive Control Processes, A Guided Tour*, Princeton University Press, Princeton, New Jersey, 1961.

[3] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[4] R. Bellman, I. Glicksberg, and O. Gross, *Some aspects of the mathematical theory of control processes*, The RAND Corporation, R-313, January 1958.

[5] R. Bellman, *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1954.

[6] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, ASME J. of Basic Engrg., March 1961.

[7] R. E. Beckwith, *Analytic and computational aspects of dynamic programming processes of high dimension*, Ph.D. Thesis, Purdue University, 1959.

[8] D. Iglehart, Ph.D. Thesis, Stanford University, 1960.

[9] M. Freimer, *A dynamic programming approach to adaptive control processes*, Lincoln Lab. Report, 54-2, 1959.

[10] R. Bellman, *A Markovian decision process*, J. Math. Mech., 6 (1957), pp. 679–684.

[11] I. G. Petrovsky, *Partial Differential Equations*, Interscience, New York, 1954.

[12] R. E. Kalman, *The Theory of Optimal Control and the Calculus of Variations*, Mathematical Optimization Techniques, University of California Press, Berkeley, California, 1963, pp. 309–331.

[13] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Interscience, New York, 1962.

J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

MULTIVARIABLE LINEAR FILTER THEORY APPLIED TO SPACE VEHICLE GUIDANCE*

GERALD L. SMITH†

Abstract. Midcourse guidance of a spacecraft involves estimating the vehicle's trajectory from noisy observations and then computing velocity corrections on the basis of this estimate. The estimation procedure is regarded as a filtering problem and a guidance system concept is developed using multivariable linear filter theory. The ability of such a system to guide the spacecraft accurately and efficiently is demonstrated by the results of a digital computer simulation.

1. Introduction. Space-vehicle guidance presents a complex and exacting design problem for which we need the most modern design techniques available. Despite the complexity, this problem is recognizable as having the features of a control problem, and we therefore seek to apply control theory methods to its solution. In particular, recent developments in multivariable filter theory have provided a useful new approach to such problems [1]. In this paper we will show how these new ideas can be employed in a space-vehicle application. This application is described in more detail in NASA papers recently published [2, 3].

SYMBOLS AND NOTATION CONVENTIONS

H = submatrix in M relating δy to δx
k = subscript denoting the kth observation
K = weighting matrix in optimal filter
M = matrix relating δy to δx^*
n = observation error vector
N = covariance matrix related to observation error n
P = covariance matrix of $\tilde{x}$
Q = covariance matrix of u
r = position deviation of spacecraft from reference trajectory
R = covariance matrix of n
u = white noise
v = velocity deviation of spacecraft from reference trajectory
x = state vector
x^* = augmented state vector
y = observation vector
Φ = transition matrix
$(\ )^{-1}$ = inverse of matrix ( )
$(\ )^{T}$ = transpose of matrix ( )

* Received by the editors July 4, 1962, and in revised form September 30, 1963.

† National Aeronautics and Space Administration, Ames Research Center, Moffet Field, California.

$E(\ )$ = expected value of ( )
$(\hat{\ })$ = estimate of ( )
$(\tilde{\ })$ = error in estimate of ( )

2. Description of the space-vehicle guidance problem. We will be concerned here with the midcourse phase of guidance. The general nature of midcourse guidance is illustrated in block diagram form in Fig. 1. It is assumed that the vehicle is in free fall following injection, except for brief periods of thrusting when corrections to the trajectory are executed. The state of the vehicle—that is, its position and velocity—is therefore a function of the injection conditions and the trajectory dynamics. Since the injection conditions are not perfect, the vehicle departs from its desired or nominal trajectory, and it is the function of the control system to correct the course so that prescribed end-point conditions will be satisfied.

The first step in performing this function is assigned to instruments or sensors which measure observables related in some known way to the state. The sensors could be, for instance, optical instruments on board the vehicle, measuring the space angles between the lines of sight to certain celestial bodies. These angles are geometrically related to the vehicle position. The measurements are of course subject to errors, represented as observation errors in the figure. Generally it is neither necessary nor desirable to make continuous measurements, so some means must be provided for deciding if and when certain measurements should be made. This amounts to the selection of an optimum observation schedule which will be discussed later.

The next step is to make use of the observational data in the best possible manner to obtain an estimate of the state. This is seen to be essentially a filtering process, and filter theory can be applied to the design of the data

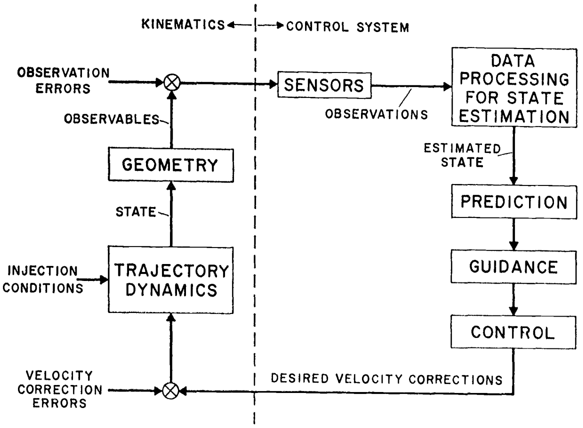


FIG. 1. *Schematic diagram of a midcourse guidance system*

processing system. The output of the filter is an estimate of the state, which is then used to estimate, or predict, what the end-point conditions would be if no course corrections were made. Next, a guidance law is employed to compute the velocity correction which would change the predicted end-point conditions to correspond to those prescribed. (The prescribed conditions might be, for instance, achievement of a given periapsis at the target moon or planet at a given time.)

Finally, a decision must be made as to whether or not the computed velocity correction should be made at the present time, and the correction implemented if the decision logic so indicates. The velocity correction when made then closes the control loop, acting through the trajectory dynamics to influence the state. The actual velocity correction, of course, is not quite the same as that intended, because of errors in the engine control mechanism.

3. Design of the guidance system. Having separated the midcourse guidance problem into distinguishable elements, we now can proceed to the application of design techniques to each of these elements. The sensor and control element designs will not be discussed, and we will dispose of the guidance law briefly by stating that in our studies we have used a linear prediction law. Thus we will concentrate on the trajectory estimation and decision aspects of the system.

First consider the trajectory estimation subsystem. Here we assume a sequence of observations, perturbed by additive errors, which are to be processed in the order in which they are received to maintain a continuous estimate of the state. The injection conditions and observation errors are not known exactly, hence can be described only probabilistically. Thus, the series of observations is regarded as a stochastic process (assumed discrete here since isolated observation times are presupposed). This stochastic process is generated by physical phenomena which can be represented in block diagram form as shown in Fig. 2. The injection conditions are actually trajectory initial conditions and thus not, strictly speaking, an input. The state is a 6 vector of position and velocity which can be expressed as a function of the injection conditions:

$$x(t) = f[x(t_0)]. \tag{1}$$

The observables constitute a vector having as components all those physical quantities to be measured by the sensors. The observation errors, $n(t)$, are represented as the output of a linear dynamic system excited by white noise, the standard engineering trick, valid when only second-order statistics are concerned. It is noted that this representation can be used for any type additive observation errors—for instance, bias type errors are

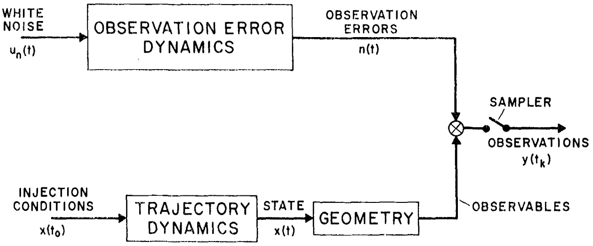


FIG. 2. *Observation process*

associated with dynamics having very long time constants. The sampler represents the selection of a particular one (or set) of the observables for measurement. The resulting observation is designated $y(t_k)$, the k subscript being used to index the time of a member of the sequence of observations. For convenience, the observation can be written as a function of $x(t_k)$ and $n(t_k)$:

$$y(t_k) = g[x(t_k), n(t_k)] \tag{2}$$

or more compactly,

$$y(t_k) = g[x^*(t_k)], \tag{3}$$

where x^* is an augmented state vector having as components all the components of both x and n.

It is now assumed that the statistics of the injection conditions $x(t_0)$ and of the white noise $u_n(t)$ are known. (If $x(t_0)$ and $u_n(t)$ are gaussian, only the means and covariance matrices are required.) If the trajectory dynamics (i.e., the vehicle equations of motion), the error dynamics, and the geometry equations are also known, then the $y(t_k)$ stochastic process is completely specified as soon as the observation schedule is stipulated.

Having defined the observation process, we now wish to develop the equations needed to process the observational data. We assume that we desire an optimal linear estimate of the state. This estimate will also contain an estimate of the observational error vector n. That is, we obtain an estimate of $x^*(t)$ which we call $\hat{x}_k{}^*(t)$. The k subscript means that the estimate is based on a total of k observations.

Assuming that we wish to process one observation at a time, the linear estimation equations are of the form

$$\hat{x}_k{}^*(t_k) = \hat{x}_{k-1}^*(t_k) + K(t_k)\{y(t_k) - g[\hat{x}_{k-1}^*(t_k)]\}, \tag{4}$$

$$\hat{x}_{k-1}^*(t_k) = f_{k,k-1}[\hat{x}_{k-1}^*(t_{k-1})], \tag{5}$$

where $K(t_k)$ is a weighting matrix to be described later, and $\hat{x}_{k-1}^*(t_k)$ is the estimate at time t_k based on the previous $k - 1$ observations. The quantity $\hat{x}^*_{k-1}(t_{k-1})$ is known as a result of processing the previous observation, $y(t_{k-1})$, and it is clear that updating this estimate to time t_k is simply a matter of using the equations which describe the dynamics of the x^* process, (5).

The computation flow diagram is as shown in Fig. 3. It is seen that the essential elements are a model of the kinematics and the matrix K. The model of the kinematics simulates the vehicle equations of motion, the error dynamics, and the geometrical relations between the state and observables. The operation of the system is described as follows. After injection, but before any observations have been made, the best estimate of x is based solely upon a priori knowledge of injection conditions, and the best estimate of n is zero. Thus, these are inserted as initial conditions on the kinematics equations. When an observation is made and the data is to be processed, the equations are integrated until computer time equals observation time. Then the estimated observation is computed from this updated estimate of x^*, compared with the actual data and the residual weighted by the matrix K to produce an incremental change in the estimated position, velocity, and observation error. The new estimated state variables serve as new conditions on the kinematics equations when the entire process is repeated to process the next observation.

The optimality of the data-processing system described obviously depends on the weighting matrix K. Linear filter theory is used to derive the equations by which K is computed. To facilitate this derivation it is convenient to linearize the equations of the observation process. This linearization is accomplished by expanding in a Taylor's series about the mean of the random variable $x^*(t)$. Thus, $x^*(t) = Ex^*(t) + \delta x^*(t)$, where $\delta x^*(t)$ now has zero mean and the same covariance matrix as $x^*(t)$.

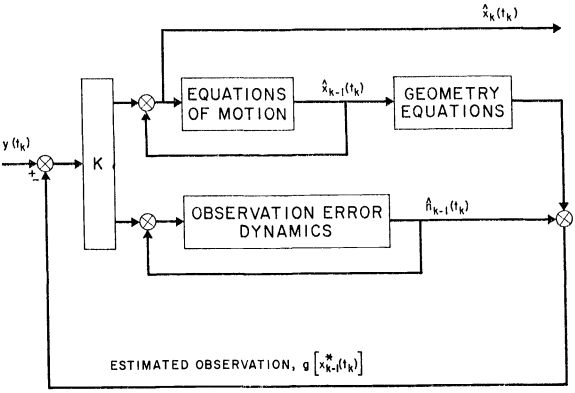


FIG. 3. *Trajectory estimation system*

The vehicle equations of motion and the observational error equations so linearized are written in terms of the transition matrices Φ_x and Φ_n.

$$\delta x(t) = \Phi_x(t, t_0)\delta x(t_0). \tag{6}$$

$$\begin{aligned} n(t) &= \Phi_n(t, t_0)n(t_0) + \int_{t_0}^{t} \Phi_n(t, \tau)u_n(\tau)\, d\tau \\ &= \Phi_n(t, t_0)n(t_0) + u'(t, t_0). \end{aligned} \tag{7}$$

In (6) there is no forcing function since the vehicle is assumed to be in free fall. In (7) the forcing function $u_n(t)$ appears under the integral, and the entire integral is replaced by a new function $u'(t, t_0)$ for convenience. It is noted that since $u_n(t)$ is uncorrelated with $n(t_0)$, $u'(t, t_0)$ is also uncorrelated with $n(t_0)$.

Equations (6) and (7) may now be combined:

$$\begin{Bmatrix} \delta x(t) \\ n(t) \end{Bmatrix} = \left[\begin{array}{c|c} \Phi_x(t, t_0) & 0 \\ \hline 0 & \Phi_n(t, t_0) \end{array}\right] \begin{Bmatrix} \delta x(t_0) \\ n(t_0) \end{Bmatrix} + \begin{Bmatrix} 0 \\ u'(t, t_0) \end{Bmatrix}, \tag{8}$$

or in more compact form:

$$\delta x^*(t) = \Phi(t, t_0)\delta x^*(t_0) + u(t, t_0). \tag{9}$$

The statistics necessary to describe the random process δx^* are the covariance matrices

$$\begin{aligned} P(t_0) &= E[\delta x(t_0)\delta x^T(t_0)], \\ R(t_0) &= E[n(t_0)n^T(t_0)], \\ Q_n(t) &= E[u_n(t)u_n^T(t)]. \end{aligned} \tag{10}$$

In combined form we write

$$\begin{aligned} P^*(t_0) &= \left[\begin{array}{c|c} P(t_0) & 0 \\ \hline 0 & R(t_0) \end{array}\right], \\ Q(t) &= \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & Q_n(t) \end{array}\right]. \end{aligned} \tag{11}$$

The observation is also expressed in terms of a deviation quantity, $\delta y(t_k)$, and the geometry equations are linearized to obtain:

$$\begin{aligned} \delta y(t_k) &= H(t_k)\delta x(t_k) + n(t_k) \\ &= [H \mid I] \begin{Bmatrix} \delta x(t_k) \\ n(t_k) \end{Bmatrix} \\ &= M(t_k)\delta x^*(t_k), \end{aligned} \tag{12}$$

where H is a matrix of partial derivatives of the observables with respect to the state variables.

The linear estimation equations are now written in terms of the deviation quantities:

$$\delta\hat{x}_k{}^*(t_k) = \delta\hat{x}_{k-1}^*(t_k) + K(t_k)[\delta y(t_k) - M(t_k)\delta\hat{x}_{k-1}^*(t_k)], \tag{13}$$

$$\delta\hat{x}_{k-1}^*(t_k) = \Phi(t_k\,,\, t_{k-1})\delta\hat{x}_{k-1}^*(t_{k-1}). \tag{14}$$

For this linear problem the equations for the computation of the optimal K have been derived by Kalman [1]:

$$K = P_{k-1}^* M[MP_{k-1}^* M^T]^{-1}, \tag{15}$$

where

$$P_{k-1}^* \equiv E(\delta\tilde{x}_{k-1}^* \delta\tilde{x}_{k-1}^{*T})$$

and

$$\delta\tilde{x}_{k-1}^* \equiv \delta x^* - \delta\hat{x}_{k-1}^* .$$

The argument of all these quantities is t_k, omitted here for simplicity. The covariance matrix P^* is computed from the recursion formulas as given by [1]:

$$P_k{}^*(t_k) = P_{k-1}^*(t_k) - K(t_k)M(t_k)P_{k-1}^*(t_k), \tag{16}$$

$$P_{k-1}^*(t_k) = \Phi(t_k\,;\, t_{k-1})P_{k-1}^*(t_{k-1})\Phi^T(t_k\,;\, t_{k-1}) + N(t_k\,,\, t_{k-1}), \tag{17}$$

where N is the covariance matrix of u. Now, although the K so computed is not the optimal K for the original nonlinear problem, it certainly is approximately the optimal K to the same degree that the linearized equations approximate the original nonlinear equations. This approximation has been demonstrated by means of computer simulation to be good as long as the actual state does not depart radically from the reference (mean value).

The computation of K is seen to be straightforward. At the time of the kth observation, the matrix M and the transition matrix $\Phi(t_k\,;\, t_{k-1})$ from the last observation must be computed. The latter can be done in a number of ways. One is to integrate a set of perturbation equations, each with a set of suitable initial conditions at the time of the previous observation, to give the several columns of the matrix. The M matrix and the coefficients of the perturbation equations are functions of the state variables and are computed using the estimated values of these variables. It should be noted that using the estimated state variables, in effect, amounts to linearizing about the estimated state. This is clearly the correct procedure because at each step of the estimation process, the mean of the conditional random vector $(x^* \mid y_1\,, \cdots, y_k)$ is $\hat{x}_k{}^*$, and $\tilde{x}_k{}^*$ thus has zero mean which is required if the foregoing is to result in an unbiased estimate. In actual practice if the actual trajectory does not depart very far from the nominal, the nominal may be used with little effect on the results. However, in some situations, such as an abort, the departure is substantial. Then either the estimated trajectory or a new nominal trajectory sufficiently close to the actual must be used.

The matrix $N = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & N' \end{array}\right]$ is also required. This can be computed from the relationship

$$N'(t_k, t_{k-1}) = \int_{t_{k-1}}^{t_k} \Phi_n(t_k, \tau) Q_n(\tau) \Phi_n^T(t_k, \tau)\, d\tau, \tag{18}$$

where $Q_n(t)$ would presumably be a stored matrix, and Φ_n is computed as part of Φ. In many cases $Q_n(t)$ may be constant, or at least only slowly time varying; also Φ_n may be a function only of $t - \tau$ (i.e., non-time-varying). In these cases the computation (18) is substantially simplified.

In the data processing scheme we have described above, each observation is a vector whose components are a set of measurements made at the same time. Now when it is seen that the observations can be processed one at a time, it is natural to consider the further possibility that the components of each observation can themselves be processed individually. In fact, they can be, and that is the reason we have chosen to write the estimation equations in the particular form given. If one piece of data (i.e., a measurement) has been processed at time t_k, then to process another measurement taken at the same time, (5) and (17) are not used since there is no time transition. Equations (4) and (16) give the new estimate and the new P^* matrix. In reference to Fig. 3 this means simply that the integration parts of the computation are not employed. We note that in processing data in this way MP^*M^T is always a scalar, so the matrix inversion required in (15) is avoided.

If the observations are uncorrelated, some simplification of the data processing equations is possible. By "uncorrelated observations" we mean that the errors in any pair of observations are statistically independent. This may be because the time between the observations is large compared to the "time constants" of the error dynamics, or because the observations are of basically different types. In any case, this means that the previously processed data contain no information regarding the present error in observation; hence, the estimate $\hat{n}(t_k)$ is zero and need not be computed. The error in estimate of $n(t_k)$, namely $\tilde{n}(t_k)$, is of course just $n(t_k)$, and the portion of the $P^*(t_k)$ matrix containing the covariance matrix of $\tilde{n}(t_k)$ is seen to be simply $N'(t_k\, ;\, t_{k-1})$. Furthermore, N' is seen in this case to be a function only of t_k, hence could be a simple stored matrix. The result is that the computation of the P^* matrix can be simplified, omitting the rows and columns having to do with n, thus reducing the order of the matrix operations required. The estimation equations can then be written in the form:

$$\hat{x}_k(t_k) = \hat{x}_{k-1}(t_k) + K(t_k)\{y(t_k) - g_k[x_{k-1}(t_k)]\}, \tag{19}$$

$$\hat{x}_{k-1}(t_k) = f_{k,k-1}[\hat{x}_{k-1}(t_{k-1})], \tag{20}$$

$$K = P_{k-1}H^T[HP_{k-1}H^T + N_k']^{-1}, \tag{21}$$

$$P_{k-1}(t_k) = \Phi(t_k\,;\,t_{k-1})P_{k-1}(t_{k-1})\Phi^T(t_k\,;\,t_{k-1}), \tag{22}$$

$$P_k = P_{k-1} - KHP_{k-1}\,. \tag{23}$$

Consider now what happens to the estimation process when a velocity correction is made. This situation is treated exactly the same as an observation. If the velocity correction is assumed to be monitored, the actual correction (Δv) differs from the desired (Δv_d) because of the control error, and from the observed (Δv_m) because of the monitoring error. The a priori statistics of the control error and monitoring error are assumed known. The estimate of Δv before the correction is Δv_d, and we obtain the new estimate by the formula

$$\Delta\hat{v} = \Delta v_d + K_v[\Delta v_m - \Delta v_d], \tag{23a}$$

where S is the covariance matrix of control error, T is the covariance matrix of monitoring error, and $K_v = S(S + T)^{-1}$. This $\Delta\hat{v}$ is added to the estimated velocity vector in the trajectory estimation system. The error in estimate, $\Delta\tilde{v} = \Delta v - \Delta\hat{v}$, has zero mean and covariance matrix $(S - K_vS)$, which adds to the velocity portion of the P matrix. The system is then ready to process new data.

This method of handling velocity corrections is seen to be quite simple and is valid even for very large corrections, such as might occur in an abort maneuver for instance. The assumption of an instantaneous correction is not valid, however, when the correction is large. In this case a continuous correction estimation, analogous but more complicated than the above procedure, would have to be implemented.

One pitfall we must avoid in using the foregoing theory is that the P matrix gives us the correct values of the error statistics only if we have employed the correct model of the observation process. In practice, our model can never be perfect. For instance, the equations of motion simulated in the system described here are only approximations in that the gravitational effects of only a few celestial bodies are included and the astrodynamic constants used in the equations are not known perfectly.

Furthermore, any digital integration routine employed is in itself an approximation to true integration and thus generates errors.

Such errors can be seen from Fig. 3 to enter the system in exactly the same manner as do observation errors. Hence, they can be estimated and compensated for in exactly the same manner as described before—at the expense of a more complex system, of course. In practice one would first determine the gross effects of such errors and implement only as sophisticated a system as is justified by the accuracy desired; that is, for the sake of simplicity, in general we would accept "off-design" performance a bit poorer than might theoretically be attainable.

An interesting by-product of the consideration of errors due to the imperfect knowledge of the astrodynamic constants is the thought that we have here a ready-made technique for obtaining by direct experiment a better estimate of these constants. For instance, a properly instrumented circumlunar vehicle could be used to improve current estimates of the earth-moon distance and the earth and moon gravitational constants. It may be noted that Pioneer V tracking data were used to obtain a good estimate of the astronomical unit [4] although the shot was not designed specifically for this experiment and the data processing technique employed was somewhat different from that described here.

To implement an estimating procedure for the astrodynamic constants, we define the uncertainties in these constants as additional random variables, augmenting the state vector with these variables. The transition matrix and the P and K matrices are likewise augmented. If we then solve the variance equation (using a digital computer), we can obtain a measure of the improvement in the knowledge of these constants which could be obtained from a prescribed sequence of observations.

Now, to complete the design of the midcourse guidance system, we must consider the selection of an appropriate schedule of observations and velocity corrections. Specifically, we would like to find an optimal schedule, where the optimality criter on must take into account practical considerations such as the cost of executing the required operations and the interactions that exist between these and other operations involved in the over-all mission. The problem is seen to be complex and, worse, rather ill-defined. Thus, an attempt to find a true optimum does not seem practical, at least at present. In our studies we have resorted instead to a cut-and-try approach. First, a reasonable operational schedule is selected and its performance is computed. Then this schedule is varied systematically and the change in performance noted. Problem solutions are obtained fairly rapidly on the digital computer (typically about 10 minutes on the IBM 7090 using the particular programs we have written). Thus, a reasonable schedule can be generated without too much effort.

To obtain a true optimum schedule it should be possible, at least in principle, to mechanize the described variational procedure so all the work is done by the computer. However, because of the varied nature of the conditions one wishes to place on the schedule, a program for doing this must necessarily be quite complex, and as yet we have not attempted it.

4. Results of simulation study. Computer simulation studies carried on at Ames Research Center have demonstrated that the midcourse guidance system described here can do an effective job in the assigned guidance problem. Some of the results of these studies will now be presented for the case of a hypothetical $6\frac{1}{2}$-day circumlunar flight in which the entire guidance system is to be carried on board the space vehicle. The assumed conditions for this mission are summarized as follows:

1. Each observation involves sighting upon either the earth or moon and measuring the direction of the line of sight (two angles) and the angle subtended by the disk of the planet.

2. Observation errors have zero mean and are uncorrelated from one observation to the next. The error statistics are represented by a diagonal covariance matrix, Q, whose elements are of the form

$$\sigma^2 = 100 + (0.001\gamma)^2 \text{ seconds of arc squared,} \tag{24}$$

where γ is half the subtended angle; that is, the errors are assumed to be greater when the vehicle is nearer the planet being observed.

3. Midcourse velocity corrections are computed using a simple linear prediction fixed-time-of-arrival scheme. The corrections are intended to null the position deviation of the vehicle from a reference perilune on the outboard leg and from a reference atmospheric entry point on the return leg.

4. Velocity correction errors are one degree rms in direction and 0.1 m/sec in magnitude. Errors in the measurement of the correction are 0.01 m/sec rms in each of three Cartesian coordinate directions.

5. Rms injection errors are 1 km and 1 m/sec in each of the three Cartesian coordinates used in the computations.

It should be pointed out that these assumptions are not intended to describe, even tentatively, any actual mission configuration. Although hypothetical, they are nevertheless realistic and can be used to illustrate the operation of the system described.

One of the trajectories studied is shown in Fig. 4. On the trajectory is indicated one specific observation and velocity correction schedule for which we have obtained performance data. No attempt has been made to optimize this schedule, which consists of 77 observations and 6 velocity corrections. Earth observations are shown as tick marks, moon observations as stars, and velocity corrections as circles.

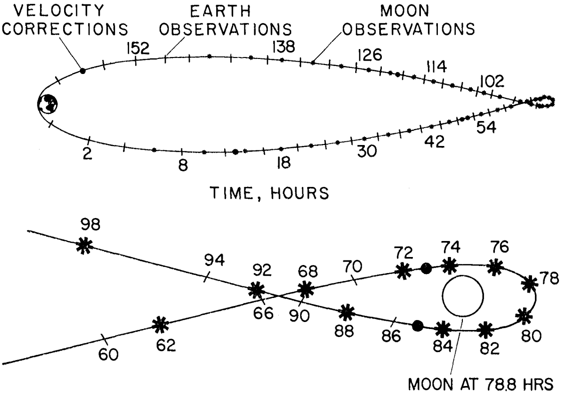


FIG. 4. *Schedule of observations and velocity corrections*

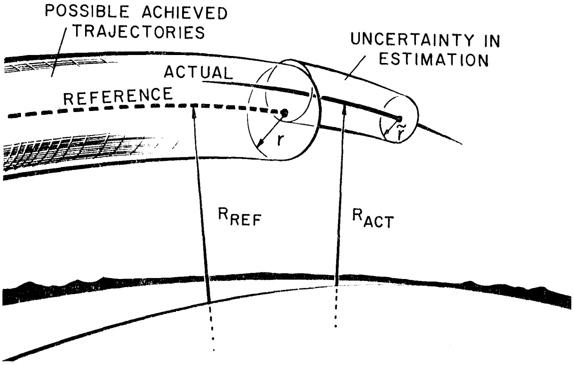


FIG. 5. *Errors at time of reference perigee*

The manner in which we describe the performance of the guidance system is shown in Fig. 5. The dotted line indicates the reference trajectory selected to provide a near passage of the moon and a safe entry into the earth's atmosphere. A measure of the guidance effectiveness is the difference between the actual and reference trajectories, the statistics of which we compute in our studies. The deviation in position is called r. The deviation in velocity is in like manner called v but not illustrated in the figure. Similarly, we represent the difference between the actual and estimated trajectories in terms of $\tilde{r}$ and $\tilde{v}$, the rms values of which are obtained from the covariance matrix of estimation errors P. We also compute the rms variation in perigee altitude, which is of significance for establishing the probability of safe entry, regardless of the achievement of a particular landing site.

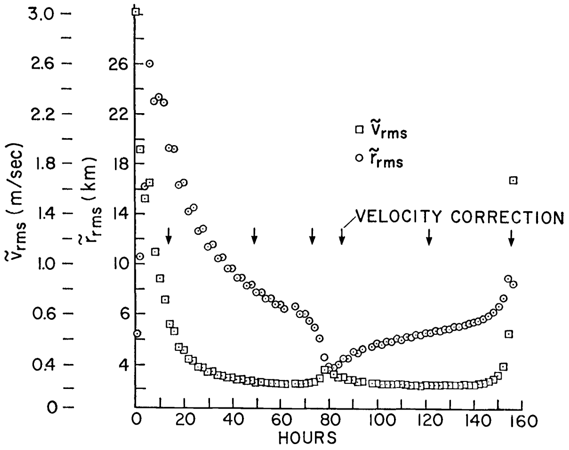


FIG. 6. *Time history of the rms estimation errors*

TABLE 1. *Results at end points—rms values*

		AT MOON	AT EARTH
MISS	PERIAPSIS VARIATION (km)	2.3	1.1
	r(km)	8.6	26.4
	v(m/sec)	2.2	23.8
UNCERTAINTY	$\tilde{r}$(km)	2.9	15.0
	$\tilde{v}$(m/sec)	0.27	13.2
	TOTAL APPLIED ΔV m/sec		20.0

Fig. 6 shows what happens to the rms errors in estimation as the flight progresses. The points indicate the rms errors after each observation, and the times at which velocity corrections are made are shown by arrows. It is seen that rms position estimation errors do not exceed 26 km. Rms velocity estimation errors are highest at the beginning of flight and rise again near the end but are never greater than 0.03 per cent of vehicle velocity. Thus, the assumption of small deviations is valid and the linearization approach employed in the analysis should be reasonable.

Table 1 summarizes the end-point data obtained for the case described, showing how well the guidance system has performed at the times of nominal perilune and perigee. In the first column are perilune results and in the second column results at perigee. Note that the rms variation in virtual perigee is only 1.1 km, indicating a high probability of safe atmospheric entry. The next two numbers of 26.4 km and 23.8 m/sec for the rms position and velocity deviations from reference are given at virtual perigee but are of the same order of magnitude at the time of actual atmosphere entry.

The next two figures are the rms values of the errors in knowledge of position and velocity, 15.0 km and 13.2 m/sec. These figures are to the terminal guidance system what the uncertainty in knowledge of injection conditions are to the midcourse guidance system. They result in an uncertainty in the landing location which we have not calculated. Of course, any tracking information acquired during the terminal phase would reduce this uncertainty.

The last figure in the table shows the total corrective velocity required for making the six corrections for the $6\frac{1}{2}$-day flight—a modest 20 m/sec rms.

The performance at perilune as shown in the second column is seen to be similar to that at perigee.

5. Conclusions. The simulation results presented demonstrate that the described guidance system concept is capable of providing excellent mission performance. It is seen that an important advantage of the system is a high degree of versatility in that a fixed observation and velocity correction schedue need not be adhered to and there is no dependence upon earth-vehicle communication. The required calculations are not overly complex, so it is felt that the on-board digital computer can be of modest size and power consumption.

REFERENCES

[1] R. E. Kalman, *New methods and results in linear prediction and filtering theory*, Technical Report 61-1, RIAS, Baltimore, Maryland, November 1960.

[2] G. L. Smith, S. F. Schmidt, and L. A. McGee, *Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle*, NASA TR R-135, 1962.

[3] J. D. McLean, S. F. Schmidt, and L. A. McGee, *Optimal filtering and linear prediction applied to a midcourse navigation system for the circumlunar mission*, NASA TN D-1208, 1962.

[4] J. B. McGuire and L. Wong, *A dynamical determination of the astronomical unit by a least squares fit to the orbit of Pioneer V*, Report 2301-0004-RU-000, Space Technology Laboratories Inc., Los Angeles, California, May 15, 1961.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

OPTIMIZATION, A MOMENT PROBLEM, AND NONLINEAR PROGRAMMING*

LUCIEN W. NEUSTADT†

1. Introduction. Certain problems from optimal control theory and the optimization of trajectories can be formulated as follows.

Given an $n \times r$ matrix Y whose elements $y_j^i(\cdot)$ are continuous, real-valued functions on [0, 1], a normed linear function space $\mathfrak{F}$ whose elements $u(\cdot)$ are Lebesgue integrable functions from [0, 1] to E_r (real r-dimensional Euclidean space), and a vector c in E_n ; find an element $u^*(t) \in \mathfrak{F}$ of minimum norm satisfying the equation

$$\int_0^1 Y(t)u^*(t)\,dt = c. \tag{1}$$

In this paper we restrict ourselves to spaces $\mathfrak{F}$ with norm defined by

$$\| u(\cdot) \|_p = \int_0^1 | u(t) |_p \, dt,$$

where $1 \leqq p \leqq \infty$, and $| u |_p$, for any vector $u = (u_1, \cdots, u_r)$, is defined by

$$\begin{aligned} | u |_p &= \left(\sum_{i=1}^r | u_i |^p \right)^{1/p}, \qquad p < \infty, \\ | u |_\infty &= \max_{1 \leqq i \leqq r} | u_i |. \end{aligned} \tag{2}$$

For such spaces $\mathfrak{F}$, the above described problem does not, in general, have a solution. It is necessary to embed $\mathfrak{F}$ (while preserving the norm) in the conjugate space $\mathfrak{B}^*$ of an appropriate Banach space $\mathfrak{B}$ (to which the vector functions $y^i(\cdot)$, which make up the rows of Y, belong) and interpret (1) as conditions to be satisfied by an element of $\mathfrak{B}^*$. We shall show that a desired element of $\mathfrak{B}^*$ with minimum norm does exist, that it can be characterized in a relatively simple manner, and that it can be thought of as corresponding to a function $u^*(t)$, which is a linear combination of "delta functions."

Before putting the initially given problem in this new formulation (in §4), we shall prove a general theorem in the theory of moments. This theorem, which is stated and proved in §2, provides the existence and characterization of the minimizing functional. When applied to the problem

* Received by the editors July 22, 1963, and in revised form October 14, 1963.

† Aerospace Corporation, Los Angeles, California. On leave of absence at College of Engineering, The University of Michigan, Ann Arbor, Michigan.

described in the preceding paragraph, it makes it possible to reduce the original variational problem to a relatively simple (nth order) problem in nonlinear programming which is particularly suitable for solution on a digital computer. This is described in detail in §5. Although the theorem, in its essential features, is well-known, the proof we shall give is novel.

In §2 we prove another general theorem, as a consequence of which we show that a minimizing "function" $u^*(t)$ can be constructed with at most n "impulses."

In §6 the general theory is applied to the problem of determining a minimum-fuel midcourse correction for a space flight. A particularly noteworthy result is that if the equations of motion of the space vehicle can be approximated by equations of a special form, and if n ($n \leqq 6$) components of the vehicle's position and velocity vectors must take on given values at a given terminal time, then a minimum-fuel maneuver for achieving these end values will consist of not more than n impulsive corrections.

2. Theorems from the theory of moments. We now prove two theorems from the theory of moments.

The first theorem deals with the problem of finding a functional (on a Banach space) of least norm taking on given values at a certain finite number of given elements (or having a finite number of given "moments"). Our original problem will be put in this formulation in §4.

THEOREM 1. A. *Let $\mathfrak{B}$ be a Banach space, let $y^1, \cdots, y^n$ be n linearly independent elements of $\mathfrak{B}$, and let $c = (c_1, \cdots, c_n)$ be a given nonzero vector in E_n. Then there exists a functional $l^0 \in \mathfrak{B}^*$ such that $l^0(y^i) = c_i$ for each i, and $\| l^0 \| = \lambda$, where*

$$(3) \qquad \lambda = \sup_{\eta \in H} \eta \cdot c,$$

and

$$(4) \qquad H = \left\{ \eta: \quad \eta \in E_n, \left\| \sum_{i=1}^{n} \eta_i y^i \right\| = 1 \right\}$$

(the dot in (3) denotes the ordinary vector dot product, and $\| \ \|$ denotes the norm either in $\mathfrak{B}$ or in $\mathfrak{B}^*$). Further, if l is any element in $\mathfrak{B}^*$ satisfying the relations $l(y^i) = c_i$, $i = 1, \cdots, n$, then $\| l \| \geqq \lambda$ (i.e., l^0 is a minimum-norm solution of the equations $l(y^i) = c_i$).

B. *The supremum in (3) is attained. If $\bar{\eta} = (\bar{\eta}_1, \cdots, \bar{\eta}_n)$ is any member of H which achieves the maximum, and l is any element of $\mathfrak{B}^*$ such that $l(y^i) = c_i$ for each i, then $\| l \| = \lambda$ if and only if*

$$(5) \qquad l\left(\sum_{i=1}^{n} \bar{\eta}_i y^i\right) = \| l \| \left\| \sum_{i=1}^{n} \bar{\eta}_i y^i \right\|$$

(by definition of H, the right-hand side in (5) is equal to $\| l \|$).

C. *An element* $l \in \mathcal{B}^*$ *satisfies the relations* $l(y^i) = c_i$ *for* $i = 1, \cdots, n$, *and* $\| l \| = \lambda$ *if and only if*

(a) *there exists a vector* $\hat{\eta} = (\hat{\eta}_1, \cdots, \hat{\eta}_n)$ *in* H *such that* $l(\sum \hat{\eta}_i y^i) = \| l \| = \hat{\eta} \cdot c > 0$, *and*

(b) $l(\sum \eta_i y^i) = 0$ *whenever* $\sum \eta_i c_i = 0$.

We note that the number λ in (3) can also be given as

$$(6) \qquad \lambda = \sup_{\eta \in E_n} \frac{\eta \cdot c}{\| \eta \cdot y \|} = \left[\inf_{\eta \in E_n} \frac{\| \eta \cdot y \|}{\eta \cdot c} \right]^{-1} \left[\inf_{\eta \cdot c = 1} \| \eta \cdot y \| \right]^{-1},$$

where $\eta \cdot y$, with $\eta = (\eta_1, \cdots, \eta_n)$, is used to denote $\sum_{i=1}^n \eta_i y^i$. Also, except possibly for a positive scalar multiple, the same vectors η achieve each of the extrema in (6) and (3).

Relations (3) and (4) (or (6)), together with (5), may be used to obtain the minimum-norm functional which is being sought. As a result of (3) (or (6)), the variational problem in $\mathcal{B}^*$ has been reduced to a variational problem in the n-dimensional space E_n, since a minimizing functional is usually easy to obtain, because of the necessity and sufficiency of relation (5), once a maximizing element in (3) has been found.

The basic result of Theorem 1 is contained in Part A. This was first proved by Hahn [1], and is a consequence of the Hahn-Banach theorem. Earlier references pertaining to the same result in particular Banach spaces, together with a proof of the general theorem, are given in Dunford and Schwartz [2, p. 86]. The same problem was also treated in detail by Krein [3, Article IV], who pointed out the necessity of relation (5) in order that l be a minimum-norm solution. Krasovskii [4, 5] first applied these results to specific optimization problems of the type mentioned in the introduction. In many of these problems, relation (5) essentially specifies the functional l (or the function $u^*(\cdot)$ in the original problem formulation), once $\bar{\eta}$ is known.

The application of Theorem 1 to particular optimization problems has also been derived independently. The case of spaces $\mathfrak{F}$ with the norm defined by

$$\| u(\cdot) \| = \left[\int_0^1 [| u(t) |_p]^p \, dt \right]^{1/p},$$

where $1 < p < \infty$, or $\| u(\cdot) \| = \sup_{0 \leqq t \leqq 1} | u(t) |_\infty$, has been extensively studied by Krasovskii [4, 5], the author [6], Reid [7], and Kreindler [8] among many others. In this problem, relation (5) usually defines the minimum-norm function $u^*(\cdot)$ uniquely, as is pointed out in the above-cited references. In [6], a successive approximation scheme is suggested for the computation of the extremum in (6).

We shall present a proof of Theorem 1 which is distinct from that given in [1] or [3], and which, because of its geometrical character, is felt to be particularly illuminating. The basic idea of this proof can be found in [6]. Very similar arguments are presented in [8]. Antosiewicz [9] has also made use of this geometric viewpoint in considering a closely related problem.

The proof given by Reid [7] is noteworthy in that it is distinct from any of the others discussed above. It is partially based on Part C of Theorem 1, specialized to the problem he considers. Some additional references to earlier work on this problem are also given in this paper.

A large number of related references from the engineering literature can be found in [8].

Proof of Theorem 1. Let y denote the n-tuple $(y^1, \cdots, y^n)$. If $l \in \mathcal{B}^*$, we shall, for ease of notation, denote the element $(l(y^1), \cdots, l(y^n))$ of E_n by $l(y)$.

Consider the linear operator T from $\mathcal{B}^*$ to E_n defined by $T(l) = l(y)$. Define $S_\alpha = \{l\colon\ l \in \mathcal{B}^*,\ \| l \| \leqq \alpha\}$ for every positive number α, and let $TS_\alpha = C_\alpha$. We denote S_1 by S and C_1 by C.

LEMMA. *For every $\alpha > 0$, C_α is a convex compact set in E_n containing the origin as an interior point.*

Proof. Since S_α is convex and T is linear, C_α is convex. It is clear that C_α is symmetric with respect to the origin. Hence, to show that the origin is an interior point of C_α, it is sufficient to show that C_α is not contained in any subspace of dimension less than n. Suppose the contrary. Then there is a nonzero vector $\eta \in E_n$ such that $\eta \cdot T(l) = l(\eta \cdot y) = 0$ for every $l \in S_\alpha$, and indeed for every $l \in \mathcal{B}^*$. But by a well-known corollary to the Hahn-Banach theorem (see, for example, [2, Corollary 14, p. 65]), this implies that $\eta \cdot y = 0$, contradicting the linear independence of the y^i.

It follows at once from the definition of the weak* topology in $\mathcal{B}^*$ [10, p. 37], that T is continuous from $\mathcal{B}^*$ to E_n in terms of the weak* topology in $\mathcal{B}^*$ and the ordinary Euclidean topology in E_n. Since S_α is compact in the weak* topology [10, Theorem 2.10.2, p. 37], C_α is compact. This completes the proof of the lemma.

Now consider the following question. What is the smallest positive number α for which $c \in C_\alpha$, or, equivalently (since $C_\alpha = \alpha C$), what is the smallest positive number α such that $\alpha^{-1}c \in C$? Since $c \neq 0$, the existence of such a number α follows from the lemma. Denote this number by λ, and let $\gamma = \lambda^{-1}c$. It is clear that γ is on the boundary of C.

There is a plane of support to C at each of its boundary points. We shall say that the nonzero vector $\eta \in E_n$ is an outward normal to C at a boundary point ζ if there is a plane P normal to η which is a support plane to C at ζ, and if η is directed away from C at ζ.

Let $\bar{\eta}$ be any outward normal to C at γ. Then

$$\bar{\eta} \cdot \gamma = \max_{\xi \in C} \bar{\eta} \cdot \xi. \tag{7}$$

If $\eta \in E_n$ ($\eta \neq 0$) is not an outward normal to C at γ, then

$$(8) \qquad \eta \cdot \gamma < \max_{\xi \in C} \eta \cdot \xi.$$

But, again using Corollary 14 in [2, p. 65],

$$(9) \qquad \max_{\xi \in C} \eta \cdot \xi = \max_{\| l \| = 1} l(\eta \cdot y) = \| \eta \cdot y \| .$$

Combining (7)–(9), and using the definition of γ, we finally obtain

$$(10) \qquad \lambda = \max_{\eta \in E_n} \frac{\eta \cdot c}{\| \eta \cdot y \|} = \max_{\eta \in H} \eta \cdot c.$$

(It is interesting to note that the maximum in (10) is attained by those, and only those, vectors η that are outward normals to C at γ.)

Now, since $\gamma = \lambda^{-1} c \in C$, there is an element $\bar{l} \in S$ such that $T(\bar{l}) = \gamma$, or, if we set $l^0 = \lambda \bar{l}$, $\| l^0 \| \leqq \lambda$ and $T(l^0) = (l^0(y^1), \cdots, l^0(y^n)) = (c_1, \cdots, c_n) = c$. On the other hand, if $l \in \mathfrak{B}^*$ and $l(y) = T(l) = c$, then $c \in C_{\| l \|}$. Then, by definition of λ, $\| l \| \geqq \lambda$. In particular, $\| l^0 \| \geqq \lambda$, or $\| l^0 \| = \lambda$.

Let $\bar{\eta}$ be any vector in H which achieves the maximum in (10), so that $\bar{\eta} \cdot c = \lambda$ and $\| \bar{\eta} \cdot y \| = 1$. If l is a member of $\mathfrak{B}^*$ such that $l(y) = c$ and $\| l \| = \lambda$, then $l(\bar{\eta} \cdot y) = \bar{\eta} \cdot c = \lambda = \| l \| \| \bar{\eta} \cdot y \|$. Conversely, if $l(y) = c$ and $l(\bar{\eta} \cdot y) = \| l \| \| \bar{\eta} \cdot y \|$, then $\| l \| = l(\bar{\eta} \cdot y) = \bar{\eta} \cdot c = \lambda$.

Finally, suppose that $l(y) = c$ and $\| l \| = \lambda$. Then $\eta \cdot c = 0$ implies that $l(\eta \cdot y) = \eta \cdot c = 0$. Conversely, suppose that there is an element $l \in \mathfrak{B}^*$ and a vector $\hat{\eta} \in H$ such that $l(\hat{\eta} \cdot y) = \| l \| = \hat{\eta} \cdot c > 0$, and such that $l(\eta \cdot y) = 0$ whenever $\eta \cdot c = 0$. Then, $\eta \cdot l(y) = 0$ whenever $\eta \cdot c = 0$, implying that $l(y) = \epsilon c$ for some scalar ϵ. But then $l(\hat{\eta} \cdot y) = \epsilon \hat{\eta} \cdot c = \hat{\eta} \cdot c > 0$, so that $\epsilon = 1$; i.e., $l(y) = c$. By what was proved above, this implies that $\| l \| \geqq \lambda$. But by (10), $\lambda \geqq \hat{\eta} \cdot c = \| l \|$, so that $\| l \| = \lambda$. This completes the proof of Theorem 1.

The following theorem will be used to characterize certain solutions of our basic problem.

THEOREM 2. *Let $y^1, \cdots, y^n$ be nonzero elements of a Banach space $\mathfrak{B}$. Let T be the map from $\mathfrak{B}^*$ to E_n defined by $T(l) = (l(y^1), \cdots, l(y^n))$. Suppose that there is a set D in $\mathfrak{B}^*$ with the following properties.*

(a) *$l \in D$ implies that $\| l \| = 1$.*

(b) *The convex hull of D is dense in the unit ball S of $\mathfrak{B}^*$ with respect to the weak* topology of $\mathfrak{B}^*$.*

(c) *If $\{l_k\}$, $k = 1, 2, \cdots$, is any sequence of elements in D, there is a subsequence $\{l_{k_j}\}$ of $\{l_k\}$, and an element l_∞ of D, such that $T(l_{k_j}) \to T(l_\infty)$ as $j \to \infty$.*

Then, if l is any element of $\mathfrak{B}^*$, there exist n elements (depending on l)

$l_1, \cdots, l_n$ *of* D *and* n *nonnegative numbers* $\lambda_1, \cdots, \lambda_n$ *such that* $T(l) = T(\sum_{i=1}^n \lambda_i l_i)$ *and* $\| \sum \lambda_i l_i \| \leqq \| l \|$.

Proof. Let K denote the convex hull of D, and let $C = TS$. Since the map T is linear, TK is convex. It follows at once from hypothesis (b) that TK is dense in C.

It is sufficient to prove the theorem for elements $l \in \mathcal{B}^*$ with $\| l \| = 1$. Thus, let $l \in \mathcal{B}^*$ where $\| l \| = 1$, and let $x' = T(l)$. If $x' = 0$, the theorem follows immediately. Thus, assume that $x' \neq 0$. Let $\alpha x' = x$ be on the boundary of C, where $\alpha \geqq 1$. Such a number α exists by virtue of the lemma used in proving Theorem 1.

Since TK is dense in C, there exist elements $x^k \in TK$ $(k = 1, 2, \cdots)$ such that $x^k \to x$ as $k \to \infty$. Since TK, the convex hull of TD, is in E_n, it follows from a theorem of Carathéodory [11, p. 35] that there are elements $l_i^k \in D$ and nonnegative numbers λ_i^k $(k = 1, 2, \cdots; i = 1, \cdots, n + 1)$ such that

$$x^k = \sum_{i=1}^{n+1} \lambda_i^k T(l_i^k), \qquad \sum_{i=1}^{n+1} \lambda_i^k = 1,$$

for each $k = 1, 2, \cdots$. Because of hypothesis (c) and the uniform boundedness of the numbers λ_i^k, we shall assume, without loss of generality, that (for each $i = 1, \cdots, n + 1$) there is a nonnegative number $\bar{\lambda}_i$, and an element $l_i \in D$, such that $\lambda_i^k \to \bar{\lambda}_i$ and $T(l_i^k) \to T(l_i)$ as $k \to \infty$. Hence,

$$x = \sum_{i=1}^{n+1} \bar{\lambda}_i T(l_i), \qquad \sum_{i=1}^{n+1} \bar{\lambda}_i = 1.$$

Let us first consider the case where the points $T(l_1), \cdots, T(l_{n+1})$ do not all lie on some hyperplane in E_n of dimension less than n. Then, these points constitute the vertices of a nondegenerate simplex $M \subset C$. Since $x \in M$ and x is on the boundary of C, x is on the boundary (i.e., a face) of M. This means that one of the numbers $\bar{\lambda}_i$—say $\bar{\lambda}_{n+1}$—vanishes. Thus

$$x = \sum_{i=1}^{n} \bar{\lambda}_i T(l_i) = T\left(\sum_{i=1}^{n} \bar{\lambda}_i l_i\right),$$

or

$$T(l) = x' = \alpha^{-1} x = T\left(\sum_{i=1}^{n} \lambda_i l_i\right),$$

where $\lambda_i = \alpha^{-1}\bar{\lambda}_i \geqq 0$, and $\sum_{i=1}^n \lambda_i = \alpha^{-1} \leqq 1$. Finally, $\| l_i \| = 1$ for each i, so that $\| \sum_{i=1}^n \lambda_i l_i \| \leqq \sum \lambda_i \leqq 1 = \| l \|$.

If the points $T(l_i)$ belong to a hyperplane of dimension less than n, we can apply the Carathéodory theorem to the convex hull of these points, and conclude that there are nonnegative numbers $\hat{\lambda}_i$ $(i = 1, \cdots, n + 1)$, at most n of which are positive, with $\sum_{i=1}^{n+1} \hat{\lambda}_i = 1$, such that $x = \sum_{i=1}^{n+1} \hat{\lambda}_i T(l_i)$. The remainder of the proof follows as above.

3. Description and characterization of some vector spaces. In order to apply Theorems 1 and 2 to our original problem, we shall introduce some linear vector spaces.

Let r be a fixed positive integer, and let p be a real number such that $1 \leqq p \leqq \infty$. Let q be the conjugate index of p: if $1 < p < \infty$, $q = p(p-1)^{-1}$; if $p = 1, q = \infty$; if $p = \infty, q = 1$.

Denote by $\mathfrak{F}_p$ the normed linear space of Lebesgue integrable functions from [0, 1] to E_r with the norm of an element $u(\cdot) \in E_r$ given by

$$\|u(\cdot)\|_p = \int_0^1 |u(t)|_p \, dt, \tag{11}$$

where $|u|_p$ is defined by (2).

Let $\mathcal{S}_q$ be the Banach space of continuous functions from [0, 1] to E_r, with the norm of an element $y(\cdot) \in \mathcal{S}_q$ defined by

$$\|y(\cdot)\|_{\infty,q} = \sup_{0 \leqq t \leqq 1} |y(t)|_q .$$

If $g(\cdot)$ is a function from [0, 1] to E_r, define the strong total p-variation (STV$_p$) of $g(\cdot)$ by

$$\mathrm{STV}_p g(\cdot) = \sup \sum_{i=1}^{\nu} |g(t_i) - g(t_{i-1})|_p ,$$

where the supremum is taken over all finite partitions $0 = t_0 < t_1 < \cdots < t_\nu = 1$ of [0, 1]. If $\mathrm{STV}_p g(\cdot) < \infty$, we say that $g(\cdot)$ is of strong bounded p-variation (see [10, p. 59]).

Let $\mathcal{G}_p$ denote the Banach space which consists of all functions $g(\cdot)$ from [0, 1] to E_r that are of strong bounded p-variation, satisfy the relation $g(0) = 0$, and are continuous from the right in (0, 1), with the norm in $\mathcal{G}_p$ given by $\|g(\cdot)\|_{v,p} = \mathrm{STV}_p g(\cdot)$.

If $r = 1$, it is well-known that $\mathcal{S}_q^*$ and $\mathcal{G}_p$ are isometrically isomorphic. Namely, to each function $g(\cdot) \in \mathcal{G}_p$ there corresponds a functional $l \in \mathcal{S}_q^*$ defined by

$$l(y(\cdot)) = \int_0^1 y(t)\, dg(t), \tag{12}$$

with $\|l\| = \|g(\cdot)\|_{v,p}$; conversely, if l is any functional in $\mathcal{S}_q^*$, then there is a unique function $g(\cdot) \in \mathcal{G}_p$ of equal norm such that l is defined by (12). If $r > 1$, the above statements are still valid, except that (12) must be replaced by

$$l(y(\cdot)) = \int_0^1 \sum_{j=1}^{r} y_j(t)\, dg_j(t) = \int_0^1 y(t) \cdot dg(t)$$

(the scalar-valued functions $g_j(\cdot)$ and $y_j(\cdot)$ represent the components in E_r of $g(\cdot)$ and $y(\cdot)$). The proof of this representation in the case $r > 1$ is straightforward and based on the case of $r = 1$, and is omitted here.

We note that there is an isometric isomorphism between $\mathfrak{F}_p$ and a linear manifold in $\mathcal{G}_p$ (or $\mathcal{S}_q^*$). Namely, to each $u(\cdot)$ in $\mathfrak{F}_p$, there corresponds a function $g(\cdot)$ in $\mathcal{G}_p$ defined by

$$(13) \qquad g(t) = \int_0^t u(s)\, ds, \qquad 0 \leqq t \leqq 1,$$

and a functional $l \in \mathcal{S}_q^*$ defined by

$$(14) \qquad l(y(\cdot)) = \int_0^1 \sum_{j=1}^r y_j(t)u_j(t)\, dt = \int_0^1 y(t)\cdot u(t)\, dt.$$

It follows directly from (14) that $\| l \| = \| u(\cdot)\|_p$. In addition, relation (13) implies that for every $y(\cdot) \in \mathcal{S}_q$,

$$(15) \qquad \int_0^1 y(t)\cdot dg(t) = \int_0^1 y(t)\cdot u(t)\, dt = l(y(\cdot)),$$

so that the isomorphism defined by (14) is the product of the isomorphism between $\mathcal{G}_p$ and $\mathcal{S}_q^*$, and the isomorphism defined by (13). Relation (15) implies that $\| l \| = \| g(\cdot)\|_{v,p} = \| u(\cdot)\|_p$, so that the isomorphisms defined by (13) and (14) are indeed isometric.

4. Solution of the optimization problem. In this section we shall apply Theorems 1 and 2 to the spaces described in §3, with the aim of obtaining a solution to our given problem.

In this and the next section we denote by $y^i(t)$ the ith row vector $(y_1^i(t), \cdots, y_r^i(t))$ of the matrix $Y(t)$ in (1).

Our original problem may now be reformulated as follows.

Given n elements $y^1(\cdot), \cdots, y^n(\cdot)$ in $\mathcal{S}_q$, and a nonzero vector $c = (c_1, \cdots, c_n)$ (the problem is trivial if $c = 0$) in E_n, find a function $u(\cdot) \in \mathfrak{F}_p$ of least norm that satisfies the equations

$$(16) \quad c_i = \int_0^1 \sum_{j=1}^r y_j^i(t)u_j(t)\, dt = \int_0^1 y^i(t)\cdot u(t)\, dt, \qquad i = 1, \cdots, n.$$

In general, this problem will have no solution. It is necessary to embed $\mathfrak{F}_p$ in $\mathcal{G}_p$ (as described in §3) and rephrase the problem as follows.

Find a function $g(\cdot) \in \mathcal{G}_p$ of minimum norm which satisfies the equations

$$(17) \qquad c_i = \int_0^1 y^i(t)\cdot dg(t), \qquad i = 1, \cdots, n.$$

Because of the isomorphism between $\mathcal{G}_p$ and $\mathcal{S}_q^*$, we may state the problem in yet a different way.

Find a functional $l \in \mathcal{S}_q^*$ of minimum norm which satisfies the relations

$$l(y^i) = c_i, \qquad i = 1, \cdots, n. \tag{18}$$

The last problem statement is precisely the one to which Theorem 1 applies. We shall show that Theorems 1 and 2 imply that there is a minimum-norm functional l^* which satisfies (18) whose corresponding function $g^*(\cdot) \in \mathcal{G}_p$ is a step function with at most n points of discontinuity. As we shall see, a step function $g^*(\cdot)$ in (17) may loosely be thought of as corresponding to a function $u^*(\cdot)$ in (16) which is a linear combination of delta (or impulse) functions (with impulses at those values of t where $g^*(\cdot)$ is discontinuous), with $\| u^*(\cdot)\|_p = \| g^*(\cdot)\|_{v,p} = \| l^* \|$.

We shall henceforth assume that the vector functions $y^i(\cdot)$ are linearly independent. By virtue of Theorem 1 and the representation of $\mathcal{S}_q^*$ discussed in §3, a minimum-norm solution $g^*(\cdot) \in \mathcal{G}_p$ of (17) does exist, and can be obtained by first finding a vector $\bar{\eta}$ that satisfies the relation

$$\bar{\eta}\cdot c = \sup_{\eta \in H} \eta\cdot c, \tag{19}$$

where

$$H = \left\{\eta: \quad \eta \in E_n, \max_{0 \leqq t \leqq 1} \left| \sum_{i=1}^{n} \eta_i y^i(t) \right|_q = 1\right\}, \tag{20}$$

and then finding any solution $g^*(\cdot) \in \mathcal{G}_p$ that satisfies (17) as well as the condition (corresponding to (5))

$$\int_0^1 \bar{y}(t) \cdot dg^*(t) = \left[\max_{0 \leqq t \leqq 1} | \bar{y}(t) |_q\right] \mathrm{STV}_p g^*(\cdot), \tag{21}$$

where

$$\bar{y}(t) = \sum_{i=1}^{n} \bar{\eta}_i y^i(t), \qquad \bar{\eta} = (\bar{\eta}_1, \cdots, \bar{\eta}_n). \tag{22}$$

Define the sets Γ_j ($j = 1, \cdots, r$) and Γ (which are closed subsets of [0, 1]), for a fixed solution $\bar{\eta}$ of (19) and the corresponding function $\bar{y}(\cdot)$ defined by (22), as follows.

$$\begin{aligned} \Gamma &= \{t: \quad | \bar{y}(t)|_q = \max_{0 \leqq \tau \leqq 1} | \bar{y}(\tau)|_q, \quad 0 \leqq t \leqq 1\}, \\ \Gamma_j &= \{t: \quad | \bar{y}_j(t)| = \max_{0 \leqq \tau \leqq 1} \max_{1 \leqq k \leqq r} | \bar{y}_k(\tau)|, \quad 0 \leqq t \leqq 1\}. \end{aligned} \tag{23}$$

(Since $\bar{\eta} \in H$, $\max_\tau | \bar{y}(\tau)|_q = 1$, and if $p = 1$, $\max_{\tau,k} | \bar{y}_k(\tau)| = 1$.) Note that $\Gamma = \bigcup_{j=1}^r \Gamma_j$ if $p = 1$.

We now make use of (21) to characterize the minimum-norm solutions of (17). It is convenient to treat the cases $p = 1$ and $p > 1$ separately. We first consider the case $p > 1$.

THEOREM 3. *Let $\bar{\eta}$ be any solution of* (19), (20), *and let $\bar{y}(\cdot)$ and Γ be correspondingly defined through* (22) *and* (23).

Then, if $g^*(\cdot) \in \mathcal{G}_p$ (where $p > 1$ is the conjugate index of q in (20)) *is any minimum-norm solution of* (17), *$g^*(\cdot)$ is constant in each open subinterval of* [0, 1] *which does not meet Γ. Also, the points of discontinuity of $g^*(\cdot)$ are all contained in Γ. If $\hat{t}$ is a point of discontinuity of $g^*(\cdot)$, there is a positive number $\alpha_{\hat{t}}$ such that the "jumps" in the components $g_j^*(\cdot)$ of $g^*(\cdot)$ are given by*

$$(24)\quad \alpha_{\hat{t}} > 0, \qquad g_j^*(\hat{t}) - g_j^*(\hat{t}^-) = \begin{cases} \alpha_{\hat{t}}\,|\,\bar{y}_j(\hat{t})\,|^{q-1}\operatorname{sgn}\bar{y}_j(\hat{t}), \\ \quad \text{if } 1 < p < \infty; \\ \alpha_{\hat{t}}\operatorname{sgn}\bar{y}_j(\hat{t}), \quad \text{if } p = \infty \\ \quad \text{and } \bar{y}_j(\hat{t}) \neq 0; \\ \text{any value in } [-\alpha_{\hat{t}}, \alpha_{\hat{t}}], \quad \text{if} \\ \quad p = \infty \quad \text{and} \quad \bar{y}_j(\hat{t}) = 0 \end{cases}$$

(*if $\hat{t} = 0$, the left-hand side of* (24) *should be replaced by $g_j^*(0^+) - g_j^*(0)$*). *In particular, if Γ is made up of a finite number of points, then $g^*(\cdot)$ is a step function (whose points of discontinuity belong to Γ and whose jumps are given by* (24)).

Conversely if $g^*(\cdot)$ is any step function in $\mathcal{G}_p$ whose points of discontinuity all belong to Γ, whose jumps are given by (24), *and which satisfies* (17), *then $g^*(\cdot)$ is a minimum-norm solution of* (17).

Proof. We first prove that if $g^*(\cdot)$ is a minimum-norm solution of (17), and $[t', t'']$ is a closed subinterval of [0, 1] that does not meet Γ, then $g^*(\cdot)$ is constant for $t' \leqq t \leqq t''$. By Theorem 1 Part B, $g^*(\cdot)$ satisfies (21). It is easily seen that

$$(25)\quad \begin{aligned} \int_0^1 \bar{y}(t)\cdot dg^*(t) \leqq{} & \underset{0 \leqq t \leqq t'}{\mathrm{STV}_p}\, g^*(\cdot)\Big[\max_{0 \leqq t \leqq t'} |\,\bar{y}(t)|_q\Big] \\ & + \underset{t' \leqq t \leqq t''}{\mathrm{STV}_p}\, g^*(\cdot)\Big[\max_{t' \leqq t \leqq t''} |\,\bar{y}(t)|_q\Big] + \underset{t'' \leqq t \leqq 1}{\mathrm{STV}_p}\, g^*(\cdot)\Big[\max_{t'' \leqq t \leqq 1} |\,\bar{y}(t)|_q\Big]. \end{aligned}$$

Since $[t', t'']$ does not meet Γ,

$$(26)\qquad \max_{t' \leqq t \leqq t''} |\,\bar{y}(t)|_q < \max_{0 \leqq t \leqq 1} |\,\bar{y}(t)|_q\,.$$

Also,

$$(27)\quad \underset{0 \leqq t \leqq 1}{\mathrm{STV}_p}\, g^*(\cdot) = \underset{0 \leqq t \leqq t'}{\mathrm{STV}_p}\, g^*(\cdot) + \underset{t' \leqq t \leqq t''}{\mathrm{STV}_p}\, g^*(\cdot) + \underset{t'' \leqq t \leqq 1}{\mathrm{STV}_p}\, g^*(\cdot).$$

Relations (21), (25), (26), and (27) can hold simultaneously only if

$$\underset{t' \leqq t \leqq t''}{\mathrm{STV}_p}\, g^*(\cdot) = 0,$$

i.e., only if $g^*(\cdot)$ is constant in $[t', t'']$. This implies that $g^*(\cdot)$ is constant in any open subinterval of $[0, 1]$ which does not intersect Γ. In particular, if $g^*(\cdot)$ has a point of discontinuity in $(0, 1)$, this point must belong to Γ. An analogous argument shows that the jumps of $g^*(\cdot)$ at a point of discontinuity are given by (24), and that $g^*(\cdot)$ is discontinuous at $t = 0$ or at $t = 1$ only if these points belong to Γ.

If $g^*(\cdot) \in \mathcal{G}_p$ is a step function whose points of discontinuity belong to Γ and whose jumps are given by (24), it follows by direct substitution that $g^*(\cdot)$ satisfies (21). If, in addition, $g^*(\cdot)$ satisfies (17), then according to Theorem 1 Part B, $g^*(\cdot)$ is a minimum-norm solution of these equations.

We now employ Theorem 2 to show that there is always a minimum-norm solution of (17) which is a step function *with at most n points of discontinuity* (all of which belong to Γ, with the jumps satisfying (24)).

Define the subset of G of $\mathcal{G}_p$ as follows: $g(\cdot) \in G$ if and only if $g(\cdot)$ is a step function with a single point of discontinuity in $[0, 1]$, and $\| g(\cdot)\|_{v,p} = 1$; i.e., G is made up of functions of the form

$$g(t) = \begin{cases} 0, & 0 \leqq t < \bar{t}, \\ \xi, & \bar{t} \leqq t \leqq 1, \end{cases} \tag{28}$$

where $\bar{t}$ is some point in $[0, 1]$ and $| \xi |_p = 1$. If $\bar{t} = 0$, an obvious modification must be made in (28). Let D denote the set of functionals in $\mathcal{S}_q^*$ that correspond to elements of G.

THEOREM 4. *The set D defined above satisfies conditions* (a), (b), *and* (c) *of Theorem* 2, *with $\mathcal{S}_q$ taken for $\mathcal{B}$.*

Proof. Condition (a) follows from the fact that $\| l \| = \mathrm{STV}_p\, g(\cdot) = | \xi |_p = 1$ when $l \in D$ corresponds to a function $g(\cdot)$ of the form (28). Condition (c) is an immediate consequence of the sequential compactness of the interval $[0, 1]$ and of the unit "sphere" $\{\xi\colon \ \xi \in E_r, | \xi |_p = 1\}$ in E_r.

Thus, it only remains to prove that the convex hull K of D is dense in the unit ball S of $\mathcal{S}_q^*$ with respect to the weak* topology of $\mathcal{S}_q^*$.

Let $\bar{l} \in S$, and suppose that $\bar{l}$ is represented by $\bar{l}(y) = \int y(t)\cdot d\bar{g}(t)$, where $\bar{g}(\cdot) \in \mathcal{G}_p$ and $\mathrm{STV}_p\, \bar{g}(\cdot) = \| \bar{l} \| \leqq 1$. Fix a weak* neighborhood N of $\bar{l}$ defined by elements $z^1, \cdots, z^m$ in $\mathcal{S}_q$ and a positive number ϵ, thus $N = \{l\colon \ | \bar{l}(z^i) - l(z^i)| < \epsilon, i = 1, \cdots, m\}$.

Choose a partition $0 = t_0 < t_1 < \cdots < t_\nu = 1$ of $[0, 1]$ such that, for each $i = 1, \cdots, m$,

$$\left| \bar{l}(z^i) - \sum_{j=0}^{\nu-1} z^i(t_j)\cdot[\bar{g}(t_{j+1}) - \bar{g}(t_j)] \right| < \epsilon. \tag{29}$$

Note that, by definition of $\mathrm{STV}_p\,\bar{g}(\cdot)$,

$$\sum_{j=0}^{\nu-1} |\,\bar{g}(t_{j+1}) - \bar{g}(t_j)|_p \leqq \mathrm{STV}_p\,\bar{g}(\cdot) \leqq 1. \tag{30}$$

Define elements l_j, $j = 0, \cdots, \nu - 1$, in D by setting

$$l_j(y(\cdot)) = y(t_j)\cdot\left[\frac{\bar{g}(t_{j+1}) - \bar{g}(t_j)}{|\,\bar{g}(t_{j+1}) - \bar{g}(t_j)\,|_p}\right], \tag{31}$$

(if the denominator in (31) vanishes, set $l_j = 0$), and let

$$\hat{l} = \sum_{j=0}^{\nu-1} |\,\bar{g}(t_{j+1}) - \bar{g}(t_j)|_p\, l_j\,. \tag{32}$$

It follows from (29), (31) and (32) that $|\,\bar{l}(z^i) - \hat{l}(z^i)| < \epsilon$ for each i; i.e., $\hat{l} \in N$. Since each nonzero $l_j \in D$ and since the origin of $\mathcal{S}_q^*$ belongs to K, it follows from (30) and (32) that $\hat{l} \in K$. Because of the arbitrariness in the choice of $\bar{l}$ and N, this implies that K is dense in S, completing the proof of Theorem 4.

COROLLARY. *There is a minimum-norm solution* $g^*(\cdot) \in \mathcal{G}_p\,(p > 1)$ *of* (17) *which is a step function with at most n points of discontinuity, all of which belong to* Γ, *the jumps at which satisfy* (24).

Proof. According to Theorem 1, a minimum-norm solution l^0 of (18) exists. By Theorem 2, there is an element $l' = \sum_{j=1}^n \lambda_j l_j$, with each $l_j \in D$, in $\mathcal{S}_q^*$ such that $l'(y^i) = l^0(y^i) = c_i$ for each $i = 1, \cdots, n$, and such that $\|\,l'\,\| \leqq \|\,l^0\,\|$. Since l^0 is of minimum norm, $\|\,l'\,\| = \|\,l^0\,\|$; i.e., l' is a minimum-norm solution of (18).

Each l_j corresponds to a function $g^j(\cdot) \in G$ with a single point of discontinuity. Hence, l' corresponds to $g^*(\cdot) = \sum_{j=1}^n \lambda_j g^j(\cdot) \in \mathcal{G}_p$, which has at most n points of discontinuity and is a minimum-norm solution of (17). The remainder of the corollary follows from Theorem 3.

We now turn to the case of $p = 1$ $(q = \infty)$. Corresponding to Theorem 3, we have the following proposition.

THEOREM 3$'$. *Let* $\bar{\eta}$ *be any solution of* (19), (20) *with* $q = \infty$, *and let* $\bar{y}(\cdot)$ *and the sets* Γ *and* Γ_j *be correspondingly defined through* (22) *and* (23).

Then, if $g^*(\cdot) \in \mathcal{G}_1$ *is any minimum-norm solution of* (17), *the component* $g_j^*(\cdot)$ *of* $g^*(\cdot)$ *is constant in every open subinterval of* [0, 1] *which does not meet* Γ_j. *Also, the points of discontinuity of* $g_j^*(\cdot)$ *are all contained in* $\Gamma_j \subset \Gamma$. *If* $\hat{t}$ *is a point of discontinuity of* $g_j^*(\cdot)$, *there is a positive number* $\alpha_{j,\hat{t}}$ *such that the jump in* $g_j^*(\cdot)$ *is given by*

$$g_j^*(\hat{t}) - g_j^*(\hat{t}^-) = \alpha_{j,\hat{t}}\,\mathrm{sgn}\,\bar{y}_j(\hat{t}), \qquad \alpha_{j,\hat{t}} > 0 \tag{33}$$

(*if* $\hat{t} = 0$, *the left-hand side in* (33) *should be replaced by* $g_j^*(0^+) - g_j^*(0)$). *In particular, if* Γ_j *is made up of a finite number of points, then* $g_j^*(\cdot)$ *is a* *step function (whose points of discontinuity belong to* Γ_j *and whose jumps are given by* (33)), *and if* Γ_j *is empty, then* $g_j^*(t) \equiv 0$.

Conversely, if $g^*(\cdot)$ *is any step function in* $\mathcal{G}_1$ *such that the points of discontinuity of* $g_j^*(\cdot)$ *belong to* Γ_j *and the jumps are given by* (33), *and if* $g^*(\cdot)$ *also satisfies* (17), *then* $g^*(\cdot)$ *is a minimum-norm solution of* (17).

The proof is very similar to that of Theorem 3, and is therefore omitted.

The fact that a minimum-norm solution $g^*(\cdot)$ is of the above form, when Γ is made up of a finite number of points, has previously been pointed out by Krasovskii [5], the author [6], and Kreindler [8].

In analogy with the set G in the case $p > 1$, we now define the set $\bar{G}$ as follows: $g(\cdot) \in \bar{G}$ if an only if $g(\cdot)$ is of the form (28) (or an obvious modification thereof in case $\bar{t} = 0$) with ξ a unit vector all but one of whose components vanish; i.e., $\bar{G}$ consists of the functions $g(\cdot)$ whose components are given by (for some index $k = 1, \cdots, r$, some number $\bar{t} \in [0, 1]$, and $\theta = +1$ or -1)

$$g_k(t) = \begin{cases} 0, & 0 \leqq t < \bar{t}, \\ \theta, & \bar{t} \leqq t \leqq 1, \end{cases}$$

$$g_j(t) \equiv 0, \qquad \text{if } j \neq k.$$

(An obvious modification must again be made if $\bar{t} = 0$.) Let $\bar{D}$ denote the set of functionals in $\mathcal{S}_\infty^*$ that correspond to elements of $\bar{G}$.

THEOREM 4$'$. *The set* $\bar{D}$ *defined above satisfies conditions* (a), (b), *and* (c) *of Theorem* 2, *with* $\mathcal{S}_\infty$ *taken for* $\mathcal{B}$.

The proof is almost identical with that of Theorem 4, and is omitted. The following corollary to Theorem 4$'$ (which follows just as the corollary to Theorem 4) yields a representation of a minimum-norm solution of (17) when $p = 1$.

COROLLARY. *There is a minimum-norm solution* $g^*(\cdot) \in \mathcal{G}_1$ *of* (17) *which is a step function such that* $g_j^*(\cdot)$ *has* n_j *points of discontinuity (all of which belong to* Γ_j, *with the jumps satisfying* (33)), *and* $\sum_{j=1}^r n_j \leqq n$.

Because a minimum-norm solution of (17) is an element $g(\cdot) \in \mathcal{G}_p$ which is a step function, there naturally arises the following question. Given a step function $g(\cdot) \in \mathcal{G}_p$ and the corresponding functional $\hat{l} \in \mathcal{S}_q^*$ (defined by (12)), do there exist functions $u(\cdot) \in \mathcal{F}_p$ such that the corresponding functionals $l \in \mathcal{S}_q^*$ (defined by (14)) in some sense approximate the functional $\hat{l}$? We shall below answer this question in the affirmative. Indeed, for any $\epsilon > 0$, there exist functions $u'(\cdot\,; \epsilon)$ and $u''(\cdot\,; \epsilon)$ in $\mathcal{F}_p$ such that

$$|l_\epsilon'(y^i) - \hat{l}(y^i)| < \epsilon, \qquad i = 1, \cdots, n, \qquad \|l_\epsilon'\| = \|\hat{l}\|,$$

and

$$l_\epsilon''(y^i) = \hat{l}(y^i), \qquad i = 1, \cdots, n, \qquad |\,\|l_\epsilon''\| - \|\hat{l}\|\,| < \epsilon,$$

where l_ϵ' and l_ϵ'' are the functionals in $\mathcal{S}_q^*$ corresponding to $u'(\cdot; \epsilon)$ and $u''(\cdot; \epsilon)$. Thus, if $g(\cdot)$ is a minimum-norm solution of (17), it is possible to find a function $u'(\cdot) \in \mathfrak{F}_p$ whose norm in $\mathfrak{F}_p$ equals $\| g(\cdot)\|_{v,p}$, and which satisfies relations (16) with an arbitrarily small error, as well as a function $u''(\cdot) \in \mathfrak{F}_p$ that satisfies relations (16) exactly and has norm arbitrarily near $\| g(\cdot)\|_{v,p}$.

To verify the above, let $g(\cdot)$ be a step function with discontinuities at the points $t_1, \cdots, t_\mu$. Let $\Delta_j g = [g(t_j) - g(t_j^-)]$ if $t_j \neq 0$; if $t_j = 0$, let $\Delta_j g = [g(0^+) - g(0)]$. If $0 < t_j < 1$, define $I_{j,\epsilon}$ for $\epsilon > 0$ to be the closed interval $[t_j - \epsilon/2, t_j + \epsilon/2]$; if $t_j = 0$, let $I_{j,\epsilon}$ be $[0, \epsilon]$; if $t_j = 1$, let $I_{j,\epsilon} = [1 - \epsilon, 1]$. We shall always assume that ϵ is sufficiently small that $I_{j,\epsilon} \subset [0, 1]$ for each j, and that $I_{j,\epsilon} \cap I_{k,\epsilon} = \emptyset$ if $j \neq k$. Denote by $\kappa_j(t; \epsilon)$ the characteristic function of $I_{j,\epsilon}$, and let

$$u'(t; \epsilon) = \frac{1}{\epsilon} \sum_{j=1}^{\mu} \kappa_j(t; \epsilon)\Delta_j g.$$

Then it is easily seen that

$$\int_0^1 y^i(t) \cdot u'(t; \epsilon)\, dt \xrightarrow[\epsilon \to 0]{} \int_0^1 y^i(t) \cdot dg(t), \qquad i = 1, \cdots, n,$$

$$\| g(\cdot)\|_{v,p} = \| u'(\cdot; \epsilon)\|_p \quad \text{for all} \quad \epsilon.$$

If the $n \times \mu r$ matrix $\bar{Y} = (Y(t_1), \cdots, Y(t_\mu))$ has rank n, it readily follows that there exist functions $u''(t; \epsilon) \in \mathfrak{F}_p$ of the form

$$u''(t; \epsilon) = \frac{1}{\epsilon} \sum_{j=1}^{\mu} \kappa_j(t; \epsilon) u^{j,\epsilon},$$

where the $u^{j,\epsilon}$ are constant r-vectors such that $u^{j,\epsilon} \to \Delta_j g$ as $\epsilon \to 0$ for each j, with

$$(34) \qquad \int_0^1 y^i(t) \cdot u''(t; \epsilon)\, dt = \int_0^1 y^i(t) \cdot dg(t), \qquad i = 1, \cdots, n,$$

and such that

$$\| u''(\cdot; \epsilon)\|_p \xrightarrow[\epsilon \to 0]{} \| g(\cdot)\|_{v,p}.$$

Thus, the functions u' and u'' have the desired properties. Loosely speaking, one may say that a step function $g(\cdot)$ in (17) corresponds to a function $u(\cdot)$ in (16) which is a linear combination of "delta-functions."

If the matrix $\bar{Y}$ has rank less than n, it is still possible to construct functions $u''(t; \epsilon)$ satisfying (34) and vanishing outside $\bigcup_{j=1}^{\mu} I_{j,\epsilon}$. This follows from the fact that if we define

$$S(j, \epsilon) = \left\{ \int_{I_{j,\epsilon}} Y(t)u(t)\, dt \colon \ \| u(\cdot) \|_p \leqq | \Delta_j g |_p \right\},$$
$$j = 1, \cdots, \mu; \epsilon > 0,$$

then $S(j, \epsilon)$ is a convex set in E_n containing the origin as a relative interior point, and $Y(t_j)\Delta_j g$ belongs to the closure of $S(j, \epsilon)$. In fact, the stated properties of $S(j, \epsilon)$ imply that there exist functions $u^j(t; \epsilon) \in \mathfrak{F}_p$, vanishing outside $I_{j,\epsilon}$, such that for each j,

$$\int_{I_{j,\epsilon}} Y(t)u^j(t; \epsilon) = \frac{1}{1+\epsilon}\int_{I_{j,\epsilon}} Y(t)\, dg(t) = \frac{1}{1+\epsilon} Y(t_j)\Delta_j g,$$

$$\| u^j(\cdot\,; \epsilon) \| \leqq | \Delta_j g |_p .$$

If we define $u''(t; \epsilon) = (1 + \epsilon)\sum_{j=1}^{\mu} u^j(t; \epsilon)$, then the functions $u''(t; \epsilon)$ satisfy (34), and $\| u''(\cdot\,; \epsilon)\|_p \leqq (1 + \epsilon)\| g(\cdot)\|_{v,p}$. In particular, if $g(\cdot)$ is a minimum-norm solution of (17), so that the $u''(\cdot\,; \epsilon)$ satisfy (16), then $\| u''(\cdot\,; \epsilon)\|_p \geqq \| g(\cdot)\|_{v,p}$, and $\| u''(\cdot\,; \epsilon)\|_p \rightarrow \| g(\cdot)\|_{v,p}$ as $\epsilon \rightarrow 0$.

The above remarks have significance for the following problem, which is a variant of our original problem.

Let Γ_α, where $0 \leqq \alpha < \infty$, be the set of elements $u(\cdot) \in \mathfrak{F}_p$ satisfying the condition $| u(t)|_p \leqq \alpha$ for all t, $0 \leqq t \leqq 1$; then find a function $u(\cdot) \in \Gamma_\alpha$ of least norm that satisfies (16).

By what was said above, there is at least one solution in Γ_α to (16) if α is sufficiently large. Then, according to results in [16], a minimum-norm solution exists in Γ_α (so that it is unnecessary to embed $\mathfrak{F}_p$ in $\mathcal{G}_p$).

Let M_∞ denote the minimum of the norms of the solutions of (17) in $\mathcal{G}_p$, and let M_α denote the minimum of the norms of the solutions of (16) that belong to Γ_α. It follows from the previous discussion that $M_\alpha \rightarrow M_\infty$ as $\alpha \rightarrow \infty$.

Under certain conditions (which will not be discussed here), if a minimum-norm solution of (17) is a step function, then the minimum-norm solutions of (16) in Γ_α approach the linear combination of delta functions corresponding to the step function. To be precise, if the step function has discontinuities at the points $t_1, \cdots, t_\mu$ (we denote the value of the jump at t_j, as before, by $\Delta_j g$), and if $u^\alpha(\cdot)$ is a minimum-norm solution of (16) in Γ_α, then, for sufficiently large α,

$$u^\alpha(t) = \sum_{j=1}^{\mu} \kappa_{\alpha,j}(t)\, \frac{u^{\alpha,j}(t)}{\delta_{\alpha,j}},$$

where $\kappa_{\alpha,j}(t)$ is the characteristic function of an interval $I_{\alpha,j}$ contained in [0, 1], containing the point t_j, and of length $\delta_{\alpha,j} > 0$. In addition, $\delta_{\alpha,j} \rightarrow 0$ as $\alpha \rightarrow 0$; $| u^{\alpha,j}(t)|_p = 1$ for all α, j, and $t \in I_{\alpha,j}$; and

$$\sup_{t \in I_{\alpha,j}} | u^{\alpha,j}(t) - \Delta_j g |_p \underset{\alpha\rightarrow\infty}{\longrightarrow} 0.$$

5. Computing the optimum solution. In this section we shall discuss a possible computational method for obtaining a solution $\bar{\eta}$ of (19), (20), and, having found an $\bar{\eta}$ and determined the corresponding sets Γ, Γ_j and the function $\bar{y}(\cdot)$, for finding a minimum-norm solution of (17) of the type described in the corollary to Theorem 4 (or 4′).

We first consider the case $p > 1$.

The problem of finding the maximum in (19), (20) is clearly equivalent to the following problem in nonlinear programming. Maximize the linear function $g(\eta) = g(\eta_1, \cdots, \eta_n) = \sum_{i=1}^{n} c_i\eta_i$ subject to the constraints (defined for every t, $0 \leqq t \leqq 1$):

$$(35) \qquad \rho(\eta; t) = \sum_{j=1}^{r} \left| \sum_{i=1}^{n} y_j^{\,i}(t)\eta_i \right|^q \leqq 1.$$

This programming problem, being given in terms of a linear objective function and a continuum of convex constraints, is in itself convex, and the absence of false local maxima is thus guaranteed. Note that the dimension of the programming problem is only n. A number of computational methods, e.g., Rosen's gradient projection method [12], exist for such problems. It is clear, however, that the continuum of constraints (35) must be approximated by a suitably large, finite subcollection, if the problem is to be solved on a digital computer.

In the nonsingular case, a solution of the above problem is given by a vector $\bar{\eta}$ which lies on the intersection of n (or fewer) surfaces $\rho(\eta; t) = 1$ which are in general position at $\bar{\eta}$; i.e., $\bar{\eta}$ maximizes $g(\eta)$ subject to the constraints (35), $\rho(\bar{\eta}; t) = 1$ for n (or fewer) values of t in [0, 1]—which, by definition, make up Γ—and $\rho(\bar{\eta}; t) < 1$ away from these values of t.

Once a solution $\bar{\eta}$ to the programming problem has been obtained, and the set Γ and the function $\bar{y}(\cdot)$ defined accordingly (by (22) and (23)), it is only necessary to find values $t_i \in \Gamma$ and positive numbers $\alpha_{t_i} = \alpha_i$ $(i = 1, \cdots, \mu; \mu \leqq n)$ such that the corresponding step function in $\mathcal{G}_p$ (i.e., with discontinuities at the points t_i and the jumps given by (24)) is a (necessarily minimum-norm) solution of (17). If $p = \infty$, additional numbers must be determined if $\bar{y}_j(t_i) = 0$ for some i and j. We shall suppose that this case does not arise in the argument that follows.

Let us first consider the nonsingular case where Γ consists of precisely n points $t_1, \cdots, t_n$, and the n surfaces whose equations are $\rho(\eta; t_j) = 1$ $(j = 1, \cdots, n)$ are in general position at their point of intersection $\bar{\eta}$. By Theorem 3, every minimum-norm solution $g^*(\cdot) \in \mathcal{G}_p$ of (17) is a step function whose points of discontinuity are included among the t_j, and whose jumps are given by (24). If $g^*(\cdot)$ is of this form, then

$$(36) \qquad \int_0^1 y^i(t) \cdot dg^*(t) = \sum_{j=1}^{n} \beta_i(t_j)\alpha_j = c_i, \qquad i = 1, \cdots, n,$$

where

$$\beta_i(t_j) = \sum_{s=1}^{r} y_s^{\,i}(t_j) |\bar{y}_s(t_j)|^{q-1} \operatorname{sgn} \bar{y}_s(t_j).$$

The quantities α_j are the only unknowns in (36), but they are uniquely determined by these equations if the matrix $(\beta_i(t_j))_{i,j}$ is nonsingular. However, the nonsingularity of this matrix is precisely the condition that the surfaces $\rho(\eta;\, t_i) = 1$ be in general position at $\bar{\eta}$. In this case, therefore, the minimum-norm solution of (17) is unique. In summary, it can be obtained as follows.

1. Solve the nonlinear programming problem of maximizing $g(\eta) = \eta \cdot c$ subject to the constraints (35).

2. Let $\bar{\eta}$ be a solution of this problem, let $t_1, \cdots, t_n$ be those values of t for which $\rho(\bar{\eta};\, t) = 1$, and let $\bar{y}(t)$ be defined by (22).

3. Solve (36) for the constants α_j $(j = 1, \cdots, n)$.

4. The minimum-norm solution $g^*(\cdot)$ of (17) is a step function whose points of discontinuity are $t_1, \cdots, t_n$, and whose jumps are given by (24) with $\alpha_{t_j} = \alpha_j$ (if $\alpha_k = 0$, $g^*(\cdot)$ is continuous at t_k).

5. The optimum solution can be approximated by an element $u(\cdot) \in \mathfrak{F}_p$ as discussed in §4.

If Γ consists of μ points $t_1, \cdots, t_\mu$, where $\mu < n$, and the corresponding surfaces given by $\rho(\eta;\, t_i) = 1$ $(i = 1, \cdots, \mu)$ are in general position at $\bar{\eta}$, equations (36) take the form

$$\sum_{j=1}^{\mu} \beta_i(t_j)\alpha_j = c_i\,, \qquad i = 1, \cdots, n. \tag{37}$$

Although this system is overdetermined $(n > \mu)$, our existence theorem guarantees that it is consistent, i.e., does have a solution for numbers α_j, and the general position condition implies that this solution is unique, which in turn means that the minimum-norm solution of (17) is unique in this case also.

The numbers α_j are analogous to the Lagrange multipliers that arise in an ordinary maximization problem in the presence of constraints. Indeed, they are precisely the multipliers for the problem of maximizing $g(\eta)$ subject to the constraints $\rho(\eta;\, t_i) \leqq 1$, $i = 1, \cdots, \mu \leqq n$.

If the surfaces $\rho(\eta;\, t) = 1$ for $t \in \Gamma$ are not in general position at $\bar{\eta}$, as must occur when Γ consists of more than n points, it is necessary to pick out some μ $(\mu \leqq n)$ values $t_j \in \Gamma$ such that the corresponding equations (37) have a solution for numbers α_j with each $\alpha_j > 0$. Such values always exist by virtue of the above-proved existence theorems.

We now turn to the case $p = 1$. The problem of finding a maximum in (19), (20) is equivalent to the problem of finding the maximum of the linear function $g(\eta)$ subject to the linear constraints

$$-1 \leqq \sum_{i=1}^{n} \eta_i y_j^i(t) \leqq 1; \qquad j = 1, \cdots, r; \qquad 0 \leqq t \leqq 1.$$

This is now a linear programming problem of dimension n. To obtain an approximate solution on a digital computer, it is again necessary to replace the continuum of constraints by a suitably large finite number chosen from them.

Once a solution $\bar{\eta}$ of this linear programming problem has been found, and the sets Γ_j and the function $\bar{y}(\cdot)$ defined accordingly, it is only necessary to find values $t_{ij} \in \Gamma_j$, and positive numbers α_{j,t_i} $(j = 1, \cdots, r;$ $i = 1, \cdots, n_j; \sum_j n_j \leqq n)$ such that the corresponding step function $g(\cdot)$ in $\mathfrak{G}_1$ (i.e., with the discontinuities of $g_j(\cdot)$ at the points t_{ij} and the jumps given by (33)) is a (necessarily minimum-norm) solution of (17). The method of finding these numbers differs only in detail from that for the case $p > 1$, and will not be presented here. This case has been discussed in some detail by Kreindler [8].

6. Applications to optimization. In this section we shall describe how the results derived in the preceding sections may be applied to optimization problems, and, in particular, to obtaining minimum-fuel space maneuvers.

We consider physical "systems" whose behavior can be described by a system of ordinary differential equations of the form

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t). \tag{38}$$

In (38), x is an m-vector whose coordinates describe the "state" of the system at any instant of time t, $A(t)$ is an $m \times m$ matrix function, $B(t)$ is an $m \times r$ matrix function, and $f(t)$ is an m-vector function; A, B, and f are assumed to be continuous known functions of t. The quantity $u(t)$ is the "control," a measurable function whose range is contained in E_r, which is constrained to be a member of a given normed linear function space $\mathfrak{F}$.

Certain problems in the theory of optimal control can be stated as follows. Given two distinct values t_0 and t_1 of t, an $n \times m$ matrix N, an m-vector x^0, and an n-vector x^1 (where $n \leqq m$); find a function $u^*(\cdot) \in \mathfrak{F}$ of minimum norm such that the solution $x(t)$ of (38) (by a *solution* we mean an absolutely continuous vector function that satisfies the equation almost everywhere) with $x(t_0) = x^0$ and $u(t) = u^*(t)$ satisfies the boundary condition $Nx(t_1) = x^1$. Without loss of generality, we may assume that $t_0 = 0$ and $t_1 = 1$.

By virtue of the variation of parameters formula for solutions of (38), the preceding problem can be restated as follows. Find a function $u^*(\cdot) \in \mathfrak{F}$ of minimum norm such that

$$\int_0^1 Y(t)u^*(t)\,dt = c. \tag{39}$$

In (39), $Y(t) = NX(1)X^{-1}(t)B(t)$ is a known continuous $n \times r$ matrix function, $X(t)$ being the $m \times m$ matrix solution of the equation

$$\dot{X} = AX, \qquad X(0) = I, \quad \text{the identity},$$

and $c = x^1 - NX(1)[x^0 + \int_0^1 X^{-1}(t)f(t)\,dt]$ is a known n-vector. In this form, this is precisely the problem described in the Introduction, and if the norm in $\mathfrak{F}$ is given by (11), we have the problem discussed in §4 and §5. The theory developed therein can be applied if $c \neq 0$ (the problem has the trivial solution $u^*(t) \equiv 0$ if $c = 0$), and provided that the vector functions $y^1(\cdot), \cdots, y^n(\cdot)$ constituting the rows of Y are linearly independent. A necessary condition for the linear independence of $y^1(\cdot), \cdots, y^n(\cdot)$ is that N have rank n; the latter together with the condition that the given system is "proper" [13, p. 12] is sufficient for linear independence of the $y^i(\cdot)$.

Let us now consider a specific type of physical system that is of particular current interest.

The equations of motion of a space vehicle subject only to gravitational and propulsive forces can be given in the form

$$\text{(40)} \qquad \ddot{r} = G(r, t) + \frac{T}{M},$$

where r is the radius-vector to the vehicle's center of gravity from the origin of some inertial coordinate system, G is the vector representing the gravitational acceleration, T represents the force vector due to the vehicle engine thrust, and M is the vehicle mass. If the thrust is due to a single rocket engine, the rate of change of mass due to thrusting is given by

$$\text{(41)} \qquad -\dot{M}(t) = \frac{|\,T(t)\,|_2}{g\,I_{\text{sp}}},$$

where g is the acceleration due to gravity at the earth's surface (a known constant) and I_{sp} is the so-called specific impulse, which we shall assume to be a known function of time.

The following problem naturally arises in the control of such vehicles. For given initial position $r(t_0)$, velocity $\dot{r}(t_0)$, and mass $M(t_0)$, find a thrust program (T as a function of t) which will achieve prescribed terminal values for (some, or possibly all of) the components of r and $\dot{r}$ (or for given functions of the components). The terminal time may be fixed or free. The optimal problem consists in finding that thrust program that results in a minimum loss of mass, or expenditure of fuel. This general optimization problem is as yet unsolved, although numerous particular cases have been treated in the engineering literature (see, for example, [14]).

For those cases where (40) can be put in the form (38), the methods and results developed in this paper can be applied. This can be done when (40) represents the motion of a vehicle near a "nominal" known free-fall trajectory, the radius-vector along which satisfies the equations $\ddot{R} = G(R, t)$. Namely, set $\delta r = r - R$, in which case $\delta\ddot{r} = G(R + \delta r, t) - G(R, t)$

$+ T/M$. Assuming that $[G(R + \delta r, t) - G(R, t)]$ can be approximated by first-order terms in δr, we obtain

$$\delta \ddot{r} = \frac{\partial G}{\partial r}(R(t), t)\, \delta r + \frac{T}{M}, \tag{42}$$

where $\partial G(R(t), t)/\partial r$ is a known matrix function of the time t. Finally, let x be the 6-vector whose first three coordinates coincide with those of δr and whose last three coincide with those of $\delta \dot{r}$. Then (42) can be put in the form

$$\dot{x} = A(t)x(t) + B(t)u, \tag{43}$$

where $u = T/(I_{\text{SP}} M)$. Suppose that initial "perturbations" from the nominal $\delta r(t_0)$ and $\delta \dot{r}(t_0)$, a terminal time t_1, an initial mass $M(t_0)$, and desired terminal values $\delta r(t_1)$, $\delta \dot{r}(t_1)$ (or certain linear combinations, less than 6 in number of them) are given. Since a minimization of the loss of mass is equivalent to a maximization of $M(t_1)$, or a minimization of

$$\int_{t_0}^{t_1} | u(t) |_2 \, dt,$$

because, by (41) and the definition of u,

$$\int_{t_0}^{t_1} | u(t) |_2 \, dt = \int_{t_0}^{t_1} \frac{| T(t) |_2}{M(t) I_{\text{SP}}} \, dt = g \ln \left[\frac{M(t_0)}{M(t_1)} \right],$$

our problem is now of the type described at the beginning of this section, with $p = 2$. The computational method described in §5 can be applied to determine an optimal thrusting program. Note that if it is only of interest to determine the minimum fuel expenditure, it is sufficient to solve the nonlinear programming problem, since the number λ in (3) here corresponds to the minimum of the values for $\int | u |_2 \, dt$.

The corollary to Theorem 4 has a particularly interesting interpretation in the present problem. Namely, if the number of coordinates of δr and $\delta \dot{r}$ (or linear combinations of them) whose end values are prescribed is n ($n \leqq 6$), there is a minimum-fuel thrust program which consists of n, or fewer, impulses. If a rendezvous with another vehicle is the desired terminal state, all the coordinates of δr and $\delta \dot{r}$ are specified at the terminal time, and $n = 6$. Note that if the entire motion of the space vehicle takes place in a plane, (43) can be put in the form of a fourth-order system, and a minimum-fuel rendezvous can be accomplished with four or fewer impulsive corrections.

Acknowledgment. The author is indebted to Dr. J. S. Meditch for his aid in formulating the minimum-fuel space vehicle problem. Some related work in this problem has previously been described in [15].

REFERENCES

[1] H. Hahn, *Über lineare Gleichungssysteme in linearen Räumen*, J. Reine Angew. Math., 157 (1927), pp. 214–229.

[2] N. Dunford and J. T. Schwartz, *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.

[3] N. I. Ahiezer and M. Krein, *Some Questions in the Theory of Moments*, Nauch. -Tekh. Izd. Ukr., Kharkov, 1938; English translation published by Amer. Math. Soc., Providence, Rhode Island, 1962.

[4] N. N. Krasovskii, *On the theory of optimum regulation*, Avtomat. i Telemeh., 18 (1957), pp. 960–970; English translation in Automation and Remote Control, 18 (1957), pp. 1005–1016.

[5] ———, *On the theory of optimum control*, Prikl. Mat. Meh., 23 (1959), pp. 625–639; English translation in J. Appl. Math. Mech., 23 (1959), pp. 899–919.

[6] L. W. Neustadt, *Minimum effort control systems*, this Journal, 1 (1962), pp. 16–31.

[7] W. T. Reid, *Ordinary linear differential operators of minimum norm*, Duke Math. J., 29 (1962), pp. 591–606.

[8] E. Kreindler, *Contributions to the theory of time-optimal control*, J. Franklin Inst., 275 (1963), pp. 314–344.

[9] H. A. Antosiewicz, *Linear control systems*, Arch. Rational Mech. Anal., 12 (1963), pp. 313–324.

[10] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, Amer. Math. Soc., Providence, Rhode Island, 1957.

[11] H. G. Eggleston, *Convexity*, Cambridge University Press, Cambridge, 1958.

[12] J. B. Rosen, *The gradient projection method for nonlinear programming, Part II. Nonlinear constraints*, J. Soc. Indus. Appl. Math., 9 (1961), pp. 514–532.

[13] J. P. LaSalle, *The time-optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. 5, Princeton University Press, Princeton, 1960.

[14] G. Leitman, ed., *Optimization Techniques*, Academic Press, New York, 1962.

[15] J. S. Meditch and L. W. Neustadt, *An application of optimal control to midcourse guidance*, Proceedings of Second Congress of International Federation of Automatic Control (IFAC), Butterworths, London, 1964.

[16] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

TIME-OPTIMAL CONTROL OF SOLUTIONS OF OPERATIONAL DIFFERENTIAL EQUATIONS*

H. O. FATTORINI†

Introduction. We consider the following problem: given two points u, v in the Hilbert space H, find f, $|f(t)| \leqq 1$, such that the solution of the operational differential equation $u_t = Au + f$, with initial condition $u(0) = u$, reaches v in the smallest possible time. We prove that such an f exists, utilizes the maximum energy available ($|f(t)| = 1$), and is unique. The finite-dimensional problem was studied by Bellman, Glicksberg and Gross [3] and others (see [6]); results for the infinite-dimensional case have been announced by Egoroff [7], who generalizes Pontryagin's maximum principle to a class of equations in Banach space.

The author wishes to acknowledge his indebtedness to Professor P. D. Lax for assistance received during the preparation of this work.

1. Existence of optimal controls. We shall use the notations

(i) $s, t, t', \cdots$ for positive real numbers, $c, d, e, \cdots$ for (Lebesgue measurable) subsets of the real line, $|c|, |d|, \cdots$ for their measure;

(ii) $H = \{u, v, w, \cdots\}$ for a Hilbert space with scalar product (u, v) and norm $|u|$;

(iii) $L_t = \{f, g, \cdots\}$ for the space $L^\infty((0, t); H)$ of all functions with domain $(0, t)$, range in H, strongly measurable and bounded, with norm

$$\| f \|_t = \text{ess. sup } \{|f(r)|, 0 \leqq r \leqq t\}.$$

Sometimes we shall write simply L, $\| f \|$, omitting the subindex t. We recall that L_t is a Banach space, dual of the space $L^1((0, t); H)$ of summable, H-valued functions in $(0, t)$. For further details see [1, p. 88] and [2].

Given a linear operator A in H with domain $D(A)$ and a function $f(t)$, $t \geqq 0$, with values in H, we will consider the initial-value problem

$$(1.1) \qquad u'(t) = Au(t) + f(t), \qquad t \geqq 0,$$

$$(1.2) \qquad u(0) = u.$$

A function $u(t)$, $t \geqq 0$, with values in H will be called a *strong* or *genuine* solution of (1.1), (1.2), if

(a) for each $t \geqq 0$, $u(t) \in D(A)$;

* Received by the editors November 11, 1963.

† Courant Institute of Mathematical Sciences, New York University, 4 Washington Place, New York, New York.

(b) the equality (1.1) is valid, where u' is the *strong* derivative of $u(t)$;

(c) for $t = 0$, $u(t)$ assumes the required initial value.

We assume that the homogeneous problem ($f = 0$) is *well-posed* in the sense that

(d) it has a genuine solution for any $u \in D(A)$;

(e) two solutions that agree for $t = 0$ agree for all values of t;

(f) the values of a solution at a time $t > 0$ depend *continuously* on the initial data, i.e., the operators $T(t)$ defined by

$$T(t)u(0) = u(t) \tag{1.3}$$

are *bounded*.

By (f) we can extend $T(t)$ (by closure) uniquely to the whole space H. We will denote these extensions by the same symbols. From the definition of $T(t)$ it follows that

$$T(0) = I. \tag{g}$$

By uniqueness of genuine solutions,

$$T(t)T(s) = T(t + s), \tag{h}$$

i.e., $T(\cdot)$ is a semigroup of bounded operators. It can be shown also [1, pp. 304–305] that

(i) $T(t)u$ is a strongly continuous function of t (u being a fixed element in H);

(j) $m(t) = \sup \{ | T(r) |, 0 \leqq r \leqq t\}$ is finite for all $t \geqq 0$;

(k) $T^*(t)$ is also a strongly continuous semigroup of bounded operators.

Under conditions (d), (e), (f), it is also true that

(l) if $f(t) \in D(A)$ for all $t \geqq 0$, $f(t)$ and $Af(t)$ are strongly continuous functions of t, then

$$u(t) = T(t)u + \int_0^t T(t - r)f(r)\, dr \tag{1.4}$$

is a genuine solution of (1.1), (1.2), for every $u \in D(A)$ (see [4]).

On this basis we define the expression (1.4) as a *weak solution* (or simply a solution) of (1.1), (1.2), integration being performed in the sense of Bochner [1, p. 76].

LEMMA 1.1. *Let* $\{t_n\}$, $\{u_n\} \subset H$, $\{f_n\} \subset L$ *be sequences such that*

$$\int_0^{t_n} T(t_n - r)f_n(r)\, dr = u_n .$$

Suppose further that $t_n \rightarrow t$, $u_n \rightarrow u$ (*weakly*), $\| f_n \| \leqq 1$. *Then there exists* $f \in L$, $\| f \| \leqq 1$, *such that*

$$\int_0^t T(t - r)f(r)\, dr = u.$$

Proof. Choose some upper bound s for $\{t_n\}$ and some element v in H. Define $g(r) = T^*(t - r)v$ if $0 \leqq r < t$, $g(r) = 0$ if $t \leqq r \leqq s$; define similarly g_n using t_n instead of t. It follows at once from the strong continuity of $T^*(\cdot)$ that $g_n \to g$ in $L^1((0, s); H)$. Observe next that, by Alaoglu's theorem [1, p. 37], the unit sphere of L_t is weakly compact. Then there exists a subsequence of $\{f_n\}$ (strictly speaking, a generalized subsequence), weakly convergent to some element f, $\| f \| \leqq 1$; denote again this subsequence by $\{f_n\}$. Noting that

$$(u_n , v) = \left(\int_0^{t_n} T(t_n - r)f_n(r)\, dr, v\right) = \int_0^s (f_n(r), g_n(r))\, dr$$

and taking limits, the lemma follows.

Given two elements u, v in H, we will call f in L an *admissible control* if the weak solution of (1.1), (1.2) reaches v at some time, i.e., if for some t,

$$T(t)u + \int_0^t T(t - r)f(r)\, dr = v, \tag{1.5}$$

and $\| f \| \leqq 1$.

The corresponding solution will be called an *admissible trajectory*. The smallest t for which equality (1.5) is valid will be the *transition time* corresponding to the control (or to the trajectory). The admissible control f will be called *optimal* if its transition time minimizes the transition times of all admissible controls; we will call also the corresponding trajectory an *optimal trajectory*.

To avoid confusion we shall often write (u, v)-admissible control, $\cdots$, etc., to specify which u, v we consider.

THEOREM 1.2. *Suppose that for u, v in H there exists an admissible control. Then there exists an optimal control.*

Proof. Let t be the infimum of all transition times of all admissible controls, t_n a sequence of transition times corresponding to admissible controls f_n, tending to t. Write (1.5) for these controls and apply Lemma 1.1.

2. Uniqueness of optimal controls. We define the subspace K_t as the set of all elements of H of the form

$$\int_0^t T(t - r)f(r)\, dr, \qquad f \in L. \tag{2.1}$$

The equalities

$$\begin{aligned} &\int_0^s T(s - r)f(r)\, dr \\ &\qquad = \int_{t-s}^t T(t - r)f(r - (t - s))\, dr, \qquad 0 < s < t, \end{aligned} \tag{2.2}$$

$$\int_0^t T(t-r)f(r)\,dr$$

$$(2.3)\qquad = \int_0^s T(s-r)\left(\frac{T(r)}{s}\int_0^{t-s} T(t-s-r')f(r')\,dr'\right)dr + \int_0^s T(s-r)f(r+t-s)\,dr, \qquad 0 < s < t,$$

$$(2.4)\qquad T(t)u = \int_0^t T(t-r)\,\frac{T(r)u}{t}\,dr,$$

imply that $K_s = K_t = K$ for all $s, t > 0$ and that $T(t)H \subseteq K$. If we now introduce the family of norms

$$|\, u \,|_t = \inf\left\{\|\, f \,\|_t\,;\ \int_0^t T(t-r)f(r)\,dr = u, f \in L_t\right\},$$

identities (2.1), $\cdots$, (2.4) give immediately

$$(2.5)\qquad |\, u \,|_s \geqq |\, u \,|_t\,,$$

$$(2.6)\qquad |\, u \,|_s \leqq \left(1 + \frac{t-s}{s}\,m(s)m(t-s)\right)|\, u \,|_t\,, \qquad 0 < s < t,$$

$$(2.7)\qquad |\, u \,| \leqq tm(t)|\, u \,|_t\,,$$

$$(2.8)\qquad |\, T(t)u \,|_t \leqq \frac{m(t)}{t}\,|\, u \,|.$$

Given now any measurable set e in the positive real line we define $L_t(e)$ as the (closed) subspace of L_t consisting of all functions with support in $e(0, t) = e \cap (0, t)$, and $K_t(e)$ as the subspace of K consisting of all elements of the form (2.1), with $f \in L_t(e)$. Our next result is concerned with points $t \in e$ for which $K_t(e) = K$.

LEMMA 2.1. *Let* $|\, e \,| > 0$. *Then for almost all* t *in* e, $K_t(e) = K$.

Proof. Applying (2.2) and (2.3) it is easy to see that any u in K can be written in the form

$$\int_s^t T(t-r)f(r)\,dr,$$

with $f \in L$, s arbitrarily close to t. Now take $t \in e$ and a sequence $t_n < t_{n+1} < t$, $t_n \to t$. If we write

$$\int_{t_1}^t T(t-r)f(r)\,dr = \sum_{n=1}^{\infty} T(t-t_{n+1})\int_{t_n}^{t_{n+1}} T(t_{n+1}-r)f(r)\,dr$$
$$= \sum_{n=1}^{\infty}\int_{e(t_{n+1},t_{n+2})} T(t-r)\left(\frac{T(r-t_{n+1})}{|\, e(t_{n+1}\,,\, t_{n+2}) \,|}\cdot\int_{t_n}^{t_{n+1}} T(t_{n+1}-r')f(r')\,dr'\right)dr,$$

we see that we can assert that $K_t(e) = K$ if we can construct a sequence $t_n < t_{n+1} < t$ such that $\lim t_n = t$ and

$$|\, e(t_n\,,\, t_{n+1})\,| \geqq p(t_{n+1} - t_n), \qquad p > 0, \tag{2.9}$$

$$\limsup \frac{t_n - t_{n-1}}{t_{n+1} - t_n} < \infty. \tag{2.10}$$

Now take $e_m = \left\{ t \in e\colon \left| e\left(t - \frac{1}{k}, t\right) \right| \geqq \frac{1}{2k}, k \geqq m \right\}$. By a well-known result in measure theory, $\bigcup_m e_m$ has full measure in e; moreover the set d of points on e which are density points of some e_m has the same property. It is now easy to show that any t in d has the required property. In fact, let t be a density point of e_m. Then we can select a sequence $\{t_n\}$ in e_m such that $t - \frac{1}{m} < t_n < t_{n+1} < t$, and $t = t - s_n + o(s_n)$, where s_n is any sequence tending monotonically to zero. But this implies (2.10), having chosen s_n adequately (for instance $s_n = \exp(-n)$), and (2.9) follows from the fact that $\{t_n\} \subset e_m$.

Before beginning the proof of our main result, we recall that, if f is an optimal control with transition time t, then f is optimal in any subinterval of $(0, t)$. In particular, the corresponding trajectory u will reach any of its points *only once*. The proof is a straightforward consequence of the definitions involved.

THEOREM 2.2. *Let f be a (u, v)-admissible control with transition time t, and suppose that $|\,f(r)\,| < 1$ for r in a non-null set $c \subseteq (0, t)$. Then f is not (u, v)-optimal.*

Proof. Suppose first that $\|\,f\,\|_t < 1$. Then taking s sufficiently close to t and applying inequalities (2.5), (2.6), and (2.8), we can conclude that

$$\left| \int_0^t T(t - r)f(r)\,dr + T(t)\,u - T(s)u \right|_s < 1,$$

which plainly implies that f is not (u, v)-optimal. Similarly, if for some s, $0 < s < t$, we have $\|\,f\,\|_s < 1$, f will not be $(u, u(s))$-optimal, and a fortiori will not be (u, v)-optimal.

We return now to the general case. For some $\epsilon > 0$ there exists a non-null set $e \subseteq c$ with $|\,f(r)\,| \leqq 1 - \epsilon$, $r \in e$. Take s in e like in Lemma 2.1, and consider the operator $M\colon L_s(e) \to K$, defined by

$$Mg = \int_{e(0,s)} T(s - r)g(r)\,dr.$$

It is not difficult to see, examining the proof of Lemma 2.1, that the equation

$$Mg = \delta \int_0^s T(s - r)f(r)\,dr \tag{2.11}$$

will have, for δ sufficiently small, a solution g in $L_t(e)$ with $\| g \| \leqq \epsilon$. Consider now the control $h(r) = (1 - \delta)f(r) + g(r)$. By (2.11), h is (u, v)-admissible. But $\| h \|_s < 1$. This shows that f itself cannot be (u, v)-optimal.

COROLLARY 2.3. (UNIQUENESS THEOREM). *Let f and g be two (u, v)-optimal controls. Then*

(a) *both transition times coincide,*

(b) *f and g are equal for almost all points in $(0, t)$.*

Proof. (a) is trivial. Suppose (b) were false. Then $\frac{1}{2}(f + g)$ would also be an optimal control with norm less than 1 in a non-null set, which is absurd.

3. Generalizations. The present methods can be applied, in some cases with slight modifications, to the following more general situations.

(a) H is a reflexive Banach space.

(b) The elements u and v are replaced by closed convex sets, for instance $| u(0) - u |, | u(t) - v | \leqq \rho$.

(c) The arrival time is specified (instead of the departure time).

(d) $A = A(t)$. In this case, the weak solutions of $u_t = A(t)u + f$, $u(s) = u$, are given by

$$u(t) = U(t, s)u + \int_s^t U(t, r)f(r)\, dr,$$

where $U(t, s)u$ is the solution of $u_t = A(t)u$, $u(s) = u$ (see [5]).

REFERENCES

[1] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-groups*, Amer. Math. Soc. Colloquium Publications, XXXI, 2nd edition, Providence, 1957.

[2] S. BOCHNER AND A. E. TAYLOR, *Linear functionals on certain spaces of abstractly-valued functions*, Ann. of Math., 2, 39 (1938), pp. 913–944.

[3] R. E. BELLMAN, I. GLICKSBERG, AND O. A. GROSS, *On the "bang-bang" control problems*, Quart. Appl. Math., 14 (1956), pp. 11–18.

[4] R. S. PHILLIPS, *Perturbation theory for linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[5] T. KATO, *On linear differential equations in Banach space*, Comm. Pure Appl. Math., 74 (1956), pp. 479–486.

[6] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.

[7] I. V. EGOROV, *Optimal control in Banach space*, Dokl. Akad. Nauk. SSSR, 150 (1963), pp. 241–244; translated in Soviet Mathematics, 4 (1963).

J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

MODES OF FINITE RESPONSE TIME CONTROL*

C. A. HARVEY†

Summary. A linear autonomous system with a single control variable is considered. There are, in general, several modes of finite response time control for such a system. The concepts of single component regulation and multiple component regulation are defined. It is then shown that a multiple component regulation problem can be transformed into a single component regulation problem. Thus it is possible to express any of the modes of control considered as control of a single input, single output system.

Introduction. The system considered is represented by the vector differential equation

$$\dot{x}(t) = Ax(t) + bu(t), \tag{1}$$

where the dot denotes differentiation with respect to time t, $x(t)$ is a column vector with elements $x_1(t)$, $x_2(t)$, $\cdots$, $x_n(t)$ which describe the state of the system, $u(t)$ is a scalar control variable, A is a constant $n \times n$ matrix, and b is a constant column vector.

It is assumed that the system (1) is completely controllable. This means that for any initial state of the system there exists a control defined on a closed finite interval of time $[0, T]$ such that the state of the system arrives at the zero state ($x = 0$) at the time T. It is known [3, pp. 483–484] that a necessary and sufficient condition for complete controllability of system (1) is that the vectors b, Ab, $\cdots$, $A^{n-1}b$ be linearly independent, i.e.,

$$\det [b, Ab, \cdots , A^{n-1}] \neq 0.$$

Single component regulation is defined as control of the system such that one component of the state vector is transferred to zero in a finite time and held zero thereafter. *Multiple component regulation* is defined as control of the system such that more than one component of the state vector is transferred to zero in a finite time and held zero thereafter. As an example of a particular type of multiple component control a time optimal multiple component regulation problem could be defined when $u(t)$ is constrained in amplitude as follows: for any initial condition find a control satisfying the amplitude constraint on the interval $(0, \infty)$ such that the components to be controlled are transferred to zero in the minimum time such that they may be held at zero thereafter. The time optimal single

* Received by the editors August 28, 1962, and in revised form December 4, 1963. Prepared under contract NASw-563 for the National Aeronautics and Space Administration.

† Minneapolis-Honeywell Regulator Company, 2600 Ridgway Road, Minneapolis, Minnesota.

component regulation problem was first discussed by Schmidt [5, pp. 40–69] and was later treated by Harvey and Lee [1], [2], [4].

The definitions of single component and multiple component regulation given above are somewhat ambiguous and are not mutually exclusive. It is possible in some cases to state the same control problem as a single component or as a multiple component regulation problem. For example, consider the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u.$$

The single component regulation problem of controlling x_1 is the same as the multiple component regulation problem of controlling x_1 and x_2 since $x_2 = \dot{x}_1$, and a necessary condition for holding x_1 at zero is that x_2 be held at zero. Thus, whether this particular control problem is viewed as a single or multiple component regulation problem depends on the desire of the analyst.

The following section is devoted to a constructive proof of this paper's principal result.

Given a multiple component regulation problem, there exists a linear transformation of the state space such that the given problem is a single component regulation problem in the transformed state variables.

This result makes possible the application of the theory related to time-optimal single component regulation [1], [2], [4], [5] to time optimal multiple component regulation. Also, the result allows the control engineer faced with a multiple component regulation problem to reformulate the problem as a single input, single output problem with which he may have more familiarity.

Development of transformations. Consider the following multiple component regulation problem for the system (1). Suppose that the components $x_1, x_2, \cdots, x_m$, $1 \leqq m \leqq n$, are to be controlled; i.e., given an arbitrary initial condition $x(0) = x^0$, find a control $u(t)$, $0 \leqq t$, depending on x^0, such that the corresponding solution of (1) satisfies $x_1(t) = x_2(t) = \cdots = x_m(t) = 0$ for $t \geqq \tau$ for some real number τ which may depend on x^0.

For convenience the following notation is introduced. The vector x will be partitioned into two vectors ξ_1 and ξ_2 with $\xi_1 = (x_1, x_2, \cdots, x_m)'$ and $\xi_2 = (x_{m+1}, x_{m+2}, \cdots, x_n)'$ where $'$ denotes transpose. Also the vector b will be partitioned into two vectors $\beta_1 = (b_1, b_2, \cdots, b_m)'$ and $\beta_2 = (b_{m+1}, b_{m+2}, \cdots, b_n)'$. The matrix A will be partitioned into four submatrices, A_1, A_2, A_3, and A_4 with $A_1 = [a_{ij}]$, $1 \leqq i \leqq m$, $1 \leqq j \leqq m$; $A_2 = [a_{ij}]$, $1 \leqq i \leqq m$, $m + 1 \leqq j \leqq n$; $A_3 = [a_{ij}]$, $m + 1 \leqq i \leqq n$,

$1 \leqq j \leqq m$; $A_4 = [a_{ij}]$, $m + 1 \leqq i \leqq n$, $m + 1 \leqq j \leqq n$. Then (1) can be written as

$$\begin{aligned} \dot{\xi}_1 &= A_1\xi_1 + A_2\xi_2 + \beta_1 u, \\ \dot{\xi}_2 &= A_3\xi_1 + A_4\xi_2 + \beta_2 u. \end{aligned} \tag{2}$$

The following theorem, which is evident from an examination of (2), is readily established.

THEOREM 1. *If the system* (1) *is completely controllable, then A_2 and β_1 are not both zero.*

Proof. Suppose that A_2 and β_1 are both zero. Then it is easy to show that the vector $A^k b$ has zeros for its first m elements, with k a nonnegative integer. Thus the matrix $[b, Ab, \cdots, A^{n-1}b]$ has m rows of zeros and hence its determinant is zero.

It may occur, as in the example cited in the introduction, that the control of ξ_1 may imply the control of certain linear combinations of components of ξ_2. To examine this possibility, consider the requirement that $\xi_1(t) = 0$ for all $t \geqq T$ for some time T. From the system (2) it is clear that for $t \geqq T$,

$$\begin{aligned} 0 &= A_2\xi_2 + \beta_1 u, \\ \dot{\xi}_2 &= A_4\xi_2 + \beta_2 u. \end{aligned} \tag{3}$$

If $\beta_1 = 0$ then $A_2\xi_2 = 0$ for $t \geqq T$. Hence control to the subspace defined by $\xi_1 = 0$ implies control to the subspace $\hat{\xi}_1 = 0$ defined by $\xi_1 = 0$ and $A_2\xi_2 = 0$. $\hat{\xi}_1$ may be obtained by adjoining to ξ_1 the linearly independent elements of $A_2\xi_2$. The problem may then be restated with $\hat{\xi}_1$ and $\hat{\xi}_2$ (the projection of x onto $\hat{\xi}_1 = 0$) replacing ξ_1 and ξ_2. The matrices A_1, A_2, A_3, A_4 and the vectors β_1 and β_2 would of course have to be replaced with corresponding matrices and vectors. In case $\beta_1 \neq 0$, it is clear from (3) that $u = -\beta_1'A_2\xi_2/\| \beta_1 \|^2$, and hence $(\| \beta_1 \|^2 A_2 - \beta_1\beta_1'A_2)\xi_2 = 0$. As in the case when $\beta_1 = 0$ the problem can be reformulated with x partitioned into vectors $\hat{\xi}_1$ and $\hat{\xi}_2$. These procedures may be repeated until it is found that control to the subspace $\xi_1 = 0$ does not imply control to any smaller subspace. The number of reformulations is finite and is in fact less than or equal to $n - m$.

Now let us assume that the problem stated at the beginning of this section is the result of necessary reformulations so that control to the subspace $\xi_1 = 0$ does not imply control to any smaller subspace. This hypothesis guarantees that

$$\beta_1 \neq 0, \qquad A_2 = \beta_1\beta_1'A_2/\| \beta_1 \|^2. \tag{4}$$

To show this suppose that $\beta_1 = 0$. Then, since the system is assumed to be completely controllable, $A_2 \neq 0$, and control to the subspace $\xi_1 = 0$ implies control to the smaller subspace, $\xi_1 = 0$ and $A_2\xi_2 = 0$, which contradicts our hypothesis. Thus $\beta_1 \neq 0$ and hence $A_2 = \beta_1\beta_1'A_2/\| \beta_1 \|^2$, because if this were not the case, control to the subspace $\xi_1 = 0$ would imply control to the smaller subspace, $\xi_1 = 0$ and $(A_2 - \beta_1\beta_1'A_2/\| \beta_1 \|^2)\xi_2 = 0$, which contradicts our hypothesis.

With (4) established, the system (2) will be transformed into a particular form, in which it is evident that the problem is a single component control problem. Let $z = Sx$ where S is an $n \times n$ matrix partitioned into the submatrices S_1, S_2, S_3 and S_4 in the same manner that was used in partitioning A. The matrices S_2 and S_3 are zero matrices of appropriate size and S_4 is the $(n - m)$th order identity matrix. The matrix S_1 is defined indirectly by defining a matrix denoted by S_1^{-1} and the nonsingularity of S_1^{-1} is established in the next theorem.

THEOREM 2. *If the system* (1) *is completely controllable and* (4) *is satisfied, then* S_1^{-1} *is nonsingular, where* S_1^{-1} *is defined as*

$$S_1^{-1} = [A_1^{m-1}\beta_1, A_1^{m-2}\beta_1, \cdots, A_1\beta_1, \beta_1].$$

The proof of this theorem will be given following the proof of Theorem 3. Partitioning the vector z into m and $n - m$ vectors ζ_1 and ζ_2, the transformation may be written as $\zeta_1 = S_1\xi_1$, $\zeta_2 = \xi_2$. The transformed system is

$$\begin{aligned} \dot{\zeta}_1 &= S_1A_1S_1^{-1}\zeta_1 + S_1A_2\zeta_2 + S_1\beta_1 u, \\ \dot{\zeta}_2 &= A_3S_1^{-1}\zeta_1 + A_4\zeta_2 + \beta_2 u. \end{aligned} \tag{5}$$

The matrix S_1 has the property that $S_1\beta_1$ is a unit vector with its first $m - 1$ elements zero. From this result and (4) it is clear that the first $m - 1$ rows of S_1A_2 are zero and the last row is $\beta_1'A_2/\| \beta_1 \|^2$. The matrix $S_1A_1S_1^{-1}$ has ones on the superdiagonal, the first column is a vector c, and all other elements are zero, where the elements c_i satisfy

$$A_1^m = \sum_{i=1}^{m} c_i A_1^{m-i}.$$

From the form of (5) it is easy to establish the next theorem.

THEOREM 3. *Regulation of* z_1 (*the first component of* z) *is equivalent to the regulation of* ζ_1.

Proof. Clearly, regulation of ζ_1 implies regulation of z_1. From (5), $z_{k+1} = \dot{z}_k - c_k z_1$, $k = 1, 2, \cdots, m - 1$. Therefore

$$z_{k+1} = z_1^{(k)} - \sum_{j=0}^{k-1} c_{k-j} z_1^{(j)},$$

where $z_1^{(j)}$ denotes the jth time derivative of z_1. Thus ζ_1 can be expressed in terms of z_1 and its first $m - 1$ derivatives and hence regulation of z_1 implies regulation of ζ_1.

Proof of Theorem 2. From (4) it is clear that $A_2\beta$ is a multiple of β_1

for any $n - m$ vector β. Let γ_{1j} and γ_{2j} denote m and $n - m$ vectors, respectively, such that

$$A^j b = \begin{bmatrix} \gamma_{1j} \\ \gamma_{2j} \end{bmatrix}, \qquad \text{for each } j \geqq 0.$$

By induction it can be shown that

$$\gamma_{1j} = \sum_{k=0}^{j} \lambda_k A_1^{j-k} \beta_1 , \tag{6}$$

where λ_k is a scalar for $k = 0, 1, \cdots, j$; $\lambda_0 = 1$, and $A_2 \gamma_{2k} = \lambda_{k+1} \beta_1$. Denoting the matrix $[\beta_1, A_1\beta_1, \cdots, A_1^{m-1}\beta_1]$ by M and the matrix $[b, Ab, \cdots, A^{n-1}b]$ by N, the determinant of N may be written as

$$\det \begin{bmatrix} \gamma_{10} & \gamma_{11} & \cdots & \gamma_{1n} \\ \gamma_{20} & \gamma_{21} & \cdots & \gamma_{2n} \end{bmatrix}.$$

Using (6), the Cayley-Hamilton theorem and the elementary properties of determinants, this determinant may be written as

$$\det \begin{bmatrix} M & 0 \\ P & Q \end{bmatrix},$$

where 0 is the $m \times (n - m)$ matrix of zeros. Thus the determinant of N is the product of the determinants of M and Q. The determinant of N is nonzero since the system (1) is assumed to be completely controllable and hence the determinant of M is nonzero. But the determinant of M is the determinant of S_1^{-1}, so that S_1^{-1} is nonsingular.

Remarks. If ζ_1 is to be held zero after the response time T, it is clear from (5) that for $t \geqq T$,

$$u(t) = -\beta_1' A_2 \zeta_2(t) / \| \beta_1 \|^2, \tag{7}$$

and

$$\dot{\zeta}_2 = (A_4 - \beta_2 \beta_1' A_2 / \| \beta_1 \|^2) \zeta_2 . \tag{8}$$

If the control $u(t)$ is required to satisfy the constraint $| u(t) | \leqq 1$ for all t, it is necessary to consider $u(t)$ given by (7) and (8) with $\zeta_2(T)$ being the initial condition for (8). Satisfying the constraint imposes constraints on the initial condition $\zeta_2(T)$. It may occur that some constraints are of the form $\eta' \zeta_2(T) = 0$, where η is a constant n-m vector. In this case the control of ζ_1 implies the control to the subspace, $\zeta_1 = 0$ and $\eta' \zeta_2 = 0$, and the problem may then be reformulated to be control to this subspace.

Conclusions. It has been shown that multiple component regulation problems can be transformed into single component regulation problems for linear constant coefficient systems with a scalar control input. This permits one to view such problems as single input, single output control problems. The development presented is of a constructive nature so that the single output of the single component formulation of the regulation problem may be determined explicitly.

REFERENCES

[1] C. A. Harvey, *Determining the switching criterion for time-optimal control*, J. Math. Anal. Appl. 5 (1962), pp. 245–257.

[2] C. A. Harvey and E. B. Lee, *On the uniqueness of time-optimal control for linear processes*, Ibid., pp. 258–268.

[3] R. E. Kalman, *On the general theory of control systems*, Proc. First International Congress on Automatic Control, Moscow, 1960, Butterworths, London, 1961, pp. 481–492.

[4] E. B. Lee, *On the time optimal control of plants with numerator dynamics*, IRE Trans. on Automatic Control, 6 (1961).

[5] S. F. Schmidt, *The analysis and design of continuous and sampled-data feedback control systems with a saturation type nonlinearity*, NASA TN D-20, 1959, pp. 40–69.

J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

OPTIMAL PROGRAMS FOR AN ASCENDING MISSILE*

G. M. EWING† AND W. R. HASELTINE‡

1. Introduction. In 1919, R. H. Goddard proposed [1, p. 10] the problem of minimizing the mass of a given propellant required to transfer a rocket along a vertical path from rest on the earth to an assigned maximal height. He identified this as an unsolved problem of the calculus of variations but attempted neither a solution nor a precise formulation.

Although this problem, in one version or another, has interested many writers, no adequate treatment of any version has been published insofar as the present authors are aware. The object here is to give one.

Literature on the problem suffers from the vagueness that plagued early-day calculus of variations. Typical approaches equate to zero a formally derived first variation of the initial mass, often without identifying the class of function-triples (v,y,m) for velocity, displacement, and mass, in which a best one is sought, or stating any restrictions on the drag D. Without essential restrictions, there need not even exist an optimal program; without such restrictions, one cannot hope to prove that a particular program, suspected of being the best, does indeed have this property. One hopes that such an approach will at least yield necessary conditions on a best program. That it may not is pointed out in §13.

Authors have often overestimated the content of their work and others have referred to this or that paper as a solution when in fact only superficial aspects of the problem have been treated. For example, a system of Euler equations and transversality conditions with as many parameters as there are boundary conditions may be mistaken for a solution of the original problem.

For an introduction to a wide class of problems for which the Goddard Problem is a prototype, see articles by D. F. Lawden, G. Leitmann, and A. Miele in [2] and [3], with the included bibliographies. The problems are interesting, difficult, and tricky; they usually involve features not covered by existing books on variational theory.

We mention the work of Miele and Cavoti, [3], [4], [5], on a generalized Goddard Problem with bounded rate of mass-flow. Their mathematical model, in contrast with ours, is a classical program of Mayer [6, pp. 187–190].

* Received by the editors in revised form December 19, 1963.

† Department of Mathematics and Astronomy, University of Oklahoma, Norman, Oklahoma. The contribution of the first-named author was supported by the Air Force Office of Scientific Research under grant AF-AFOSR-211-63.

‡ Research Department, U. S. Naval Ordnance Test Station, China Lake, California.

They recognize the need for sufficient conditions, that the real objective is a global and not a local extremum, and show special cases, termed linear, for which sufficiency for the global extremum apparently follows from Green's theorem. They do not mention the class of programs in comparison with which this method identifies the best, nor examine the validity, for their procedure, of the changes in independent variables that are required, nor find conditions on drag D under which an extremum necessarily exists, nor deal with sufficiency in general nonlinear cases.

We also mention the work of Pontryagin [7] and others. His Maximum Principle is a necessary and not a sufficient condition. Our formulation of the Goddard Problem admits functions with many discontinuities as do the control problems of the Russian school, but our functional to be minimized does not fall under existence theorems based on weak compactness such as Theorem 1 of Lee and Markus [8], and certainly not under existence theorems requiring equicontinuity.

2. Formulation of the problem. We use the particle idealization, a flat stationary earth, and a uniform gravitational field. Like Hamel [9], Tsien and Evans [10], and Leitmann [11], we use a single stage rather than the continuous staging of Goddard's original description and of Leitmann [12].

Suppose we are given the positive numbers g, c, M, Y, the nonnegative number V, and a real-valued function D of v and y with suitable properties to be listed later.

That an ordered triple (v,y,m) of functions on an interval $[0,T]$ to the reals is an *admissible program* will mean that the following conditions hold.

(2.1) v is Lebesgue summable over $[0, T]$.

$$y(t) = \int_0^t v(s)\, ds. \tag{2.2}$$

For all t_1, $t_2 \in [0, T]$,

$$\begin{aligned} m(t_1) \exp \frac{v(t_1) + gt_1}{c} &= m(t_2) \exp \frac{v(t_2) + gt_2}{c} \\ &\quad + \frac{1}{c} \int_{t_1}^{t_2} D[v(t), y(t)] \exp \frac{v(t) + gt}{c}\, dt. \end{aligned} \tag{2.3}$$

(2.4) $m(t)$ is monotonic nonincreasing.

$$v(0) = 0, \quad y(0) = 0, \quad v(T) = V, \quad y(T) = Y, \quad m(T) = M. \tag{2.5}$$

For the moment, extend v by setting $v(t) = 0$ for $t < 0$ and $v(t) = V$ for $t > T$. As a consequence of (2.4) and (2.3), there is a possible countable set of t-values, all on the closed interval $[0, T]$, at each of which m has a negative jump and v a positive jump. Left limits $m(t-)$, $v(t-)$ exist everywhere, as do right limits $m(t+)$, $v(t+)$. Functions m and v are continuous except for the possible countable set.

We now adopt the convention that

$$\begin{aligned} &m(0) = m(0-), \qquad v(0) = v(0-), \\ &\qquad\qquad\qquad m(T) = m(T+), \qquad v(T) = v(T+). \end{aligned} \tag{2.6}$$

Values $m(t)$, $v(t)$ at a discontinuity are of no consequence; the one-sided limits are all that are needed and $m(t)$, $v(t)$ need not even be defined. It simplifies the exposition, however, to regard $m(t)$, $v(t)$ as defined everywhere. In order to satisfy (2.3) and (2.4), without exceptions, we require that

$$m(t-)e^{v(t-)/c} = m(t)e^{v(t)/c} = m(t+)e^{v(t+)/c}. \tag{2.7}$$

Clearly (2.6) is no real restriction.

Henceforth, we always consider the common domain of v, y, m to be a closed interval $[0, T]$ of the reals, as already stated above (2.1). When we mention $m(t)$ or $v(t)$, (2.7) is understood to hold. We shall also make statements involving one-sided limits.

Since the derivative $\dot{m}$ of a monotone function m exists and is finite a.e. (almost everywhere), it follows from (2.3) that $\dot{v}$ also exists and is finite a.e. and that

$$m\dot{v} + c\dot{m} + D(v, y) + mg = 0, \qquad \text{a.e.} \tag{2.8}$$

This familiar equation is not very useful unless m and v are both AC (absolutely continuous). Our m and v need not even be continuous, hence the formal integration by which one is tempted to go from (2.8) to (2.3) is not valid. The integral formulation (2.3) is essential in order to admit the largest possible class of programs (v, y, m).

By the *Goddard Problem* we mean the following questions:

(A) *Does there exist an admissible program* (v_0, y_0, m_0) *such that, in comparison with all admissible programs* (v, y, m), $m_0(0)$ *is the least value of* $m(0)$*?*

(B) *If so, what is the program* (v_0, y_0, m_0) *and is it unique?*

We are asking for the existence and characterization of the absolute or global minimum.

3. Existence of a minimizing program. We require of D that its domain consist of all ordered pairs (v,y) of reals, that it be of class C' in (v,y), that it be increasing in v for each y and nonincreasing in y for each v, that $D(v,y) > 0$ for $v > 0$, and that $D(0,y) = 0$.

Denote by K the class of admissible triples (v, y, m). Clearly K is not empty. It is shown in §15 that, if (v, y, m) is any admissible program, there always exists an admissible program (u, x, μ) with $u(t) \geqq 0$ on its interval and such that $\mu(0) \leqq m(0)$, with the strict inequality holding unless the original $v(t)$ is nonnegative. Therefore, in the search for a least value of $m(0)$, we need consider only the subclass K_1 of K consisting of those triples in K such that $v(t) \geqq 0$, or equivalently, such that $y(t)$ is nondecreasing. We henceforth use these properties of v and y without explicit mention of the restriction to triples in K_1.

The set of values $m(0)$ for triples in K_1 has an infimum $M_0 > M$. There necessarily exists a sequence (v_n, y_n, m_n) with domain $[0, T_n]$, $n = 1, 2, \cdots$, of triples in K_1 such that $m_n(0) \rightarrow M_0$ as $n \rightarrow \infty$. The sequence of numbers $m_n(0)$ necessarily has a finite upper bound $\hat{M}$.

From (2.3) with $t_1 = t$ and $t_2 = 0$, it follows that

$$m_n(t) \exp \frac{v(t) + gt}{c} < m_n(0);$$

hence that $v_n(t)$ has an upper bound $\hat{V}$,

$$\hat{V} = c \log \hat{M}/M. \tag{3.1}$$

With t_1, t_2 in (2.3) replaced by 0, T_n, observe that the numbers T_n have a finite upper bound $\hat{T}$; otherwise $m_n(0)$ cannot be bounded.

As a consequence of §14, under the conditions on D stated above, the total variation of v_n on $[0, T_n]$ is below a real number depending on $\hat{M}$, $\hat{V}$, and $\hat{T}$ but independent of n. We wish to apply a theorem of Helly [13, p. 29] on sets of uniformly bounded functions of uniformly bounded variation. To that end, extend v_n to the interval $[0, \hat{T}]$ by setting $v_n(t) = v_n(T_n) = V$ for $t > T_n$. By the Helly Theorem there exists a subsequence of v_n converging pointwise to a limit function v_0 on $[0, \hat{T}]$. The bounded sequence T_n, $n = 1, 2, \cdots$, may not converge but some subsequence will converge to a limit T_0. Let $v_n(t)$, $t \in [0, T_n]$, $n = 1, 2, \cdots$, now denote a sequence such that $m_n(0) \rightarrow M_0$, $v_n(t) \rightarrow v_0(t)$ on $[0, \hat{T}]$, and $T_n \rightarrow T_0$.

Lebesgue's Bounded Convergence Theorem applies to (2.2) with $v_n(s)$ as integrand and we define y_0 by the relation

$$y_0(t) = \int_0^t v_0(s)\, ds, \qquad 0 \leqq t \leqq T_0.$$

The same convergence theorem then applies to (2.3), written for (v_n, y_n, m_n), and we define m_0 by (2.3) with $t_2 = T_0$, $t_1 = t$, $v = v_0$, and $y = y_0$.

Since v_n, y_n, m_n satisfy boundary conditions (2.5) it follows that the respective limits v_0, y_0, m_0 satisfy (2.5). Since $m_n(t)$ is nondecreasing in t, so also must be the limit $m_0(t)$. Therefore (v_0, y_0, m_0) is admissible. Convergence of $m_n(t)$ to $m_0(t)$ applies in particular for $t = 0$, hence $m_0(0) = M_0$ and (v_0, y_0, m_0) is a triple such that, among all admissible triples, it furnishes the least possible value for $m(0)$.

4. Additional restrictions on D. The Existence Theorem used results from §14 and §15 depending only on the mild restrictions stated at the beginning of §3. Our sufficiency and uniqueness arguments require additional properties of D, which we state for reference.

Let h be a nonnegative constant and require of $D(v, y)$, for $v \geqq 0$:

$$D(v, y) = D_0(v)e^{-hy}, \tag{4.1}$$

$$D_0 \text{ is of class } C''', \text{ that is, } D_0''' \text{ is continuous,} \tag{4.2}$$

$$D_0'(0) = 0, \tag{4.3}$$

$$D_0''(v) + (1/c)\, D_0'(v) > 0. \tag{4.4}$$

We anticipate that suitable drag-functions D, not exponential in y, can be used with no change in our principal conclusions but the details will be more complex.

5. Thrust-free flight. One type of optimal program (v_0, y_0, m_0) will include a coast after burnout. In studying this type of motion it is convenient to shift the time-origin so that 0 now corresponds to the assigned terminal values Y, V of $y(t)$, $v(t)$. The differential equations for motion with no thrust are

$$\dot{v} = -g - D(v, y)/M, \qquad \dot{y} = v, \tag{5.1}$$

and the terminal conditions are

$$y(0) = Y > 0, \qquad v(0) = V \geqq 0. \tag{5.2}$$

Various existence theorems ensure that there is a pair of functions v, y satisfying (5.1) and (5.2) on some maximal interval $(t_1, 0]$. It is clear from the form of (5.1), that if $t_1 = -\infty$, then $v(t) \rightarrow +\infty$ and $y(t) \rightarrow -\infty$ as $t \rightarrow t_1$, and that there is a unique negative time T_a such that $y(T_a) = 0$.

If $t_1 > -\infty$, a case which occurs, for example, if $D_0(v) = v^3 + v^2$ for $v \geqq 0$, there is then a number y_1 such that $v(t) \rightarrow +\infty$ and $y(t) \rightarrow y_1$ as $t \rightarrow t_1$ from above. If $y_1 < 0$, we again define T_a by the relation $y(T_a) = 0$; if $y_1 \geqq 0$, we define T_a as t_1.

Let $y = \gamma$, $v = \dot{\gamma}$ denote the solution of (5.1), (5.2) on the closed interval $[T_a, 0]$ or the half-open interval $(T_a, 0]$, according as $y(T_a) = 0$ or $T_a = t_1$. Let Γ denote the oriented path in the (t, y) plane defined by $y = \gamma(T)$ with the positive sense determined by increasing t.

6. An Euler equation. Set

$$f(t, y, v) = (1/c)D(v, y) \exp \frac{v + gt}{c}. \tag{6.1}$$

An important role will be played by solutions of the Euler equation

$$\frac{d}{dt} f_v(t, y, \dot{y}) = f_y(t, y, \dot{y}), \tag{6.2}$$

in which subscripts denote partial differentiation.

Heuristic reasons for suspecting the relevancy of (6.2) to our problem are found in such papers as [9] and [10], to which we are indebted. If we were to replace our class of admissible triples by all triples (v, y, m) such that v and m are possibly discontinuous at take-off but are continuously differentiable everywhere else and if we require (2.2), (2.3) and (2.5), but not (2.4), then the procedure of [10] yields (6.2) as a necessary condition on $v = \dot{y}$ and y for $0 < t < t_b =$ burnout time.

If we add restriction (2.4) on m to those stated above, the problem of minimizing $m(0)$ can be expressed as a classical problem of Bolza [6, p. 189], and (6.2) again appears, this time by way of the Multiplier Rule.

These remarks are suggestive but no more. Neither of the problems described above is our Goddard Problem.

7. Some consequences of §2, §4, and §6. With primes denoting differentiation, set

$$F(v) = D_0'(v) + D_0(v)/c, \tag{7.1}$$

$$G(v) = vF(v) - D_0(v), \tag{7.2}$$

$$H(v) = F'(v) + F(v)/c. \tag{7.3}$$

From (4.1) and (6.1), Euler equation (6.2) is seen to be equivalent to the system,

$$H(v)\dot{v} = hG(v) - gF(v)/c, \qquad \dot{y} = v. \tag{7.4}$$

Conditions on drag D in §2 and §4 ensure that

$$D_0(0) = F(0) = G(0) = 0, \qquad H(0) > 0, \tag{7.5}$$

and that

$$D(v, y) > 0, \quad F(v) > 0, \quad G(v) > 0, \quad H(v) > 0, \quad \text{for} \quad v > 0. \tag{7.6}$$

Moreover,

$$G(v) \to +\infty \quad \text{as} \quad v \to +\infty, \tag{7.7}$$

$$F(v)/H(v) < c, \quad G(v)/H(v) < cv, \quad \text{for} \quad v \geqq 0, \tag{7.8}$$

and

$$F(v)/H(v) = O(v), \quad G(v)/H(v) = O(v^2), \quad \text{for small positive } v. \tag{7.9}$$

When $\dot{m}(t)$ and $\dot{v}(t)$ both exist and are finite, then (2.8) is meaningful and correct. If system (7.4) holds and $v(t) \geqq 0$, we then find, with form (4.1) for $D(v, y)$, that

$$(7.10) \qquad c\dot{m} = -D_0 e^{-hy} - m[g(D_0'' + D_0'/c) + hG]/H.$$

The right member is not positive and therefore

$$(7.11) \qquad \dot{m}(t) \leqq 0 \quad \text{a.e.}$$

It can be verified that Euler equations (7.4) and trajectory equation (2.8), taken together, have, as a first integral, the relation

$$(7.12) \qquad [G(v)e^{-hy} - mg] \exp \frac{v + gt}{c} = \text{constant}.$$

8. Construction of a field in the large. We use the time-scale of §5, in which the interval for a triple (v,y,m) is $[T,0]$, $T < 0$. The following conditions on such an interval are equivalent to the defining properties (2.1) through (2.5) of an admissible program:

$$(8.1) \qquad v \text{ is Lebesgue summable,}$$

$$(8.2) \qquad y(t) = Y - \int_t^0 v(s)\, ds,$$

$$(8.3) \qquad \text{condition (2.3) for } t_1, t_2 \text{ in } [T, 0],$$

$$(8.4) \qquad m(t) \text{ is monotonic nonincreasing,}$$

$$(8.5) \quad v(T) = 0, \quad y(T) = 0, \quad v(0) = V, \quad y(0) = Y, \quad m(0) = M.$$

With T_a defined as in §5, let

$$(8.6) \qquad t_\alpha = \begin{cases} \alpha, & T_a < \alpha \leqq 0, \\ 0, & 0 < \alpha \leqq V/g. \end{cases}$$

Define, with reference to §5 for γ,

$$(8.7) \qquad u_0(\alpha) = \begin{cases} \dot{\gamma}(t_\alpha), & T_a < \alpha \leqq 0, \\ V - \alpha g, & 0 < \alpha \leqq V/g, \end{cases}$$

$$(8.8) \qquad x_0(\alpha) = \begin{cases} \gamma(t_\alpha), & T_a < \alpha \leqq 0, \\ Y, & 0 < \alpha \leqq V/g. \end{cases}$$

Denote by

$$(8.9) \qquad v = u(t, \alpha), \qquad y = x(t, \alpha), \qquad T_a < \alpha \leqq V/g,$$

a solution of system (7.4), and hence of the Euler equation (6.2), satisfying the conditions

$$(8.10) \qquad u(t_\alpha, \alpha) = u_0(\alpha), \qquad x(t_\alpha, \alpha) = x_0(\alpha).$$

A sketch showing graphs of $y = x(t, \alpha)$ intersecting Γ tangentially from the left, for $T_a < \alpha < 0$, and issuing from $(t, y) = (0, Y)$ to the left with slopes $V - \alpha g$ at $(0, Y)$, for $0 \leqq \alpha \leqq V/g$, will clarify the above choice of notation.

These remarks apply to both of the cases, $y(T_a) = 0$ and $y(T_a) > 0$, of §5. In the first of these, extend the family (8.9) by including $\alpha = T_a$.

Standard existence theorems in the small for differential equations ensure that a solution (8.9) exists for each α and for t on some interval to which t_α is interior. For present purposes, restrict each such solution to a time-interval terminating at t_α as is customary in envelope theorems of the calculus of variations [14, p. 131, pp. 140–141].

It is clear from the form of (7.4) and the first inequality (7.8) that $du/dt > -g$. There are moreover positive numbers a_1 and v_1 such that $du/dt < -a_1 u$ if $0 \leqq u \leqq v_1$; hence if for any α on the interval $(T_a, V/g]$, solution (8.9) could not be extended to an arbitrarily long interval $(t, T_a]$, we could reach a contradiction. We may also conclude that for $\alpha < V/g$ and $t \leqq t_\alpha$, $u(t, \alpha)$ is positive and bounded from zero, and therefore that there is a t such that $x(t, \alpha) = 0$. Moreover it is easy to see that this t can be made to be as close to T_a on the left as we please, simply by taking α sufficiently close to T_a on the right.

Observe in particular of the family (8.9) that $u(t, V/g) \equiv 0$ and $x(t, V/g) \equiv Y, t \leqq 0$, and that, for each such $t, u(t, \alpha) \rightarrow 0$ and $x(t, \alpha) \rightarrow Y$ as $\alpha \rightarrow V/g$. According to existence theorems for differential equations, u and x are of class C' in t, t_α, $u_0(\alpha)$, $x_0(\alpha)$ for α in $[T_a, V/g]$ or $(T_a, V/g)$, depending on which case of §5 we may have. Moreover, t_α, $u_0(\alpha)$, and $x_0(\alpha)$ are continuous and continuously differentiable in α with the exception of $\alpha = 0$ when the assigned terminal velocity V is positive, in which event they are continuous and right and left differentiable at $\alpha = 0$.

Denote by R the subset of the (t, y) plane bounded by the halflines $y = Y$, $t \leqq 0$ and $y = 0$, $t < T_a$, together with the path Γ and, in the event that y_1 of §5 is positive, by the vertical segment $t = T_a$, $0 \leqq y \leqq y_1$. The definition of R is completed by assigning to it all of its boundary points except those on the possible vertical segment.

It follows from the properties of $u(t, \alpha)$ already discussed and the properties (7.5), (7.6) of F, G, and H that R is simply covered by the family $y = x(t, \alpha)$, except for the point $(0, Y)$, which is the common right terminal of $y = x(t, \alpha)$, $0 \leqq \alpha \leqq V/g$. With this exception, there is a slope-function p such that $p(t, y) = x_t(t, \alpha)$ if $y = x(t, \alpha)$ is the unique member of the family through (t, y) in R. $p(t, y)$ is continuous in (t, y) in R, and its first partial derivatives are continuous in R except along Γ and curve C_0 : $y = x(t, 0)$ (for $V > 0$).

9. Invariance of the Hilbert integral. Let R_1 denote R with the point $(0,Y)$ deleted. Let S_1 denote the class of all piecewise smooth oriented paths C: $y = y(t)$, that are contained in R_1 with the positive sense on C determined by increasing t.

The usual considerations show that the Hilbert integral

$$I^*(C) = \int [f(t, y, p) + (\dot{y} - p)f_v(t, y, p)]\, dt \tag{9.1}$$

is independent of the choice of C in S_1 joining given endpoints so long as C does not include points on both sides of $C_0 : y = x(t, 0)$. At points of C_0, p is continuous but, if $V > 0$, p has distinct one-sided derivatives.

Given a path C in S_1 joining points on opposite sides of C_0, one verifies that there is a path C_1 in S_1 having exactly one point in common with C_0 and such that $I^*(C_1) = I^*(C)$. It follows that $I^*(C)$ has the same value for all C in S_1 joining points on opposite sides of C_0.

Consider next two piecewise smooth paths $C_1 : y = y_1(t)$ and C_2: $y = y_2(t)$ in R with the common endpoints (t_0, y_0) and $(0, Y)$. If C_1 and C_2 coincide on some subinterval $[t_1, 0]$ of $[t_0, 0]$, then clearly $I^*(C_1) = I^*(C_2)$. If not, let t_n, $n = 1, 2, \cdots$, be a sequence in $(T_a, 0)$ converging to 0 and such that $y_1(t_n) \neq y_2(t_n)$. If $y_1(t_n) < y_2(t_n)$, a line segment of slope $2\dot{\gamma}(t_n)$ from $[t_n, y_1(t_n)]$ to a point (t_n^*, Y), joined to the part of C_1 that terminates at $[t_n, y_1(t_n)]$ defines a path C_{1n} in S_1. Construct similarly a path C_{2n} in S_1. Now $I^*(C_{1n}) = I^*(C_{2n})$ by preceding results and, if we let $n \to \infty$, we are led to the conclusion that $I^*(C_1) = I^*(C_2)$.

Finally let C be a path in R defined by the second component y of any admissible triple. Since y is an integral (2.2) of a summable function v, we know that y is absolutely continuous on its interval. Let C_n be a sequence of piecewise smooth paths in R, coterminal with C and defined by functions y_n, $n = 1, 2, \cdots$, converging in length to C. The difference $I^*(y_n) - I^*(y)$ is the sum of integrals,

$$\int [f(t, y_n, p_n) - f(t, y, p)]\, dt,$$

$$\int [pf_v(t, y, p) - p_n f_v(t, y_n, p_n)]\, dt,$$

$$\int \dot{y}[f_v(t, y_n, p) - f_v(t, y, p)]\, dt,$$

$$\int (\dot{y}_n - \dot{y})f_v(t, y_n, p_n)\, dt,$$

in which p and p_n denote $p(t, y)$ and $p(t, y_n)$ and for which the suppressed limits are endpoints of the common interval of y and y_n.

The various terms in f and f_v are bounded, hence the first two integrals converge to zero as n becomes infinite. The third integral tends to zero since the second factor of the integrand tends to zero and its first factor is summable. We reach the same conclusion for the last integral as a consequence of the boundedness of the second factor and of the fact [15, p. 247] that convergence in length of y_n to y implies that $\int |\dot{y}_n - \dot{y}|\, dt$ converges to zero. Therefore $I^*(y_n) \to I^*(y)$ and, since $I^*(y_n)$ is independent of n, $I^*(y_n) = I^*(y)$.

Integral (9.1) thus has the same value for all AC functions y defining paths C in R having the same endpoints.

10. Further necessary properties of an optimal program. We require that D have all properties stated in §3 and §4.

By §3, there exists a minimizing triple (v_0, y_0, m_0); as a consequence of §15, $v_0(t)$ is nowhere negative. The principal results of this section are that v_0 and m_0 are both necessarily continuous on the interior of their common interval, and that $y = y_0(t)$, $T \leqq t \leqq 0$ is a member of a certain family of extremals.

Let C be a path defined by the second component y of an admissible triple. As a consequence of §15 and §16, we may restrict attention to the case in which C is in the subset R of the (t, y) plane introduced in §8. In the light of §15, we need consider only the case in which $y(t)$ is nondecreasing. Let $[T, 0]$ be the domain of y as in §5 and §8. We have remarked in §9 that y is AC on $[T, 0]$.

With C fixed, there exists a path E_α coterminal with C, where E_α is defined by $y = x(t, \alpha)$, introduced in (8.9), for $t \leqq \alpha$ and, in the event that $\alpha < 0$, E_α coincides with Γ for $\alpha \leqq t \leqq 0$. Denote by $M_0(C)$, $M_0(E_\alpha)$ the respective values of initial mass corresponding to C and E_α; by $I(C)$, $I(E_\alpha)$ the respective values of $\int f(t, y, \dot{y})\, dt$, where f is given by (6.1); and by $I^*(C)$, $I^*(E_\alpha)$ the respective values of the Hilbert integral. We now prove that

$$(10.1)\quad [M_0(C) - M_0(E_\alpha)]e^{gT/c} = \int_T^0 E\{t, y(t), p[t, y(t)], \dot{y}(t)\}\, dt,$$

in which the integrand is the Weierstrass E-function,

$$E(t, y, p, \dot{y}) = f(t, y, \dot{y}) - f(t, y, p) - (\dot{y} - p)f_v(t, y, p).$$

By (2.3),

$$(10.2)\qquad M_0(C)e^{gT/c} = Me^{V/c} + I(C),$$

and

$$(10.3)\qquad M_0(E_\alpha)e^{gT/c} = Me^{V/c} + I(E_\alpha).$$

Along E_α, except at $t = 0$ and possibly at t_α, $\dot{y}(t) = p[t, y(t)]$; hence $I(E_\alpha) = I^*(E_\alpha)$, while $I^*(E_\alpha) = I^*(C)$ by §9. It then follows from (10.2) and (10.3) that $[M_0(C) - M_0(E_\alpha)]e^{gT/c} = I(C) - I^*(C)$, and this is relation (10.1).

The E-function, interpreted relative to the indicatrix $z = f(t, y, v)$ in the (v, z) plane for each fixed (t, y), [16, p. 77], is the difference between the ordinate to the indicatrix and that to its tangent line for $v = p(t, y)$. Since $f_{vv} = H(v) \exp[-hy + (v + gt)/c]$ is positive for all nonnegative v, the integrand in (10.1) is nonnegative on $[T, 0]$ and indeed strictly positive on a subset of positive measure of that interval unless C and E_α are identical. It then follows that

$$M_0(C) > M_0(E_\alpha), \qquad \text{if} \qquad C \neq E_\alpha. \tag{10.4}$$

As a consequence, the first two components of the minimizing triple must be in the one-parameter family (8.9); the third component is then given by (2.3). According respectively as $T_a < \alpha \leqq 0$ or $0 < \alpha \leqq V/g$, v_0 and m_0 will have a single discontinuity at T or discontinuities at both T and 0.

11. Characterization of the optimal program. It remains only to minimize $M_0(E_\alpha)$ with respect to the parameter α.

Symbols $u(t, \alpha)$, $x(t, \alpha)$ are as defined by (8.9) if $\alpha \geqq 0$; if $\alpha < 0$, we interpret $u(t, \alpha)$, $x(t, \alpha)$ as the extensions described preceding (10.1). Thus $y = x(t, \alpha)$, possibly extended, determines the path E_α and, if $\alpha < 0$, E_α coincides in part with Γ.

We have remarked in §8 that $u(t, \alpha)$ is, for $\alpha \neq V/g$, always positive, hence the relation $x = x(t, \alpha)$ determines $t = t(x, \alpha)$. Define $w(x, \alpha)$ as $u[t(x, \alpha), \alpha]$. Then $w(x, \alpha) > 0$ if $x < Y$. Moreover if the terminal velocity V is positive, $w(x, \alpha)$ is bounded from zero, while if $V = 0$, $w(x\ \alpha)$ behaves like $\sqrt{Y - x}$ for x near Y. Also define $\mu(x, \alpha)$ as $m[t(x, \alpha), \alpha]$. Thus t, w, and μ are functions of (x, α) for $0 \leqq x \leqq Y$ and $T_a \leqq \alpha < V/g$, with

$$t(x, \alpha) = -\int_x^Y dz/w(z, \alpha), \tag{11.1}$$

and

$$\mu(x, \alpha) \exp \frac{w(x, \alpha) + gt(x, \alpha)}{c} = Me^{V/c} + (1/c) \int_x^Y f(t, z, w)\, dz/w(z, \alpha). \tag{11.2}$$

Using the time scale of §8, denote by T_α the negative time such that $x(T_\alpha, \alpha) = 0$. Thus $T_\alpha = t(0, \alpha)$ and $dT_\alpha/d\alpha = t_\alpha(0, \alpha)$. Subscript α on t, w, or μ denotes partial differentiation.

From (11.1),

$$t_\alpha(0, \alpha) = \int_0^Y w_\alpha(z, \alpha)\, dw/w^2(z, \alpha).$$

Observe that $\mu(0, \alpha) = M_0(E_\alpha)$. We find by differentiation of (11.2) with reference to §7 that

$$(11.3) \quad \mu_\alpha(0, \alpha) = (1/c)e^{-gT_\alpha/c} \int_0^Y [G(w)e^{-hx} - \mu g](w_\alpha/w^2) \exp \frac{w + gt}{c}\, dy.$$

For $T_a < \alpha < V/g$, we have $T_a < t_\alpha \leqq 0$. Also $w_\alpha(z, \alpha) = 0$ if $\gamma(t_\alpha) < z < Y$, while $w_\alpha(z, \alpha) < 0$ if $0 < z < \gamma(t_\alpha)$; hence the upper limit of the integral (11.3) can be replaced by $\gamma(t_\alpha)$. The statement about the sign of w_α may be justified as follows: First, $dw/dx = (1/u[t(x, \alpha), \alpha])du/dt$, since u is nowhere zero for $\alpha < V/g$. Second, at $x = \gamma(t_\alpha)$ for $T_a < \alpha < 0$, $dw/dx > \partial\dot\gamma(t(x, \alpha))/\partial x$, because $\dot m < 0$ along $y = x(t, \alpha)$; and $d\gamma(t_\alpha)/d\alpha = \dot\gamma(t_\alpha) > 0$ for $T_a < \alpha < 0$; while for $0 \leqq \alpha < V/g$, $w_\alpha > 0$ at $x = y$. Then (7.4), (7.9), and (4.2), together with standard theorems, ensure that w_α exists in the interior of R, except along C_0 where the right and left derivatives exist, and furthermore that $w_\alpha < 0$.

As a consequence of (7.12),

$$(11.4) \qquad [G(w)e^{-hx} - \mu g] \exp \frac{w + gt}{c}$$

is constant for $0 < x < \gamma(t_\alpha)$. As $x \to \gamma(t_\alpha)-$, $\mu(x, \alpha) \to Me^{+\alpha g}$ or M according as $0 \leqq \alpha < V/g$ or $T_a < \alpha < 0$ respectively; hence $\mu[\gamma(t_\alpha)-, \alpha]$ is monotonic increasing in α.

Now from (7.2), (7.3), $G_v = vF' + F - D_0' = v(D_0'' + D_0'/c) + D_0/c$. Hence, by (4.4) and (7.6), $G_v > 0$ for $v > 0$. At t_α, $d\dot\gamma/d\alpha < 0$, and x_α is positive for $T_a < \alpha < 0$, and zero for $0 < \alpha < V/g$. Thus the value of $G(w)e^{-hx}$ at $x = \gamma(t_\alpha)-$ is a monotonic decreasing function of α, which is zero at $\alpha = V/g$. For α sufficiently near V/g, the bracket in (11.4) is negative for all x on the interval $(0, \gamma(t_\alpha))$.

It may happen that for some α_0 between T_a and V/g, the bracketed expression vanishes for $x = \gamma(t_\alpha)$. It is necessarily so if the y_1 of §5 is positive. Whenever there is such an α_0, the integrand of (11.3) is positive or negative for $0 < x < \gamma(t_\alpha)$ according as $\alpha > \alpha_0$ or $\alpha < \alpha_0$. If there is no such α_0, the integrand is positive for all $\alpha > T_a$ and $M_0(E_\alpha)$ assumes its smallest value at $\alpha = T_a$.

In any event there is a unique α, corresponding to which $M_0(E_\alpha)$ is a minimum.

12. The zero-drag case. This essentially trivial and well-known case is excluded by (4.4) from parts of our theory. If $D(v,y) = 0$, the path Γ of §5 clearly exists and can be extended downward from the summit arbitrarily far. From (2.3) with $D(v, y) = 0$,

$$m(0) = M \exp \frac{V + gT}{c}. \tag{12.1}$$

For any admissible program (v, y, m) such that y is not identical with γ, the time of flight will exceed the positive number $-T_a$. It follows that (12.1) is a minimum if and only if $\alpha = T_a$.

13. Description of the optimal program. If there is no drag or if the assigned height Y and the effect of drag are small enough, the optimal program consists of an initial boost from $v(0) = 0$ to $v(0+) > 0$ followed by a coast to height Y and velocity V.

If this case does not occur and if, for a given Y, V is small enough, the minimizing program consists of an initial boost followed by a propulsive phase in which the Euler equation (6.2) is obeyed and then a coast. For $V = 0$, this is the case exhibited by Tsien and Evans [10] without using a monotonicity restriction on m or any restrictions on D or showing for any class of programs including this m that it is the best.

If both Y and V are large enough the best program consists of an initial boost, a variable thrust phase subject to (6.2) and a terminal boost at height Y with no coast.

The three cases correspond respectively to $\alpha_0 = T_a$, $T_a < \alpha_0 \leqq 0$, and $0 < \alpha_0 < V/g$.

In variational problems without side-conditions that introduce boundaries in function-space or otherwise restrict the functions that are admitted, stationarity is a necessary condition on whatever it is desired to minimize.

In the present problem, consider a particular case for which the minimizing program has a coasting phase covering the altitude range $0 < y_b < y \leqq Y$, where y_b is the burnout altitude. By regarding t, v, m as functions of y, and then allowing variations of v on $y_b < y < Y$, it is easy to construct a family of varied admissible programs depending on a parameter b, such that the minimizing program is that member of the family specified by $b = 0$, changes in t, y, m are uniformly small of first order in b, and $dm(0, b)/db\,|_{b=0} > 0$. The solution program, in other words, is not even weakly stationary.

14. Velocities of bounded variation. This section together with the next two contain results already used at crucial points in the theory. In this section drag D is required to have the properties stated at the beginning of §3. We show that if (v, y, m) is admissible, then v is a function of bounded variation. We use the time scale of §2, hence the domain of v, y, and m is an interval $[0,T]$.

Given a triple (v, y, m) satisfying (2.1), (2.2), and (2.3), together with numbers t_1, t_2 on $[0, T]$, then $m(t_1)$ in (2.3) increases strictly with $m(t_2)$.

If (v^*, y^*, m^*) also satisfies (2.1) through (2.3) with $v^*(t) \equiv v(t)$ on $[t_1, t_2]$ and $y^*(t_2) = y(t_2)$, then by (2.3),

$$\begin{aligned}\frac{m^*(t_1)}{m^*(t_2)} - \frac{m(t_1)}{m(t_2)} = \frac{1}{c}\left[\frac{1}{m^*(t_2)} - \frac{1}{m(t_2)}\right] \\ \cdot\exp\frac{-[v(t_1) + gt_1]}{c} \int_{t_1}^{t_2} D(v, y) \exp\frac{v + gt}{c}\, dt.\end{aligned} \tag{14.1}$$

If $t_1 < t_2$ and if $m(t_1) \geqq m(t_2)$, $m^*(t_2) \leqq m(t_2)$, and $v(t) \geqq 0$ on the open interval (t_1, t_2), it follows from (14.1) that

$$m^*(t_1) \geqq m^*(t_2). \tag{14.2}$$

We also obtain this conclusion in the form

$$m^*(t_1) \geqq m^*(t_2-), \tag{14.3}$$

if t_2 in the hypotheses above is replaced by t_2-, that is, if the hypotheses are in terms of left limits at t_2.

Let (v, y, m) be admissible in the sense of §2. Suppose that $t_1 < t_2$ and $v_1 = v(t_1) > v(t_2) = v_2$. Then

$$v_1 - v_2 < ce^{-v_2/c}(e^{v_1/c} - e^{v_2/c}). \tag{14.4}$$

By (2.3) the difference in the parentheses on the right of (14.4) is the sum of three terms A_1, A_2, A_3, where

$$A_1 = -\left[1 - \frac{m(t_2)}{m(t_1)}\right] \exp\frac{v(t_2) + g(t_2 - t_1)}{c},$$

$$A_2 = e^{v_2/c}[e^{g(t_2-t_1)/c} - 1],$$

$$A_3 = \frac{1}{cm(t_1)} e^{-gt_1/c} \int_{t_1}^{t_2} D(v, y) \exp\frac{v + gt}{c}\, dt.$$

A_1 is negative. If $v(t_2)$ is negative, v is also negative on some maximal open interval to the left of t_2, say (t_1, t_2). With this choice of t_1 and t_2, A_3 is negative. (For our conditions on D stated at the beginning of §3 require that $D(v, y) < 0$ if $v < 0$.) Then

$$v_1 - v_2 < ce^{-v_2/c} A_2 < ce^{gT/c}.$$

Hence v is bounded below. It is even easier to see that v must be bounded above for m to exist almost everywhere. Let $\bar{V}$ be an upper bound of $|v|$.

Let $k_1 = \sup | D(v, y) |$ for $| v | < \bar{V}$, $| y | < \bar{V}T$. Again taking $t_1 < t_2$, $v(t_1) > v(t_2)$, we find by estimation of A_2 and A_3 that

$$(14.5) \qquad v_1 - v_2 < \left(g + \frac{k_1}{M}\right) \exp \frac{2\bar{V} + gT}{c} (t_2 - t_1).$$

In the event that $v_1 < v_2$, one finds by a similar argument that

$$(14.6) \quad v_2 - v_1 < \{k_1(t_2 - t_1) + c[m(t_1) - m(t_2)]\}\left(\exp \frac{2\bar{V} + gT}{c}\right) \Big/ M.$$

It follows from (14.5) and (14.6) that the total variation of v on $[0, T]$ has a bound depending only on $m(0)$, $\bar{V}$, and T in addition to the constants g, c, and M.

15. Elimination of negative velocities. Conditions on D are those stated in the opening paragraph of §3.

Let (v, y, m) be admissible in the sense of §2. Since y is continuous on $[0, T]$, there is a least value $\hat{t}$ of t such that $y(\hat{t}) = Y$. Then $v(\hat{t}-) \geqq 0$ and $y(t) < Y$ if $t < \hat{t}$.

Suppose that there is a value t_a in $(0, \hat{t})$ such that $v(t_a) < 0$. There is then a least value t_1 of t, $0 \leqq t_1 < t_a$, such that $y(t)$ has its maximum value on the interval $[0, t_a]$ at t_1. Then $v(t_1+) \leqq 0$, but $v(t_1-) \geqq 0$, hence $v(t_1) = 0$. Clearly $y(t_a) < y(t_1) < Y$ and $y(t) \leqq y(t_1)$ if t is in the closed interval $[t_1, t_a]$.

There must also exist a largest value t_2 of t such that $y(t_2) = y(t_1)$ and $y(t) \leqq y(t_1)$ for $t_1 \leqq t \leqq t_2$. Then $v(t_2-) \geqq 0$.

Consider the function v_1,

$$(15.1) \qquad v_1(t) = \begin{cases} v(t), & 0 \leqq t \leqq t_1, \\ v(t + t_2 - t_1), & t_1 < t \leqq T - (t_2 - t_1). \end{cases}$$

Then define $y_1(t)$ by (2.2) and $m_1(t)$ by (2.3). It follows that y_1 and m_1 are related to y and m respectively in the same way that v_1 is related to v in (15.1).

The step from (v, y, m) to (v_1, y_1, m_1) consists of deletion of the half-open interval $(t_1, t_2]$, followed by drawing together the separated parts of the t axis and replacement of the original interval $[0, T]$ by a shorter interval.

We wish to show that

$$(15.2) \qquad m_1(t_1) < m(t_1).$$

If $v(t_2-) > 0$, there is an open subinterval (t_3, t_2) of (t_1, t_2) such that $v(t_3-) \leqq 0$ and on which $v(t)$ is positive. By (2.3),

$$(15.3) \qquad m(t_3-) > m(t_2+) \exp \frac{v(t_2+)}{c}.$$

If $v(t_2-) = 0$, then

$$m(t_2-) = m(t_2+) \exp \frac{v(t_2+)}{c}, \tag{15.4}$$

and there is an open subinterval (t_3, t_4) of (t_1, t_2) for which, again, $v(t_3-) \leqq 0$ and on which $v(t)$ is positive. By (2.3),

$$m(t_3-) > m(t_4-). \tag{15.5}$$

From monotonicity of $m(t)$,

$$m(t_4-) \geqq m(t_2-), \tag{15.6}$$

and, in either case,

$$m(t_1-) \geqq m(t_3-). \tag{15.7}$$

Hence

$$m(t_1) > m(t_2+) \exp \frac{v(t_2+)}{c}. \tag{15.8}$$

But

$$m_1(t_1) = m(t_2+) \exp \frac{v(t_2+)}{c}, \tag{15.9}$$

and (15.8) and (15.9) imply (15.2). We note that $m_1(t) = m(t + t_2 - t_1)$ on $t_1 < t \leqq T - (t_2 - t_1)$, and $m_1(t) < m(t)$ on $0 \leqq t \leqq t_1$. In particular,

$$m_1(0) < m(0). \tag{15.10}$$

If all negative values of $v(t)$ happened to occur on the deleted interval $(t_1, t_2]$, so that $v_1(t) \geqq 0$ on its interval, we could then identify (v_1, y_1, m_1) with (v^*, y^*, m^*) of §14, let the interval $[t_1, t_2]$ of that section be any subinterval of the present $[0, t_1]$, and conclude with the aid of (14.2) that (v_1, y_1, m_1) is admissible; hence, by (15.10), (v_1, y_1, m_1) would be a better program for our idealized missile than (v, y, m). More generally, if negative values of $v(t)$ could always be enclosed in a finite number of half-open subintervals of $[0, T]$, we could apply the above procedure to the leftmost, then to the next one to the right, etc. After a finite number of steps—say N—we could have an admissible triple (v_N, y_N, m_N) such that $v_N(t)$ is never negative and $m_N(t_1-) < m(t_1)$. In particular, and with convention (2.6), $m_N(0) < m(0)$. The remainder of this section is concerned with the difficult case in which there are infinitely many deleted intervals.

One verifies that no two such intervals have a common point, that indeed any two are separated by a positive distance. Since they are all subintervals of $[0, T]$, they are denumerable. Let $J_n = (t_{1n}, t_{2n}]$ be a fixed sequentialization of these intervals, $n = 1, 2, \cdots$. There is in general no leftmost J_n; if J_1 is deleted to obtain (v_1, y_1, m_1), then m_1 may not be monotone.

Let $E = \{t{:}0 \leqq t \leqq \hat{t},\ \ v(t) < 0\}$. Clearly $E \subset \bigcup J_n$. Let A_n, B_n, and B respectively denote the characteristic functions of J_n, $\bigcup_1^n J_m$, and $\bigcup_1^\infty J_m$. Then $B_n \to B$ as $n \to \infty$. Let

$$\tau_n(t) = \int_0^t [1 - B_n(s)]\, ds, \qquad \tau(t) = \int_0^t [1 - B(s)]\, ds.$$

Function τ is nondecreasing and absolutely continuous on $[0, T]$, hence, with $\bar{\tau}$ denoting $\tau(T)$ and with τ on the interval $[0, \bar{\tau}]$, the equation $\tau(t) = \tau$ holds either for a unique t or for all t on an interval. Let $t(\tau)$ be the single-valued inverse of $\tau(t)$ obtained by assigning $t(\tau)$ the leftmost solution of $\tau(t) = \tau$. Thus $t(\tau)$ increases with τ, has a possible countable set of discontinuities, and is everywhere left-continuous on $[0, \bar{\tau}]$. We observe that $\tau[t(\tau)] = \tau$, and $t[\tau(t)] \leqq t$, with equality holding in the latter if $t \in [0, T] - \bigcup J_n$. If $t \in \bigcup J_n$, then $t(\tau) = t$ has no solution.

Define

$$u(s) = v(t(s)). \tag{15.11}$$

If $t(s_1)$ is not the left endpoint t_{1n} of some J_n, $t(s)$ is continuous at s_1, and $u(s_1-) = v(t(s_1)-) \leqq v(t(s_1)+) = u(s_1+)$. If $t(s_1)$ is a left endpoint t_{1n} of a J_n, $t(s_1+) = t_{2n}$. Then, since $t(s)$ is left-continuous, $u(s_1-) = v(t(s_1)-) = v(t_{1n})$, and $v(t_{1n}) \leqq v(t_{2n}) = v(t(s_1+)) = u(s_1+)$.

In any case,

$$u(s_1-) \leqq u(s_1+). \tag{15.12}$$

Since v is, by §14, of bounded variation, so is u, which is therefore summable on $[0, \bar{\tau}]$.

Define

$$x(\tau) = \int_0^\tau u(s)\, ds. \tag{15.13}$$

Then by [17, §33.3, §35.3, §38.1],

$$\begin{aligned} x[\tau(t)] &= \int_0^t u[t(s)][1 - B(s)]\, ds \\ &= \int_0^t v(s)[1 - B(s)]\, ds. \end{aligned}$$

If $t \in [0, T] - \bigcup J_n$, then $x[\tau(t)] = y(t)$.

Set

$$F(\tau) = \frac{1}{c} \int_\tau^{\bar{\tau}} D[u(s), x(s)] \left[\exp \frac{u(s) + gs}{c}\right] [1 - B(s)]\, ds.$$

Then

$$F[\tau(t)] = \frac{1}{c}\int_t^T D[v(s),\ y(s)]\left[\exp\frac{v(s) + g\tau(s)}{c}\right][1 - B(s)]\,ds.$$

Define a mass-function μ by (2.3), namely by the relation

$$\mu(\tau)\exp\frac{u(\tau) + g\tau}{c} = M\exp\frac{V + g\bar{\tau}}{c} + F(\tau). \tag{15.14}$$

We can also define $t_n(\tau)$, $u_n(s)$, $x_n(\tau)$, $F_n(\tau)$, and $\mu_n(\tau)$ relative to $\tau_n(t)$ and $B_n(t)$ in exactly the same way as $t(\tau)$ through $\mu(\tau)$ are related above to $\tau(t)$ and $B(t)$. All statements above covering the latter hold for the corresponding expressions with index n.

However, function u has the property, not shared by u_n, that $u(\tau) \geqq 0$ for $0 \leqq \tau \leqq \hat{\tau} = \tau(\hat{t})$. For $t \notin \bigcup J_m$, $\tau_n(t) \to \tau(t)$ and $u_n[\tau_n(t)] = u[\tau(t)] = v(t)$.

Since v is bounded, the integrand of F_n is bounded uniformly in n and converges to that of F. By the Bounded Convergence Theorem, $\mu_n[\tau_n(t)] \to \mu[\tau(t)]$ as $n \to \infty$.

Now if $t \notin \bigcup_1^\infty J_n$, $\mu_1(\tau_1(t)) \leqq m(t)$, and in particular, $\mu_1(0) < m(0)$. Similar statements may be made of the relation of each (μ_n, τ_n), $n = 2, 3, \cdots$, to its predecessor. Hence, for $t \notin \bigcup J_n$

$$\mu(\tau(t)) \leqq m(t), \tag{15.15}$$

and

$$\mu(0) < m(0). \tag{15.16}$$

Though the $\mu_n(s)$ may not be nonincreasing in s, we proceed to show that $\mu(s)$ is.

If $\hat{\tau} \leqq t \leqq \bar{\tau}$, then $\mu(t) = m(t + T - \bar{\tau})$ and is nonincreasing on $[\hat{\tau}, \bar{\tau}]$ as a result of that property of m on the corresponding t-interval. If $0 \leqq \tau_1 < \hat{\tau}$, $\tau_1 < \tau_2 \leqq \bar{\tau}$ and $u(\tau_1) = 0$, let $\tau_3 = \min(\hat{\tau}, \tau_2)$. Then

$$\mu(\tau_1) = \mu(\tau_3)\exp\frac{u(\tau_3) + (\tau_3 - \tau_2)g}{c} + [F(\tau_1) - F(\tau_3)]\exp\frac{-g\tau_1}{c}.$$

Since $u(t) \geqq 0$, it follows that $F(\tau_1) \geqq F(\tau_3)$, hence that $\mu(\tau_1) \geqq \mu(\tau_3) \geqq \mu(\tau_2)$. With τ_1, τ_2 as last stated and $u(\tau_1) > 0$, let t_3 be the supremum of those t such that $t(\tau_1) < t \leqq T$ and such that if $t(\tau_1) \leqq t' < t$, then $v(t') > 0$. Set $t_1 = t(\tau_1)$, $t_2 = t(\tau_2)$. Now $t_3 = T$ or $v(t_3) = 0$ or both; hence $t_1 < t_3$, $[t_1, t_3] \cap (\bigcup J_n)$ is empty, and $v(t) > 0$ for all t on the interval $t_1 \leqq t < t_3$. If $t_2 \leqq t_3$ or $t_3 \geqq \hat{t}$ or both, then $\tau(t) - \tau(t_1) = t - t_1$ for $t_1 \leqq t \leqq t_2$. Moreover

$$\mu[\tau(t_1)]\exp\frac{v(t_1) + gt_1}{c} = \mu[\tau(t_2)]\exp\frac{v(t_2) + gt_2}{c}\,\{F[\tau(t_1)] - F[\tau(t_2)]\},$$

and

$$\mu[\tau(t_2)] \leqq m(t_2).$$

Moreover, if $t_2 \leqq t_3$, $v(t) > 0$ on $[t_1, t_2)$, and, just as we deduced (14.2), we find that $\mu(\tau_1) = \mu[\tau(t_1)] \geqq \mu[\tau(t_2)] = \mu(\tau_2)$. If, on the other hand, $t_2 > t_3$, then $v(t_3) = 0$. With $\tau_3 = \tau(t_3)$, we have already proved that $\mu(\tau_3) \geqq \mu(\tau_2)$, while $\mu(\tau_1) \geqq \mu(\tau_3)$ by the immediately preceding argument. We have thus shown in all cases that

$$\mu(\tau_1) \geqq \mu(\tau_2), \qquad 0 \leqq \tau_1 \leqq \tau_2 \leqq \bar{\tau}. \tag{15.17}$$

If $\hat{t}$ defined in the opening paragraph of this section is T, the discussion is complete. It remains to consider the case $\hat{t} < T$, in which case either $u(\tau) \equiv 0$ on $(\hat{\tau}, \bar{\tau})$ or $u(\tau)$ is positive for some τ and negative for some. The cases $u(\hat{t}-) \leqq V$ and $u(\hat{t}-) > V$ are handled separately.

In the first case let τ_1 be the supremum of those τ such that $\hat{\tau} \leqq \tau \leqq \bar{\tau}$ and $u(\tau) \leqq u(\hat{\tau}-)$. We have immediately that $\hat{\tau} \leqq \tau_1 \leqq \bar{\tau}$, that $u(\tau_1-) \leqq u(\hat{\tau}-)$, and that $\mu(\tau_1-) \leqq \mu(\hat{\tau}-)$. If $\tau_1 < \bar{\tau}$, $u(\tau)$ is nonnegative on $(\tau_1, \bar{\tau})$, and therefore $\mu(\tau_1-) \exp\,(u(\tau_1-)/c) > Me^{V/c}$. If, on the other hand, $\tau_1 = \bar{\tau}$, there exist τ_2, τ_3, with $\hat{\tau} \leqq \tau_2 < \tau_3 \leqq \tau_1$, such that $u(\tau_2-) \leqq 0$, and $u(\tau)$ is nonnegative on (τ_2, τ_3). Hence

$$\mu(\hat{\tau}-) \geqq \mu(\tau_2-) > \mu(\tau_3-) \geqq \mu(\tau_1-) = M \exp \frac{V - u(\tau_1-)}{c}.$$

In any event, we have, for this case

$$\mu(\hat{\tau}-) \exp \frac{u(\hat{\tau}-)}{c} > Me^{V/c}. \tag{15.18}$$

If we define $\tilde{u}(\tau) = u(\tau)$, $\tilde{x}(\tau) = x(\tau)$ for $0 \leqq \tau < \hat{\tau}$, and $\tilde{u}(\hat{\tau}) = V$, $\tilde{y}(\hat{\tau}) = Y$, $\tilde{\mu}(\hat{\tau}) = M$, and if $\tilde{\mu}(\tau)$ is then calculated from (2.3) for $0 \leqq \tau < \hat{\tau}$, the triple $(\tilde{u}, \tilde{x}, \tilde{\mu})$ is admissible and $\tilde{\mu}(0) < \mu(0)$.

In the second case $V < u(\hat{\tau}-)$. Then we cannot without contradiction have $\mu(\hat{\tau}-) \leqq M$. We attempt to replace all of the program for $\tau > \hat{\tau}$ and part of it for $\tau < \hat{\tau}$ with a coasting phase at mass M. There is a least τ_1 such that $u(\tau) > 0$ for τ on $(\tau_1, \hat{\tau})$, and on this interval we may use x as independent variable instead of τ, with x on $(x(\tau_1), Y]$. Let $w = w(x)$, $x \leqq Y$, be the velocity for a coasting trajectory, at mass M, which terminates at $x = Y$ with velocity V. We seek a pair x_2, τ_2 with $x(\tau_1) \leqq x_2 < Y$, $0 \leqq \tau_2 < \tau$, such that $x_2 = x(\tau_2)$ and $u(\tau_2-) \leqq w(x_2)$. Such a pair always exists. Take x_2, τ_2 to be the pair of largest such values. Set $s(x) = \int_{x_2}^{x} dz/w(z)$, and $s(Y) = S$. Therefore on $[0, S]$ the function s can be inverted to give $x = s^{-1}(s)$. Let

$$\tilde{u}(\tau) = u(\tau), \quad \text{on} \quad 0 \leqq \tau < \tau_2\,;$$

$$\tilde{u}(\tau) = w(s^{-1}(\tau - \tau_2)), \quad \text{on} \quad \tau_2 \leqq \tau \leqq \tau_2 + S.$$

Then define

$$\tilde{x}(\tau) = \int_0^\tau \tilde{u}(s)\, ds,$$

and $\tilde{\mu}(\tau)$ by (2.3), as usual, taking $\tilde{\mu}(\tau_2 + S) = M$. By now familiar methods it may be shown that $(\tilde{u}, \tilde{x}, \tilde{\mu})$ is admissible and that $\tilde{\mu}(0) < \mu(0)$.

We have now completed the proof that to any admissible program with its velocity function v anywhere negative there corresponds at least one other admissible program having nonnegative v and a smaller initial mass. In seeking a minimizing program we may confine our attention to programs with nonnegative v.

16. Admissible trajectories are on one side of Γ. In this section we again impose the conditions on drag D stated at the beginning of §3. It is convenient to use axes in the (t, y) plane which are oppositely directed to those used heretofore and with the origin at the point called (T, Y) in §2 or $(0, Y)$ in §5. The trajectory Γ of §5 now issues from the origin into the first quadrant and is convex. The region R of §8 now lies to the right of Γ and between the lines $y = 0$ and $y = Y$. Properties (2.1) through (2.4) of admissible programs now apply as stated with the one exception that g in (2.3) is replaced by $-g$. For $t_1 = 0$, $t_2 = t$, we have from (2.3) that

$$\text{(16.1)} \qquad m(t) \exp \frac{v(t) - gt}{c} = M \exp \frac{V}{c} + \frac{1}{c} \int_0^t D(v, y) \exp \frac{v(s) - gs}{c}\, ds.$$

The boundary conditions (2.5) are now

$$\text{(16.2)} \qquad v(0) = V, \quad y(0) = 0, \quad m(0) = M, \quad v(T) = 0, \quad y(T) = Y.$$

As a consequence of §15, we can restrict attention to admissible triples (v, y, m) such that $v(t) \geqq 0$ on $[0, T]$.

The thrust-free trajectory of §5 now satisfies the equations

$$\text{(16.3)} \qquad \dot{u} = g + D(u, x)/M, \qquad \dot{x} = u,$$

and the initial conditions

$$\text{(16.4)} \qquad x(0) = 0, \qquad u(0) = V \geqq 0.$$

System (16.3), (16.4) has a solution x, u on a maximal interval $[0, t_1)$, and either this interval includes a value T_0 such that $x(T_0) = Y$ or there is a value $y_1 \leqq Y$ such that $x(t) \rightarrow y_1$ and $u(t) \rightarrow \infty$ as $t \rightarrow t_1$.

Given an admissible triple (v, y, m), suppose there is a value t_2 in $[0, T)$ such that $y(t_2) > x(t_2)$. Clearly $t_2 \neq 0$ and, since x and y are continuous, there is a $t' < t_2$ with $y(t) > x(t)$ on (t', t_2). If t_3 is the infimum of such t', then $y(t_3) = x(t_3)$ and $v(t_3+) \geqq u(t_3)$. Set

$$\eta(t) = v(t+) - u(t). \tag{16.5}$$

Case 1. $v(t_3+) = u(t_3)$. Since $y(t) - x(t) > 0$ for t in (t_3, t_2) and arbitrarily near t_3, there exists $t_4 \in (t_3, t_2)$ and arbitrarily near t_3 such that $v(t_4) > u(t_4)$. With t_4 having these properties fixed, define t_5 as the infimum of those t satisfying the relations $t_3 \leqq t \leqq t_4$ and $\eta(t) \geqq \eta(t_4)$. Clearly $t_3 < t_5$, while $\eta(t_5) = \eta(t_4)$ as a result of the fact that η can have only negative jumps; therefore $t_3 < t_5 \leqq t_4$.

Observe that

$$\eta(t_5) < c\left[\exp\frac{v(t_5+)}{c} - \exp\frac{u(t_5)}{c}\right],$$

and that, by (16.1), the right member can be expressed in the form

$$\exp\frac{gt_5}{c}\sum_1^4 F_i(t_3, t_5), \tag{16.6}$$

where

$$F_1(t_3, t_5) = -c\left[1 - \frac{m(t_3)}{m(t_5)}\right]\exp\frac{u(t_3) - gt_3}{c},$$

$$F_2(t_3, t_5) = -\frac{m(t_5) - M}{Mm(t_5)}\int_{t_3}^{t_5} D(v, y)\exp\frac{v - gt}{c}\,dt,$$

$$F_3(t_3, t_5) = \frac{1}{M}\int_{t_3}^{t_5}[D(v, y) - D(v, x)]\exp\frac{v - gt}{c}\,dt,$$

$$F_4(t_3, t_5) = \frac{1}{M}\int_{t_3}^{t_5}[D(v, x)e^{v/c} - D(u, x)e^{u/c}]e^{-gt/c}\,dt.$$

Each of the first two terms in (16.6) is at most zero. On $[t_3, t_5]$, $y(t) - x(t) < \eta(t_5)(t_5 - t_3)$, hence the third term is below $k_2(t_5 - t_3)^2\eta(t_5)/M$, where k_2 denotes the product of $\exp \hat{V}/c$, in which $\hat{V}$ is an upper bound for $v(t)$, times the supremum of $[D(v, y) - D(v, x)]/(y - x)$ on the class of all real values of x, y, v such that $0 \leqq x < y \leqq Y$ and $0 \leqq v \leqq \hat{V}$. Similarly the fourth term is dominated by $k_3(t_5 - t_3)\eta(t_5)/M$, where k_3 is the supremum of $[D(v, x)e^{v/c} - D(u, x)e^{u/c}]/(v - u)$ on the class of real triples x, u, v satisfying the conditions $0 \leqq x \leqq Y, u \neq v$, and $0 \leqq u, v \leqq \max[\hat{V}, u(t_2)]$.

It follows that

$$e^{-gt_5/c}\eta(t_5) < (Tk_1 + k_2)(t_5 - t_3)\eta(t_5)/M. \tag{16.7}$$

Now k_1 and k_2 do not involve t. Recall that t_4 is arbitrarily near t_3 and $t_3 < t_5 \leqq t_4$. We are therefore free to suppose that t_4 has been so chosen that $(Tk_1 + k_2)(t_5 - t_3)/M < e^{-gt/c}$. Relation (16.7) is then a contradiction and we infer that there can exist no t_2 in the half-open interval $[0, T)$ satisfying the relation $y(t_2) > x(t_2)$.

Case 2. $v(t_3+) > u(t_3)$. A similar argument leads to a similar contradiction.

17. Concluding comments. It may be of interest to record that the authors, first singly and more recently in conjunction, have had troubles with one corner or another of this problem over a period of years. It has not been possible as yet to find the solution if restriction (4.4) is essentially relaxed, or to solve the problem if $v(0)$ is specified to be greater than that of the minimizing solution of the present case. An encompassing theory of global extrema for the class of problems mentioned in the introduction would clearly be desirable but this appears to be well beyond reach.

If one wishes to place a bound on the rate of mass-flow and yet to admit the largest class of programs (v, y, m) with this restriction, simply add a Lipschitz condition on m to our (2.1) through (2.5). The limit function m_0 in our existence theorem then necessarily satisfies the Lipschitz condition and we have an existence theorem as a corollary to §3. Characterization of the minimizing triples (v_0, y_0, m_0) among all those now admitted will of course involve considerable work. Triples (v, y, m) are now all AC but not in general piecewise smooth. [4] and [5] suggest important parts but by no means all of a solution of this characterization problem.

REFERENCES

[1] R. H. Goddard, *A method of reaching extreme altitudes*, Smithsonian Misc. Collections, (1919), and reprint by Amer. Rocket Soc., 1946.

[2] Richard Bellman, ed., *Mathematical Optimization Techniques*, University of California Press, Berkeley, 1963.

[3] G. Leitmann, ed., *Optimization Techniques with Applications to Aerospace Systems*, Academic Press, New York, 1963.

[4] A. Miele, *Generalized variational approach to the optimum thrust programming for the vertical flight of a rocket*, Part I, Z. Flugwiss., 6 (1958), pp. 69–77.

[5] A. Miele and C. R. Cavoti, *Generalized variational approach to the optimum thrust programming for the vertical flight of a rocket*, Part II, Ibid., 6 (1958), pp. 102–109.

[6] G. A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[7] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience, Wiley, New York, 1962.

[8] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[9] G. Hamel, *Über eine mit dem Problem der Rakete zusammenhängende Aufgabe der Variationsrechnung*, Angew. Math. Mech., 7 (1927), pp. 451–452.

[10] H. S. Tsien and R. C. Evans, *Optimum thrust programming for a sounding rocket*, J. Amer. Rocket Soc., 21 (1951), pp. 97–107.

[11] G. Leitmann, *The problem of optimum thrust programming*, NOTS Technical Note 5038-11, 1955.

[12] G. Leitmann, *Solution of Goddard's problem*, Astronaut. Acta, 2 (1956), pp. 55–62.

[13] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, 1941.

[14] G. A. Bliss, *Calculus of Variations*, Open Court, Chicago, 1925.

[15] T. Radó, *Length and Area*, Amer. Math. Soc. Colloquium Publications, 30 (1948).

[16] O. Bolza, *Lectures on the Calculus of Variations*, University of Chicago Press, 1904; and reprints by Stechert, New York, 1931; and recently by Chelsea, New York.

[17] E. J. McShane, *Integration*, Princeton University Press, Princeton, 1944.

J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1964

OPTIMAL PURSUIT STRATEGY PROCESSES WITH RETARDED CONTROL SYSTEMS*

M. NAMİK OĞUZTÖRELI†

Summary. Recently D. L. Kelendzheridze [4, 9] investigated an optimal pursuit problem for systems described by ordinary differential equations. We present here an extension of his results to systems described by linear differential-difference equations with retarded argument, the control functions and the initial conditions being allowed to vary in given closed compact and convex sets. We also establish here generalizations of some of the results of J. P. LaSalle [5], L. W. Neustadt [6] and the author [7].

1. Introduction. We consider two control systems X and Z, given, in the n-dimensional phase-space, by linear differential-difference equations with retarded argument of the form

$$x'(t + c_m) + \sum_{i=0}^{m} A_i(t)x(t + c_i) = A(t)u(t), \tag{1.1}$$

and

$$z'(t + d_k) + \sum_{j=0}^{k} B_j(t)z(t + d_j) = B(t)v(t), \tag{1.2}$$

where t is a real variable (time), $' = \frac{d}{dt}$, c_i and $d_j (i = 0, 1, \cdots, m;$ $j = 0, 1, \cdots, k)$ are given constants such that

$$0 = c_0 < c_1 < \cdots < c_m \quad \text{and} \quad 0 = d_0 < d_1 < \cdots < d_k, \tag{1.3}$$

$A_i(t)$ and $B_j(t) (i = 0, 1, \cdots, m; j = 0, 1, \cdots, k)$ are given $n \times n$ continuous matrix functions, $A(t)$ is a given continuous $n \times r$ matrix function, $B(t)$ is a given continuous $n \times s$ matrix function, $x(t)$ and $z(t)$ are n-dimensional vectors which describe the states of the control systems X and Z, respectively, at time t, $u(t)$ is an r-dimensional vector function controlling the motion of the system X and $v(t)$ is an s-dimensional vector function controlling the motion of system Z. The components of u and v will be denoted by $u_1, \cdots, u_r$ and $v_1, \cdots, v_s$, respectively.

Let U be a set of r-dimensional vector functions $u(t)$ piecewise continuous on each finite interval $[t_0, t]$ and V be a set of s-dimensional vector functions $v(t)$ piecewise continuous on each finite interval $[t_0^*, t]$; U and V are the "control regions" for the systems X and Z respectively. We shall

* Received by the editors October 1, 1963 and in revised form December 31, 1963.

† Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia.

suppose the U and V are closed, compact, bounded, convex and contain the origin. Vector functions $u(t)$, defined in U, and vector functions $v(t)$, defined in V, will be called *admissible control functions* for the systems X and Z respectively.

Let Φ be a closed, compact, bounded and convex subset of the set of all real-valued n-dimensional vector functions $\phi(t)$, continuous in the *initial interval* $t_0 \leqq t \leqq t_0 + c_m$ and having the property

$$\phi(t_0) = x_0 , \tag{1.4}$$

where x_0 is given. The elements of the set Φ will be called *admissible initial conditions for the system* X.

Similarly, we shall denote by Ψ a compact, closed, bounded and convex subset of all real-valued n-dimensional vector functions $\psi(t)$, continuous in the *initial interval* $t_0^* \leqq t \leqq t_0^* + d_k$ and having the property

$$\psi(t_0) = z_0 , \tag{1.5}$$

where z_0 is given. Functions $\psi(t)$ which belong to the set Ψ will be called *admissible initial conditions for the system* Z.

A solution $x(t)$ of the system (1.1) which satisfies the *initial condition*

$$x(t) = \phi(t), \qquad t_0 \leqq t \leqq t_0 + c_m , \qquad \phi \in \Phi, \tag{1.6}$$

obviously depends on the choice of functions $u(t)$ and $\phi(t)$. To indicate this relationship explicitly we shall denote by $x(t, \phi, u)$ the solution of (1.1) satisfying the initial condition (1.6) with the selected control function $u = u(t)$. It is well known [1] that there is a unique continuous solution of (1.1) for $t \geqq t_0$ which satisfies the initial condition (1.6).

A continuous solution $z(t)$ of the system (1.2), with the selected control function $v = v(t)$, which satisfies the initial condition

$$z(t) = \psi(t), \qquad t_0^* \leqq t \leqq t_0^* + d_k , \qquad \psi \in \Psi, \tag{1.7}$$

will be denoted, as above, by $z(t, \psi, v)$.

The system X will be called the *pursuing system* and the system Z the *pursued system*.

For an arbitrary admissible control $v(t)$ and arbitrary admissible initial condition $\psi(t)$ let us assume that there exists an admissible pair $u(t)$ and $\phi(t)$ such that the trajectories $x(t, \phi, u)$ and $z(t, \psi, v)$ of (1.1) and (1.2) corresponding to the controls u, v and initial conditions ϕ, ψ, respectively, satisfy the equation

$$x(T, \phi, u) = z(T, \psi, v) \tag{1.8}$$

for some

$$T > \max \, [t_0 + c_m , \quad t_0^* + d_k] \tag{1.9}$$

and

$$x(t, \phi, u) \neq z(t, \psi, v), \qquad t < T. \tag{1.10}$$

The quantity T depends on the chosen controls $u(t)$ and $v(t)$ and the chosen initial conditions $\phi(t)$ and $\psi(t)$; therefore we may write $T = T(u, \phi; v, \psi)$. This time T will be called the *pursuit time.*

If an admissible pair $v(t)$ and $\psi(t)$ for the pursued system Z is chosen, the pursuing system X should be controlled in such a manner that the corresponding pursuit time $T(u, \phi; v, \psi)$ will assume its minimal value. Denote it by

$$T_{v,\psi} = \min_{u \in U, \phi \in \Phi} T(u, \phi; v, \psi). \tag{1.11}$$

The system Z should choose an admissible pair $v(t)$, $\psi(t)$ which maximizes the quantity $T_{v,\psi}$. This maximum will be denoted by

$$T^0 = \max_{v \in V, \psi \in \Psi} \min_{u \in U, \phi \in \Phi} T(u, \phi; v, \psi). \tag{1.12}$$

In the present paper, we wish to investigate the following optimization problem.

Find the admissible controls $u(t) \in U$, $v(t) \in V$ and the admissible initial conditions $\phi \in \Phi$, $\psi \in \Psi$ for which the corresponding pursuit time $T(u, \phi; v, \psi)$ satisfies

$$T(u, \phi; v, \psi) = T^0. \tag{1.13}$$

The above problem for systems X and Z involving no time delay has recently been considered by Kelendzheridze [4]. His main objective is Pontryagin's maximum principle. We shall follow here a method which is a synthesis of that used by Kelendzheridze and another developed by LaSalle [5] and Neustadt [6]. This method has been used recently by the author [7] for a time optimal control problem with time delay.

We shall generally assume, as mentioned above, that the sets U and V are bounded, closed, compact, convex and contain the origin as an interior point. Particularly, we shall consider the case in which U and V consist of piecewise continuous vector functions such that

$$\begin{aligned} U&: \quad \{|u_i(t)| \leqq 1, \qquad i = 1, 2, \cdots, r\}, \\ V&: \quad \{|v_j(t)| \leqq 1, \qquad j = 1, 2, \cdots, s\}. \end{aligned} \tag{1.14}$$

We shall generally suppose that the sets Φ and Ψ of continuous initial conditions are closed, compact, convex and contain the origin as an interior point. Only in §5 shall we consider a special case in which the sets Φ and Ψ will consist of piecewise continuous functions such that

$$\begin{aligned} \Phi&: \quad \{|\phi_i(t)| \leqq 1, \ t_0 \leqq t \leqq t_0 + c_m, \qquad i = 1, 2, \cdots, n\}, \\ \Psi&: \quad \{|\psi_i(t)| \leqq 1, \ t_0^* \leqq t \leqq t_0^* + d_k, \qquad i = 1, 2, \cdots, n\}. \end{aligned} \tag{1.15}$$

Note that, if the retardations c_i and d_j in (1.1) and (1.2) all approach zero, our optimization problem reduces to the problem which is considered by Kelendzheridze.

2. The functionals $\Omega^*(t, \phi, u)$ and $\Theta^*(t, \psi, v)$. Let $V(s, t)$ and $W(s, t)$ be Bellman-Cooke kernel matrices [1] of the systems (1.1) and (1.2) respectively, and $H(s, t)$ and $K(s, t)$ be the kernel matrices, introduced by the author in [7], of the homogeneous systems

$$x'(t + c_m) + \sum_{i=0}^{m} A_i(t)x(t + c_i) = 0, \tag{2.1}$$

and

$$z'(t + d_k) + \sum_{j=0}^{k} B_j(t)z(t + d_j) = 0, \tag{2.2}$$

which correspond to (1.1) and (1.2) respectively.

Consider now the functionals

$$\Omega(t, \phi, u) = \int_a^b H(s, t)\, \phi(s)\, ds + \int_a^t V(s, t)\, A(s)\, u(s)\, ds \tag{2.3}$$

and

$$\Theta(t, \psi, v) = \int_c^d K(s, t)\, \psi(s)\, ds + \int_c^t W(s, t)\, B(s)\, v(s)\, ds, \tag{2.4}$$

where

$$a = t_0, \qquad b = t_0 + c_m, \qquad c = t_0^*, \qquad d = t_0^* + d_k. \tag{2.5}$$

As is shown in [7], we have the following representations of the solutions $x(t, \phi, u)$ and $z(t, \psi, v)$.

$$x(t, \phi, u) = \Omega^*(t, \phi, u), \qquad z(t, \psi, v) = \Theta^*(t, \psi, v), \tag{2.6}$$

where

$$\Omega^*(t, \phi, u) = \Omega(t - c_m, \phi, u), \qquad \Theta^*(t) = \Theta(t - d_k, \psi, v). \tag{2.7}$$

From (1.8) and (2.6) we can write

$$\Omega^*(T, \phi, u) = \Theta^*(T, \psi, v), \tag{2.8}$$

where T is the pursuit time, defined in §1. Obviously, $T = T(u, \phi; v, \psi)$. By its definition, T is single-valued.

Consider now the sets

$$C(t) = \{\Omega^*(t, \phi, u); \quad \phi \in \Phi, \quad u \in U\}, \tag{2.9}$$

and

$$E(t) = \{\Theta^*(t, \psi, v); \quad \psi \in \Psi, \quad v \in V\}. \tag{2.10}$$

In the remainder of the discussion, we shall need the following properties of the sets $C(t)$ and $E(t)$, proved in [7].

(I) *$C(t)$ and $E(t)$ are compact and convex.*

(II) *If $\tilde{U}$ and $\tilde{V}$ are the sets of all bang-bang control functions, and if U and V are defined by* (1.14), *then* $C(t) = \{\Omega^*(t, \phi, \tilde{u}), \phi \in \Phi, \tilde{u} \in \tilde{U}\}$, $E(t) = \{\Theta^*(t, \psi, \tilde{v}), \psi \in \Psi, \tilde{v} \in \tilde{V}\}$.

(III) *If Ω is an interior point of $C(t)$, then there exists an $\epsilon > 0$ such that $N_\epsilon(\Omega) \subset C(\tau)$ for all τ in $(t - \epsilon, t]$, where $N_\epsilon(\Omega)$ is a neighborhood of Ω of radius ϵ.*

3. Existence of optimal strategies.

We shall prove now the following existence theorem, which is an extension of that due to Kelendzheridze [4, 9].

THEOREM 1. *If for an arbitrary admissible pair $v(t)$ and $\psi(t)$ there exists an admissible pair $u(t)$ and $\phi(t)$, such that $x(t, \phi, u) = z(t, \psi, v)$, then there exist two pairs of functions $u^0 \in U$, $\phi^0 \in \Phi$ and $v^0 \in V$, $\psi^0 \in \Psi$, which are optimals, that is,*

$$T^0 = T(u^0, \phi^0; v^0, \psi^0),$$

where T^0 is defined by (1.2).

Proof. By hypothesis the set

$$\Gamma = \{T, x(T, \phi, u) = z(T, \psi, v); \qquad u \in U, \quad v \in V, \quad \phi \in \Phi, \quad \psi \in \Psi\} \tag{3.1}$$

is not empty.

Let us choose arbitrarily an admissible pair $v^*(t) \in V$, $\psi^*(t) \in \Psi$, and consider the following subset of Γ.

$$\Gamma^* = \{T^* = T(u, \phi; v^*, \psi^*); \qquad u \in U, \phi \in \Phi, T \in \Gamma\}. \tag{3.2}$$

By hypothesis Γ^* is not empty.

Let T_{v^*,ψ^*} be the greatest lower bound of all $T^* \in \Gamma^*$.

$$T_{v^*,\psi^*} = \inf_{u \in U, \phi \in \Phi} T(u, \phi; v^*, \psi^*). \tag{3.3}$$

By definition, we have $\Omega^*(T^*, \phi, u) = \Theta^*(T^*, \psi^*, v^*) \in C(T^*)$, $T^* \in \Gamma^*$.

Let the sequence $T_i^* \in \Gamma^*$, $i = 1, 2, \cdots$, be selected so that

$$\lim_{i \to \infty} T_i^* = T_{v^*,\psi^*}.$$

Consider now the sequence $\{\Omega^*(T_i^*, \phi^i, u^i)\}$, $\{\Omega^*(T_{v^*,\psi^*}, \phi^i, u^i)\}$, $i = 1, 2, \cdots$, where $\phi^i = \phi^i(t)$ and $u^i = u^i(t)$ are admissible. As shown in [7], for $T_i^* - T_{v^*,\psi^*} < \delta$, we have $\|\Omega^*(T_i^*, \phi^i, u^i) - \Omega^*(T_{v^*,\psi^*}, \phi^i, u^i)\| < \epsilon$, where ϵ is an arbitrarily small positive number and

$$\delta = \min\left\{\frac{\epsilon}{3m_1 m_2}, \qquad \frac{\epsilon}{3c_m m_1 m_3}\right\},$$

where

$$m_1 = \max_{a \leqq s \leqq t} \|V(s, t)A(s)\|, \qquad m_2 = \sup_{u \in U, t \geqq t_0} \|u\|, \qquad m_3 = \sup_{\phi \in \Phi} \|\phi\|,$$

the norm being defined as follows.

$$(3.4) \quad \|\alpha\| = \begin{cases} \max\{|\alpha_i|\}, \quad i = 1, \cdots, \lambda, \text{ if } \alpha \text{ is a } \lambda\text{-vector with} \\ \quad \text{components } \alpha_i, \\ \max\{|\alpha_{ij}|\}, \quad i = 1, \cdots, \mu; j = 1, \cdots, \nu; \text{ if } \alpha \text{ is a} \\ \quad \mu \times \nu \text{ matrix with elements } \alpha_{ij}. \end{cases}$$

In the definitions of m_1 and m_2 we suppose that t is sufficiently large.

Since the set $C(T_{v^*,\psi^*})$ is closed and compact, we can extract subsequences

$$\{\phi^{i_k}(t)\}, \qquad \{u^{i_k}(t)\}, \qquad k = 1, 2, \cdots,$$

from the sequences $\{\phi^i(t)\}$ and $\{u^i(t)\}$ so that they converge to the functions $\phi^0(t) \in \Phi$ and $u^0(t) \in U$ respectively. Therefore

$$(3.5) \qquad \Theta^*(T_{v^*,\psi^*}, \psi^*, v^*) = \Omega^*(T_{v^*,\psi^*}, \phi^0, u^0) \in C(T_{v^*,\psi^*}),$$

where T_{v^*,ψ^*} is defined by (3.3), $v^* \in V$, $\psi^* \in \Psi$ being selected arbitrarily so that $T^* = T(u, \phi; v^*, \psi^*) \in \Gamma^*$. Obviously $T_{v^*,\psi^*} = T(u^0, \phi^0; v^*, \psi^*)$.

Consider now the set

$$(3.6) \qquad \Gamma^0 = \{T_{v,\psi}; \quad v \in V, \quad \psi \in \Psi\}.$$

Let T^0 be the least upper bound of all $T_{v,\psi} \in \Gamma^0$.

$$(3.7) \qquad T^0 = \sup_{v \in V, \psi \in \Psi} T(u^0, \phi^0; v, \psi) = \sup_{v \in V, \psi \in \Psi} T_{v,\psi},$$

which is equal to T^0 defined by (1.12). This optimal time T^0 may be finite or infinite according as the set Γ^0 is bounded or unbounded.

Consider first the case in which T^0 is finite. Let the sequence $T_j^0 \in \Gamma^0$, $j = 1, 2, \cdots$, be selected so that

$$\lim_{j \to \infty} T_j^0 = T^0.$$

Consider the sequence $\{\Theta^*(T_j^0, \psi^j, v^j)\}$, $\{\Theta^*(T^0, \psi^j, v^j)\}$, $j = 1, 2, \cdots$, where $v^j = v^j(t)$ and $\psi^j = \psi^j(t)$ are admissible. If

$$m_4 = \max_{c \leqq s \leqq t} \|W(s, t)B(s)\|, \; m_5 = \sup_{v \in V, t \geqq c} \|v\|, \; m_6 = \sup_{\psi \in \Psi} \|\psi\|,$$

t sufficiently large, we may easily show that, for

$$T^0 - T_j^0 < \min\left\{\frac{\epsilon}{3m_4 m_5}, \frac{\epsilon}{3d_k m_5 m_6}\right\},$$

we have $\|\Theta^*(T^0, \psi^j, v^j) - \Theta^*(T_j^0, \psi^j, v^j)\| < \epsilon$, where ϵ is an arbitrarily small positive number. Since the set $E(T^0)$ is compact and closed, we can extract subsequences $\{\psi^{j_k}(t)\}$, $\{v^{j_k}(t)\}$, $k = 1, 2, \cdots$, from the sequences $\{\psi^j(t)\}$ and $\{v^j(t)\}$ so that they converge to the functions $\psi^0(t) \in \Psi$ and $v^0(t) \in V$ respectively. Therefore,

$$(3.8) \qquad T^0 = T(u^0, \phi^0; v^0, \psi^0),$$

and

$$(3.9) \qquad \Omega^*(T^0, \phi^0, u^0) = \Theta^*(T^0, \psi^0, v^0) \in E(T^0).$$

Consider now the case $T^0 = \infty$. Let the sequence $T_k^0 \in \Gamma^0$, $k = 1, 2, \cdots$, be so chosen that

$$T_k^0 \xrightarrow[k\to\infty]{} \infty \quad \text{monotonically.}$$

Since $T_k^0 \in \Gamma^0$, we have

$$(3.10) \qquad \Theta^*(T_k^0, v, \psi) = \Omega^*(T_k^0, \phi^0, u^0), \qquad k = 1, 2, \cdots,$$

for some $v \in V$, $\psi \in \Psi$, for each k. Let us select an admissible pair $v^k(t)$, $\psi^k(t)$ which satisfies (3.10). Therefore

$$(3.11) \qquad T_k^0 = T(u^0, \phi^0; v^k, \psi^k), \qquad k = 1, 2, \cdots.$$

Consider the sequences $\{\psi^k(t)\}$ and $\{v^k(t)\}$, $k = 1, 2, \cdots$. Since $v^k(t) \in V$, $\psi^k(t) \in \Psi$ and since the sets V and Ψ are compact and closed, we can extract subsequences $\{\psi^{k_r}(t)\}$ and $\{v^{k_r}(t)\}$, $r = 1, 2, \cdots$, from the sets $\{\psi^k(t)\}$ and $\{v^k(t)\}$ so that they converge to the functions $\psi^0(t) \in \Psi$, $v^0(t) \in V$, respectively. Therefore $T(u^0, \phi^0; v^0, \psi^0) = +\infty$.

Remark. Let P be the topological space of points $p = (u, \phi; v, \psi)$, where $u \in U$, $\phi \in \Phi$, $v \in V$, $\psi \in \Psi$ with the metric defined by

$$\rho(p_1, p_2) = \begin{cases} \sup\limits_{t \geqq a} \|u_1 - u_2\|, \ \sup\limits_{a \leqq t \leqq b} \|\phi_1 - \phi_2\|, \\ \sup\limits_{t \geqq c} \|v_1 - v_2\|, \ \sup\limits_{c \leqq t \leqq d} \|\psi_1 - \psi_2\|. \end{cases}$$

Consider now the function $f(t, p) = \|\Omega^*(t, \phi, u) - \Theta^*(t, \psi, v)\|$, $t \geqq e$, where $e = \max [b, d]$. It is shown in [7] that

(i) $t \to f(t, p)$ is continuous for $t \geqq e$ and for fixed $p \in P$;

(ii) $p \to f(t, p)$ is continuous for $p \in P$ and for fixed $t \geqq e$.

Let $p_0 \in P$ and consider the equation $f(t, p_0) = 0$. Let T_0 be the greatest lower bound of all the solutions of this equation.

Therefore, $f(T_0, p_0) = 0$, and $f(t, p_0) \neq 0$ for $t < T_0$, $T^0 \geqq e$. Suppose that

(iii) $f'(t, p) \geqq c(>0)$ for $|t - T_0| < \delta_0$, $p \in N_0$, where $\delta_0 > 0$ and N_0 is some neighborhood of p_0;

$$\text{(iv)} \qquad \lim_{p \to p_0} \sup_{t \geqq e} |f(t, p) - f(t, p_0)| = 0.$$

Suppose now that p is close to p_0 and consider the solution of $f(t, p) = 0$. For small h we have $f(T_0 + h, p) = f(T_0, p) + hf'(T_0 + \theta h, p)$, $(0 \leqq \theta \leqq 1)$. By (ii) we can choose a neighborhood $N_1 \subset N$ of p_0 such that $0 \leqq f(T_0, p) < \frac{1}{2}\delta_1 c$ for $p \in N_1$. If $p \in N_1$ and $0 \leqq h < \delta_1$, then from (iii) we can get $f(T_0 + h, p) \geqq ch > 0$. Similarly, if $-\delta_1 < h \leqq 0$, then $f(T_0 + h, p) \leqq \frac{1}{2}\delta_1 c + ch$, which is negative if $h = -\frac{3}{4}\delta_1$. It follows from (i) that there exists a solution t of $f(t, p) = 0$ satisfying $| t - T_0 | \leqq \frac{3}{4}\delta_1$. In fact, the argument shows that a solution t exists satisfying

$$| t - T_0 | \leqq \text{const.}\, f(T_0, p),$$

which tends to zero as $p \to p_0$, by (ii).

If T_p is the lower bound of such solutions we have therefore $T_p \leqq T_0 + \eta(p)$, where $\eta(p) \to 0$ as $p \to p_0$. On the other hand, by (i), given any $\xi > 0$, there exists an $A(\xi) > 0$ such that $f(t, p_0) \geqq A(\xi)$ for $0 \leqq t \leqq T_0 - \xi$. So, by (iv), we have for some neighborhood $N_\xi \subset N_0$ of p_0,

$$f(t, p) \geqq \tfrac{1}{2}A(\xi) > 0 \quad \text{for} \quad 0 \leqq t \leqq T_0 - \xi,\ p \in N_\xi.$$

It follows that $T_p > T_0 - \xi$ if $p \in N_\xi$. So, finally, $| T_p - T_0 | \to 0$ as $p \to p_0$. (Note that the case $f'(t, p) \leqq c\ (< 0)$ for $| t - T_0 | < \delta_0$, $p \in N_0$, can be treated in a similar way.)

We see that under the hypotheses (iii) and (iv) the functional $T(u, \phi; v, \psi)$ is continuous in all its arguments. We shall assume the hypotheses (iii) and (iv) in the following sections.

4. Properties of optimal strategies. In all the theorems which we shall prove in this and the next section we shall assume (without specifying it each time) that the pursuit time T^0 is finite and that the convex and compact set $C(T^0)$ has interior points. The latter can be proved under suitable hypotheses. (Kelendzheridze [9] made use of slightly different assumptions to prove this fact.) Let Λ be the *capture point* at which the system X (with the admissible pair $\{u(t), \phi(t)\}$) encounters the system Z (with the admissible pair $\{v(t), \psi(t)\}$) at time $t = T(u, \phi; v, \psi)$. The point Λ depends upon the choice of the functions u, ϕ, v, and ψ. Let Λ^0 be the *optimal capture point* which corresponds to the optimal strategy $u = u^0(t)$, $\phi = \phi^0(t)$, $v = v^0(t)$, and $\psi = \psi^0(t)$ and to the time $t = T^0$. Hence

$$\Lambda = \Omega^*(T, \phi, u) = \Theta^*(T, \psi, v) \quad \text{and} \tag{4.1}$$
$$\Lambda^0 = \Omega^*(T^0, \phi^0, u^0) = \Theta^*(T^0, \psi^0, v^0).$$

If $v(t)$ and $\psi(t)$, selected arbitrarily from the sets V and Ψ respectively, are kept fixed, the corresponding optimal policy of the system X will be described by $u = u^0(t)$ and $\phi = \phi^0(t)$ and the capture will occur at $t = T_{v\psi}$. Therefore,

$$\Omega^*(T_{v\psi}, \phi^0, u^0) = \Theta^*(T_{v\psi}, \psi, v), \tag{4.2}$$

and

$$(4.3)\qquad \Omega^*(t, \phi^0, u^0) \neq \Theta^*(t, \psi, v), \quad \text{for} \quad t < T_{v\psi},$$

where

$$(4.4)\qquad T_{v\psi} = \min_{u \in U, \phi \in \Phi} T(u, \phi; v, \psi) = T(u^0, \phi^0; v, \psi).$$

As shown in [7], the point $\Lambda_{v\psi} = \Omega^*(T_{v\psi}, \phi^0, u^0)$, which is the point Λ for $u = u^0(t)$, $\phi = \phi^0(t)$, $v = v(t)$ and $\psi = \psi(t)$, is a boundary point of the set $C(T_{v\psi})$ and there exists a unit vector $\eta = (\eta_1, \cdots, \eta_n)$ of the n-dimensional Euclidean space R^n such that

$$(4.5)\qquad \eta\cdot\Omega^*(T_{v\psi}, \phi, u) \leqq \eta\cdot\Omega^*(T_{v\psi}, \phi^0, u^0)$$

for all $\Omega^*(T_{v\psi}, \phi, u) \in C(T_{v\psi})$. Clearly, the vector η depends upon the choice of the vector functions $v(t) \in V$ and $\psi(t) \in \Psi$.

By the remark at the end of §3, the functional $T_{v\psi}$ varies continuously when the functions $v(t)$ and $\psi(t)$ vary continuously in V and Ψ respectively. Since $\Omega^*(t, \phi^0, u^0)$ is continuous in t (see [7]), if the admissible pair $\{v(t), \psi(t)\}$ varies continuously, the point $\Lambda_{v\psi} = \Omega^*(T_{v\psi}, \phi^0, u^0)$ will vary continuously, in such a manner that it will always be a boundary point of the set $C(T_{v\psi})$. Let $S(t)$ denote the boundary of the set $C(t)$. Hence $\Lambda_{v\psi} \in S(T_{v\psi})$ for every admissible pair.

Let $\{v^j(t), \psi^j(t)\}$, $j = 1, 2, \cdots$, be a sequence of admissible pairs such that $v^j(t) \rightarrow v^0(t)$ and $\psi^i(t) \rightarrow \psi^0(t)$ uniformly, where $\{v^0(t), \psi^0(t)\}$ is the optimal pair for the system Z. Consider the sequence of times

$$(4.6)\qquad \{T_j = T_{v^j\psi^j} = T(u^0, \phi^0; v^j, \psi^j)\}_1^\infty,$$

which corresponds to the sequence $\{v^j(t), \psi^j(t)\}_1^\infty$. Since

$$(4.7)\qquad T^0 = \max_{v \in V, \psi \in \Psi} T_{v\psi} = T(u^0, \phi^0; v^0, \psi^0),$$

taking a subsequence if necessary, we may assume that

$$(4.8)\qquad T_j < T_{j+1} \quad \text{and} \quad \lim_{j\to\infty} T_j = T^0.$$

Consider now the point Λ^0 defined by (4.1) and the sequence of points

$$(4.9)\qquad \{\Lambda_j = \Omega^*(T_j, \phi^0, u^0) = \Theta^*(T_j, \psi^j, v^j)\}_1^\infty.$$

Since the functionals $\Omega^*(t, \phi, u)$, $\Theta^*(t, \psi, v)$ and $T(u^0, \phi^0; v, \psi)$ are continuous in all their arguments, we have

$$(4.10)\qquad \lim_{j\to\infty} \Lambda_j = \Lambda^0.$$

We shall now prove the following.

THEOREM 2. *The point Λ^0 is a boundary point of the set $C(T^0)$.*

Proof. Suppose, on the contrary, that Λ^0 is an interior point of the set $C(T^0)$. Then, by property (III) in §2, there exists an $\epsilon > 0$ such that $N_\epsilon(\Lambda^0) \subset C(\tau)$ for all τ in the interval $T^0 - \epsilon < \tau < T^0$, where $N(\Lambda^0)$ is an ϵ-neighborhood of Λ^0. The continuity of $\Theta^*(t, \psi^0, v^0)$ at $t = T^0$ implies that there exists a $\delta > 0$ such that $\Theta^*(t, \psi^0, v^0) \in N_\epsilon(\Lambda^0)$ for all t in $T^0 - \delta < t \leqq T^0$. Let $2\gamma = \min(\delta, \epsilon)$. Then $\Theta^*(T^0 - \gamma, \psi^0, v^0) \in N(\Lambda^0) \subset C(T^0 - \gamma)$. But this is impossible, since $T^0 - \gamma < T^0$ and T^0 is the minimum value of t such that $\Theta^*(t, \psi^0, v^0) \in C(t)$.

Making use of Theorem 2 and property (I) in §2, the first part of the following theorem, which is an extension of Kelendzheridze's main theorem [4, 9] can be easily proved.

THEOREM 3. *There exists a unit vector $\eta^0 = (\eta_1{}^0, \cdots, \eta_n{}^0)$ of the n-dimensional Euclidean space R^n such that*

$$\text{(I)} \qquad \eta^0 \cdot \Omega^*(T^0, \phi, u) \leqq \eta^0 \cdot \Omega^*(T^0, \phi^0, u^0),$$

for all admissible pairs $\{u(t), \phi(t)\}$, $\Omega^*(T^0, \phi, u) \neq \Omega^*(T^0, \phi^0, u^0)$;

$$\text{(II)} \qquad \eta^0 \cdot \Theta^*(T^0, \psi, v) \leqq \eta^0 \cdot \Theta^*(T^0, \psi^0, v^0),$$

for all admissible pairs $\{v(t), \psi(t)\}$ which are sufficiently near to the optimal pair $\{v^0(t), \psi^0(t)\}$ and $\Theta^*(T^0, \psi, v) \in C(T^0)$;

$$\text{(III)} \qquad \eta^0 \cdot G(\Lambda^0, v^0(T^0), T^0) \leqq \eta^0 \cdot F(\Lambda^0, u^0(T^0), T^0),$$

where

$$F(x(t), u(t), t) = -\sum_{i=1}^{m} A_i(t - c_m)x(t - c_m + c_i) + A(t - c_m)u(t - c_m), \tag{4.11}$$

and

$$G(z(t), v(t), t) = -\sum_{j=1}^{k} B_j(t - d_k)z(t - d_k + d_j) + B(t - d_k)v(t - d_k). \tag{4.12}$$

Proof. Consider, following the general lines of the method of proof due to Kelendzheridze, the union Σ of all the sets $C(t)$ for $t \geqq a$. Σ is an open set and $\Lambda^0 \in C(T^0) \subset \Sigma$. Therefore, we can find a number $t^* < T^0$ such that $\Theta^*(t, \psi^0, v^0) \in \Sigma$ for $t^* \leqq t \leqq T^0$. For every t in this interval let $\tau = \tau(t)$ be the number which is defined by the relation $\Theta^*(t, \psi^0, v^0) \in S(\tau)$, where $S(\tau)$ is the boundary of the set $C(\tau)$. Since the elements of the closure of $C(t)$ are continuous in all the arguments, $S(t)$ varies continuously with t. From the continuity of the function $\Theta^*(t, \psi^0, v^0)$, we see that

$\tau = \tau(t)$ is continuous in the interval $t^* \leqq t \leqq T^0$. We can easily show, as in the case considered by Kelendzheridze, that

$$(4.13) \qquad \tau(t) > t \quad \text{for} \quad t < T^0, \quad \text{and} \quad \tau(T^0) = T^0.$$

Since $\Theta^*(t, \psi^0, v^0)$ is on the boundary $S(\tau)$ of the convex body $C(\tau)$, there exists a support hyperplane Π_τ to $C(\tau)$ at $\Theta^*(t, \psi^0, v^0)$. Let η_τ be the unit vector orthogonal to this support hyperplane which is directed from $\Theta^*(t, \psi^0, v^0)$ into the halfspace which does not contain the convex body $C(\tau)$. For every point $\Omega^*(t, \phi, u) \in C(\tau)$, the vector $\Omega^*(t, \phi, u) - \Theta^*(t, \psi^0, v^0)$ is directed into the halfspace which contains $C(\tau)$. Hence

$$(4.14) \qquad (\Omega^*(t, \phi, u) - \Theta^*(t, \psi^0, v^0), \eta_\tau) \leqq 0,$$

for $\Omega^*(t, \phi, u) \in C(\tau)$.

Consider now the sequence of times $\{T_j\}_1^\infty$ defined by (4.6) which has the property (4.8). If j is sufficiently large, $t^* < T_j < T^0$. It follows, from (4.8), (4.13) and from the continuity of $\tau(t)$, that

$$(4.15) \qquad \lim_{j\to\infty} \tau(T_j) = T^0.$$

Since in n-dimensional Euclidean space the unit sphere is compact, there exists a convergent subsequence of the sequence of unit vectors $\{\eta_{\tau(T_j)}\}_1^\infty$ which converges to the unit vector η_{T^0} which is orthogonal to the support hyperplane Π_{T^0} to the set $C(T^0)$ at the boundary point Λ^0, directed into the halfspace which does not contain the convex body $C(T^0)$.

If $\Omega^*(t, \phi, u)$ is an interior point of the set $C(T^0)$, $\Omega^*(t, \phi, u) \in C(T_j)$ for sufficiently large j, by property (III) in §2. From (4.14) we see that

$$(\Omega^*(t, \phi, u) - \Theta^*(T_j, \psi^0, v^0), \eta_{\tau(T_j)}) \leqq 0.$$

Taking the relations (4.1) and (4.15) and the continuity of $\tau(t)$ into account, for $j \to \infty$ we obtain

$$(4.16) \qquad (\Omega^*(t, \phi, u) - \Omega^*(T^0, \phi^0, u^0), \eta_{T^0}) \leqq 0.$$

Let $\{v(t), \psi(t)\}$ be an admissible pair for the pursued system Z. $\Theta^*(t, \psi, v)$ can be captured by the pursuing system X with optimal policy at $t = T_{v\psi} < T^0$. Suppose that the pair $\{v(t), \psi(t)\}$ is sufficiently near to the optimal pair $\{v^0(t), \psi^0(t)\}$ and is such that $\Theta^*(T_{v\psi}, \psi, v)$ is an interior point of the set $C(T^0)$. Then, by property (III) in §2 and by (4.8), $\Theta^*(T_{v\psi}, \psi, v)$ is an interior point of the sets $C(T_j)$ for all large j. Therefore $\Theta^*(T_j, \psi, v) \in C(T_j)$.

Consider now the points Λ_j, $j = 1, 2, \cdots$, defined by (4.9). Since $\Lambda_j \in S(T_j)$, it follows that $\Theta^*(T_j, \psi^j, v^j) \in S(T_j)$.

Since $\Theta^*(T_j, \psi, v)$ is an interior point of $C(T_j)$, and since $\Theta^*(T_j, \psi^j, v^j)$

is on the boundary of $C(T_j)$, the vector $\Theta^*(T_j, \psi, v) - \Theta^*(T_j, \psi^j, v^j)$, which passes through Λ_j, is directed into the halfspace which contains the convex body $C(T_j)$. Consequently

$$(4.17) \qquad (\Theta^*(T_j, \psi, v) - \Theta^*(T_j, \psi^j, v^j), \eta_{T_j}) \leqq 0,$$

for sufficiently large j. Passing to the limit as $j \to \infty$ in (4.17), we obtain

$$(4.18) \qquad (\Theta^*(T^0, \psi, v) - \Theta^*(T^0, \psi^0, v^0), \eta_{T^0}) \leqq 0.$$

Using the linearity of the functional $\Theta^*(t, \psi, v)$, we may write $\Delta\Theta^* = \Theta^*(T^0, \psi, v) - \Theta^*(T^0, \psi^0, v^0) = \Theta^*(T^0, \psi - \psi^0, v - v^0)$. Then the inequality (4.18) can be written in the form

$$(4.19) \qquad (\Delta\Theta^*, \eta_{T^0}) \leqq 0.$$

The formula (4.19) is true for every admissible pair $\{v(t), \psi(t)\}$ which is sufficiently near to the optimal pair $\{v^0(t), \psi^0(t)\}$ and is such that $\Theta^*(t, \psi, v) \in C(T^0)$.

Consider now the function

$$(4.20) \qquad \zeta_j(t) = (\Omega^*(t, \phi^0, u^0) - \Theta^*(t, \psi^0, v^0), \eta_{\tau(T_j)}).$$

Since $\Omega^*(t, \phi^0, u^0) \in C(t)$ for every t, it follows from (4.14) that $\zeta_j(T_j) \leqq 0$. Since $\Omega^*(T^0, \phi^0, u^0) = \Theta^*(T^0, \psi^0, v^0)$, we have $\zeta_j(T^0) = 0$. Since $u^0(t)$ and $v^0(t)$ are piecewise continuous and since the kernel functions $H(s, t)$ and $K(s, t)$ are continuously differentiable in $T_j \leqq t \leqq T^0$ if j is sufficiently large, the function $\zeta_j(t)$ has a continuous derivative in the interval $T_j \leqq t \leqq T^0$. Hence $\zeta_j'(\lambda_j) \geqq 0$ for some λ_j such that $T_j < \lambda_j < T^0$. Therefore

$$(4.21) \qquad \lim_{j\to\infty} \zeta_j'(\lambda_j) \geqq 0.$$

Consequently

$$(4.22) \qquad (w, \eta_{T^0}) \leqq 0,$$

where

$$(4.23) \qquad w = G(\Lambda^0, v^0(T^0), T^0) - F(\Lambda^0, u^0(T^0), T^0),$$

F and G being defined by (4.11) and (4.12).

Let $C^*(T^0)$ be the convex hull of $C(T^0)$ and w, and let K^* be the convex cone, with vertex Λ^0, of the vectors $\Delta\Theta^*$ which emanate from Λ^0. Since, by the above argument, the convex set $C(T^0)$ and the vectors $\Delta\Theta^*$ and w all lie on one side of the support hyperplane Π_{T^0} to the convex body $C^*(T^0)$ at Λ^0, the set $C^*(T^0)$ and the vector $-\Delta\Theta^*$ lie in two opposite closed halfspaces defined by Π_{T^0}. Hence, the vector $-\Delta\Theta^*$, which emanates from Λ^0, does not pass through interior points of the convex body $C^*(T^0)$.

The vectors $-\Delta\Theta^*$ form a convex cone K which is symmetric to K^* with respect to Λ^0. Therefore, K does not intersect the interior of the convex body $C^*(T^0)$. Since $C^*(T^0)$ has interior points (because $C(T^0)$ has), $C^*(T^0)$ and K are separated by a hyperplane Π^0. Therefore the convex hull $C^*(T^0)$ and the convex cone K^* lie in one closed halfspace defined by Π^0 and the cone K is contained in the other. Let η^0 be the unit vector which emanates from Λ^0, is orthogonal to Π^0, and is directed into the halfspace which contains K. Thus, for this vector η^0 the relations (4.16), (4.19) and (4.22) are satisfied, namely

(I) $(\Omega^*(t, \phi, u) - \Omega^*(T^0, \phi^0, u^0), \eta^0) < 0$, for $\Omega^*(t, \phi, u) \in C(T^0)$;

(II) $(\Delta\Theta^*, \eta^0) < 0$ for $\Delta\Theta^* \in K^*$;

(III) $(w, \eta^0) < 0$.

This completes the proof of Theorem 3.

5. Optimal strategies in a particular case. If the control regions U and V are defined by (1.14), and if the systems (1.1) and (1.2) are such that no component of $\eta^0 \cdot H(t, T^0 - c_m)$ or $\eta^0 \cdot K(t, T^0 - d_k)$ is identically zero on an interval of positive length for $\eta^0 \neq 0$, we can easily show that (see [5]) optimal control functions $u^0(t)$ and $v^0(t)$ are of the form

$$u^0(t) = \operatorname{sgn} [\eta^0 V(t, T^0 - c_m) A(t)], \tag{5.1}$$

for $a \leqq t \leqq T^0 - c_m$, and

$$v^0(t) = \operatorname{sgn} [\eta^0 W(t, T^0 - d_k) B(t)], \tag{5.2}$$

for $c \leqq t \leqq T^0 - d_k$.

In §1, the sets Φ and Ψ of initial functions were defined as closed compact subsets of the sets of all real-valued n-dimensional vector functions $\phi(t)$ and $\psi(t)$ continuous in the initial intervals $a \leqq t \leqq b$ and $c \leqq t \leqq d$ respectively. If $\phi(t)$ and $\psi(t)$ are measurable functions in their intervals of definition, the analysis in the previous sections is still valid. If, in addition to this,

$$\Phi: \quad \{|\phi_i(t)| \leqq 1, \quad a \leqq t \leqq b, \quad i = 1, \cdots, n\}, \tag{5.3}$$

and

$$\Psi: \quad \{|\psi_i(t)| \leqq 1, \quad c \leqq t \leqq d, \quad i = 1, \cdots, n\}, \tag{5.4}$$

we can show without any difficulty that

$$\phi^0(t) = \operatorname{sgn} [\eta^0 \cdot H(t, T^0 - c_m)], \quad a \leqq t \leqq b, \tag{5.5}$$

and

$$\psi^0(t) = \operatorname{sgn} [\eta^0 \cdot K(t, T^0 - d_k)], \quad c \leqq t \leqq d, \tag{5.6}$$

provided no component of $\eta^0 \cdot H(t, T^0 - c_m)$ or $\eta^0 \cdot K(t, T^0 - d_k)$ is identically zero on an interval of positive length for $\eta^0 \neq 0$.

Thus, if the vector η^0 is known, the optimal strategy $\{u^0(t), \cdots, \psi^0(t)\}$ is completely determined. We shall now develop a method for finding η^0.

Let the functions $u_\eta(t)$, $\phi_\eta(t)$, $v_\eta(t)$, and $\psi_\eta(t)$ be defined by (5.1), (5.2), (5.5), and (5.6), respectively, with η replacing η^0, namely,

$$(5.7)\quad \begin{aligned} u_\eta(t) &= \operatorname{sgn}\{\eta V(t - c_m, T^0 - c_m)A(t)\}, \qquad b \leqq t \leqq T^0, \\ \phi_\eta(t) &= \operatorname{sgn}\{\eta H(t, T^0 - c_m)\}, \qquad a \leqq t \leqq b, \end{aligned}$$

and

$$(5.8)\quad \begin{aligned} v_\eta(t) &= \operatorname{sgn}\{\eta W(t - d_k, T^0 - d_k)B(t)\}, \qquad c \leqq t \leqq T^0, \\ \psi_\eta(t) &= \operatorname{sgn}\{\eta K(t, T^0 - d_k)\}, \qquad c \leqq t \leqq d. \end{aligned}$$

Clearly, more than one η may determine the same strategy

$$\{u_\eta(t), \cdots, \psi_\eta(t)\} \quad \text{and} \quad u_{\eta^0}(t) = u^0(t), \cdots, \psi_{\eta^0}(t) = \psi^0(t).$$

Note that the functions $u_\eta(t), \cdots, \psi_\eta(t)$ depend continuously upon η, disregarding sets of measure zero. Consequently, the functionals $\Omega^*(t, \phi_\eta, u_\eta)$, $\Theta^*(t, \psi_\eta, v_\eta)$ and $T(u_\eta, \phi_\eta; v_\eta, \psi_\eta)$ are continuous in η as well as in t. Let us also note that, if $T(u_\eta, \phi_\eta; v_\eta, \psi_\eta) = T^0$ for some vector η, the vector η and η^0 determine the same optimal strategy $\{u^0(t), \cdots, \psi^0(t)\}$.

THEOREM 4. *There exist two positive numbers γ and δ such that $\Omega^*(t, \phi_\eta, u_\eta)$ and $\Theta^*(t, \psi_\eta, v_\eta)$ are boundary points of the set $C(T^0)$ for all t in $T^0 - \gamma < t \leqq T^0$ and for all η in $\|\eta - \eta^0\| < \delta$, provided $\Omega^*(t, \phi_\eta, u_\eta) \in C(T^0)$ and $\Theta^*(t, \psi_\eta, v_\eta) \in C(T^0)$.*

Proof. Suppose that $\Theta^*(t, \psi_\eta, v_\eta)$ is an interior point of the set $C(T^0)$ for some t and for some η. Hence, by property (III) in §2, there exists an $\epsilon > 0$ such that $N_\epsilon(\Theta^*(t, \psi_\eta, v_\eta)) \subset C(\tau)$ for all τ in $T^0 - \epsilon < \tau \leqq T^0$, where $N_\epsilon(\Theta^*)$ is an ϵ-neighborhood of $\Theta^*(t, \psi_\eta, v_\eta)$. Consequently,

$$(5.9)\qquad \|\Theta^*(t, \psi_\eta, v_\eta) - \Lambda^0\| \geqq \epsilon,$$

since Λ^0 lies on the boundary of $C(T^0)$. From the continuity of $\Theta^*(t, \psi_\eta, v_\eta)$ at $t = T^0$ and $\eta = \eta^0$ and since $\Lambda^0 = \Theta^*(T^0, \psi_{\eta^0}, v_{\eta^0})$, we can find two positive numbers γ and δ such that

$$(5.10)\qquad \|\Theta^*(t, \psi_\eta, v_\eta) - \Lambda^0\| < \epsilon$$

for all t in $T^0 - \gamma < t \leqq T^0$ and for all η in $\|\eta - \eta^0\| < \epsilon$. Therefore, $\Theta^*(t, \psi_\eta, v_\eta)$ cannot be an interior point of $C(T^0)$ for $T^0 - \gamma < t \leqq T^0$ and $\|\eta - \eta^0\| < \delta$, because, in the contrary case, the inequality (5.9) must be satisfied, which contradicts the inequality (5.10). Since $\Theta^*(t, \psi_\eta, v_\eta) \in C(T^0)$, $\Theta^*(t, \psi_\eta, v_\eta)$ is a boundary point of $C(T^0)$.

It can similarly be shown that $\Omega^*(t, \phi_\eta, u_\eta)$ lies on the boundary of the set $C(T^0)$.

Consider now the time $T_\eta = T(u_\eta, \phi_\eta, v_\eta, \psi_\eta)$ at which $\Omega^*(t, \phi_\eta, u_\eta)$ encounters $\Theta^*(t, \psi_\eta, v_\eta)$. As we mentioned above, T_η is continuous almost everywhere in η. Accordingly, for almost all η in $\|\eta - \eta^0\| < \delta$, we have $T^0 - \gamma < T_\eta \leqq T^0$.

Suppose now that γ and δ satisfy the conditions of Theorem 4. Then, $\Omega^*(t, \phi_\eta, u_\eta)$ and $\Theta^*(t, \psi_\eta, v_\eta)$ lie on the boundary $S(T^0)$ of the set $C(T^0)$ and coincide for $t = T_\eta$. Let S^0 be the portion of $S(T^0)$ described by the points $\Omega^*(t, \phi_\eta, u_\eta)$ and $\Theta^*(t, \psi_\eta, v_\eta)$ for $T^0 - \gamma < t \leqq T^0$ and $\|\eta - \eta^0\| < \delta$.

Consider the convex set H^0 of all vectors η, orthogonal to the support hyperplanes Π to S^0 and directed into the halfspaces (defined by these hyperplanes Π), which do not contain the convex body $C(T^0)$. Clearly $\eta^0 \in H^0$. Thus,

$$\eta\Omega^*(t, \phi, u) < \eta\Omega^*(T_\eta, \phi_\eta, u_\eta), \tag{5.11}$$

and

$$\eta\Theta^*(t, \psi, v) < \eta\Theta^*(T_\eta, \psi_\eta, v_\eta) \tag{5.12}$$

for all $\Omega^*(t, \phi, u) \in C(T^0)$, $\Theta^*(t, \psi, v) \in C(T^0)$, $T^0 - \gamma < T_\eta \leqq T^0$, and $\eta \in H^0$.

Define the function $w_\eta(t)$ by

$$w_\eta(t) = G(\Theta^*(t, \psi_\eta, v_\eta), v_\eta(t), t) - F(\Omega^*(t, \phi_\eta, u_\eta), u_\eta(t), t), \tag{5.13}$$

where F and G are given by (4.11) and (4.12). As in the proof of Theorem 3, we can easily show that

$$(w_\eta(t), \eta) \leqq 0 \tag{5.14}$$

for all $\eta \in H^0$.

Consider now the function $V(t, \eta)$ defined by

$$V(t, \eta) = \max\{V_1(t, \eta), V_2(t, \eta), V_3(t, \eta)\}, \tag{5.15}$$

where

$$\begin{cases} V_1(t, \eta) = (\Omega^*(t, \phi_\eta, u_\eta) - \Lambda^0, \eta), \\ V_2(t, \eta) = (\Theta^*(t, \psi_\eta, v_\eta) - \Lambda^0, \eta), \\ V_3(t, \eta) = (w_\eta(t), \eta). \end{cases} \tag{5.16}$$

From the continuity of the functions $V_i(t,\eta)$, $i = 1, 2, 3$, for all t ($\geqq e = \max(b, d)$) and for almost all η, we can easily see that the function $V(t, \eta)$ is continuous in $t(\geqq e)$ and in η, disregarding a set of measure zero. We have also

$$V(T^0, \eta) = 0, \quad \text{for} \quad \eta \in H^0. \tag{5.17}$$

Let H be the set of all vectors η for which

$$V(e, \eta) < 0, \tag{5.18}$$

and denote by H^1 the subset of H whose elements η verify the inequality

$$V(T^0, \eta) \leqq 0. \tag{5.19}$$

By the inequalities (5.11)–(5.14), $H^0 \subset H$. Clearly,

$$V(T^0, \eta) > 0 \quad \text{for} \quad \eta \in H - H^1. \tag{5.20}$$

Suppose now that $V(t, \eta)$ is strictly increasing at $t = T^0$ for every $\eta \in H^1$. From (5.17), (5.18) and (5.20), we see that there exists a unique $T(\eta)$ such that

$$V(T(\eta), \eta) = 0, \tag{5.21}$$

for t in a neighborhood of T^0 and in a neighborhood of $\eta^0 \in H^1$. Clearly, if $\eta \in H^0$, $T(\eta) = T^0$, and if $\eta \in H - H^1$, $T(\eta) < T^0$ according to (5.20). Thus, we have the following theorem, which is an extension of a theorem due to Neustadt [6] as well as of a theorem due to the author [7].

THEOREM 5. *Suppose that the control regions U and V and the sets of initial functions Φ and Ψ of the systems X and Z consist of measurable functions satisfying the conditions* (1.14), (5.3) *and* (5.4). *Let T^0 be pursuit time for the systems X and Z. Then the unique optimal strategy is given by* (5.7) *and* (5.8) *with some vector η in some set H^1. If for every $\eta \in H^1$ the function $V(t, \eta)$ defined by* (5.15) *is strictly increasing with t at $t = T^0$, then for η in a neighborhood of H^0 and t in a neighborhood of T^0 the vectors $\eta \in H$ maximize the time for which $V(t, \eta) = 0$.*

This theorem is very close to Kelendzheridze's main result in [4, 9], when the retardations in X and Z all vanish.

Note that Theorem 5 only gives a necessary condition for an optimal strategy.

6. Remark. If Φ, as in §1, consists only of continuous functions, and if we know the optimal functions $u^0(t)$, $v^0(t)$, and $\psi^0(t)$, the optimal initial function $\phi^0(t)$ can be obtained by solving the integral equation

$$\int_a^b H(s, T^0 - c_m)\phi^0(s)\, ds = P, \tag{6.1}$$

where

$$P = \Theta^*(T^0, \psi^0, v^0) - \int_a^{T^0 - c_m} V(s, T^0 - c_m)A(s)u^0(s)\, ds. \tag{6.2}$$

With obvious modifications, we can state a similar result for $\psi^0(t)$ when $u^0(t)$, $\phi^0(t)$, and $v^0(t)$ are known.

7. Acknowledgment. The author wishes to express his appreciation to the referee for his valuable comments.

REFERENCES

[1] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.

[2] H. G. Eggleston, *Convexity*, Cambridge Tracts in Math. and Math. Phys., No. 47, Cambridge.

[3] G. L. Haratishvili, *The maximum principle in the theory of optimal processes involving delay*, Dokl. Akad. Nauk SSSR, 136 (1961), pp. 39–42.

[4] D. L. Kelendzheridze, *On the theory of optimal pursuit*, Dokl. Akad. Nauk SSSR, 138 (1961), pp. 529–532. (English trans. in Soviet Math.-Dokl., 2 (1961), pp. 654–656.)

[5] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillation, 5, pp. 1–24.

[6] L. W. Neustadt, *Synthesizing time optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–493.

[7] M. N. Oğuztöreli, *A time optimal control problem for systems described by differential-difference equations*, this Journal, 1 (1963), pp. 290–310.

[8] ———, *Relay type control systems with retardation and switching delay*, this Journal, 1 (1963), pp. 275–289.

[9] L. S. Pontryagin, V. C. Boltyanskii, R. W. Gamgrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, (English trans., L. W. Neustadt and K. N. Trirogoff), Interscience, New York, 1962.

J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1962

ON THE DIFFERENTIAL EQUATIONS SATISFIED BY CONDITIONAL PROBABILITY DENSITIES OF MARKOV PROCESSES, WITH APPLICATIONS*

HAROLD J. KUSHNER†

1. Introduction and summary. Consider the vector stochastic differential equation,

$$(1)\qquad dx_i = f_i(x)dt + \sum_k F_{ik}(x)dz_k(t), \qquad i = 1, \cdots, n,$$

where each $z_i(t)$ is an independent Brownian motion process with unit variance parameter. Let x, f and z be vectors with components x_i, f_i and z_i, respectively; let $F(x)$ be the matrix with components $F_{ij}(x)$, and $V(x)$ the matrix with components $v_{ij}(x)$, where $V = FF'$. Let $\hat{P}(a, t)$ be the probability density of $x(t)$ given only the density of $x(t_0)$, $t \geqq t_0$. Under suitable conditions on f and F, it is well-known that (for almost all $z(\cdot)$ functions) there exists a unique solution to (1) which is a Markov process. If $\hat{P}$ is suitably differentiable, then Kolmogorov's forward equation,

$$(2)\qquad \frac{\partial \hat{P}(a, t)}{\partial t} = -\sum_{i=1}^{n} (f_i(a)\hat{P}(a, t))_{a_i} + \frac{1}{2} \sum_{i,j=1}^{n} (v_{ij}(a)\hat{P}(a, t))_{a_i a_j},$$

is satisfied, where the subscript a_i denotes the partial derivative.

A problem of great practical importance arises when noise corrupted observations on x are taken; i.e., the vector‡ $dy = g(x)dt + dw$ is available, where w is a vector Brownian motion process. For example, x may represent a signal stochastic process and dy/dt the (nonlinear function of the) signal plus noise, or x may represent the evolution of a dynamical system driven by a noise process and the interest may be in the estimation of various properties of x or, perhaps, the control of x. In these cases it would be very desirable to have an expression for the probability density of x conditioned upon the observations, as well as upon the initial data. The existence of such an equation is suggested by theorems|| in [3, pp. 287–291]. Here, we derive a partial differential equation satisfied by this conditional density. The equation is of the form (2) with an additional term which contains the ob-

* Received by the editors January 24, 1964. This research was supported by the United States Air Force, under Contracts No. AF 33(657)-8559, and No. AF 49(638)-1206.

† Research Institute for Advanced Studies (RIAS), 7212 Bellona Avenue, Baltimore, Maryland.

‡ This could be written without differentials as $b = g(x) + \Psi$, where Ψ is the white Gaussian noise dw/dt.

|| The relation between our results and these theorems is further discussed in the Appendix.

servation in a linear manner, and in many cases, is amenable to convenient analog or digital simulation; hence, the actual conditional density may be obtained as it evolves in time. The equation promises to be of great usefulness in communications and control problems.

The principal result is the following. For any function of time $s(t)$, define $\delta s(t) = s(t + \Delta) - s(t)$ and $ds(t) = s(t + dt) - s(t)$. Let $E\delta w \delta w' = \Sigma\Delta$ and $E\delta z\delta w' = C\Delta$ and assume Σ is nonsingular.[†] Let $P(a, t \mid t)$ be the conditional density of $x(t)$ given all observations up to t, and let

$$\begin{aligned} d\bar{f}(a, t) &= f(a, t)dt + FC\Sigma^{-1}(dy - g(a)dt), \\ \bar{V} &= V - FC\Sigma^{-1}(FC)', \end{aligned} \tag{3}$$

$$dQ(a, t) = P(a, t \mid t)\cdot(dy - Eg(a)dt)'\Sigma^{-1}(g(a) - Eg(a)), \tag{4}$$

where the expectation E is the conditional expectation using $P(a, t \mid t)$. Then $P(a, t \mid t)$ satisfies

$$\begin{aligned} &P(a, t + dt \mid t + dt) - P(a, t \mid t) \\ &\qquad = dP(a, t \mid t) = dQ(a, t) - \sum_1^n (d\bar{f}_i(a, t)\cdot P(a, t \mid t))_{a_i} \\ &\qquad\qquad + \tfrac{1}{2} \sum_{i,j=1}^n (\bar{v}_{ij}(a) \cdot P(a, t \mid t))_{a_i a_j} dt. \end{aligned} \tag{5}$$

In certain cases (discussed in §3j) which are reducible to the case where a takes on only values $x^1, \cdots, x^s$, (5) becomes

$$dP(i \mid t) = P(i \mid t)\cdot(dy - Eg(i, t))'\Sigma^{-1}(g(i, t) - Eg(i, t)). \tag{5$'$}$$

Equation (5$'$) is generally rigorously verifiable.

Although (5) can be rigorously verified in a number of cases, it is, of course, still formal in general (see Appendix), being derived under the assumption that P exists and is suitably differentiable[‡]. If there is no correlation between the observation noise dw and the noise dz, then $C = 0$ and $d\bar{f}_i = f_i dt$ and $\hat{v}_{ij} = v_{ij}$. In this case the last two terms on the right of (5) are the same as in (2), and (5) differs from (2) only in that the former contains the observation term dQ, where dQ is *linear* in the differential observation dy.

The same problem was considered in [1], where x was scalar and $g(x) = x$. A more general problem was discussed in [2] but, as discussed in [1], the results in [2] are incorrect through the omission of certain significant terms. Since the writing of [1], substantial and surprising simplifications (which were initially inapparent) in the form of the scalar equation have been

[†] Σ and C are assumed to be independent of x; if Σ depended on x, the problem appears to degenerate to one where x can be determined exactly at every t.

[‡] When $f = F = 0$, $dP = dQ$ and is simple to verify.

obtained. In this paper, taking advantage of these simplifications, the results for the general vector case with nonlinear observations are derived. These results include, as special cases, many important situations (as will be illustrated) that cannot be represented by the scalar case format.

The derivation is performed in §2. Section 3 discusses several special cases and extensions. The results include as special cases known results [4] for the filtering problem where the signal and noise are Gaussian and finite order Markovian.

Usually, when one has a stochastic differential equation, one seeks properties of the random functions which they define. In this paper, the inverse problem occurs initially: given a random function, what stochastic differential equation does it satisfy? The Appendix contains a discussion of this problem and of the sense in which such an equation is meaningful—as well as of other points which are important in the derivation.

2. The main result. The derivation proceeds by assuming the finite difference model (6) and taking formal limits subsequently.

$$\begin{aligned} \delta x &= f(x)\Delta + F(x)\delta z, \\ \delta y &= g(x)\Delta + \delta w. \end{aligned} \tag{6}$$

Let Y denote the $y(\tau)$, $\tau \leqq t$, the entire set of observations up to t; $\delta y = y(t + \Delta) - y(t)$ is the observation at t given by (6).

The following notation will be used. Let a and α be the generic value of x, and let M and N be any random quantities. Let $P(a, t; M)$ denote the joint density of $x(t)$ and M; $P(a, t \mid M)$ denotes the density of $x(t)$ conditioned upon M; $P(a, t \mid Y)$ will also be written as $P(a, t \mid t)$ or as P; $P(a, t \mid t + \Delta)$ denotes $P(a, t \mid Y, \delta y)$, the density of $x(t)$ conditioned upon the set of past observations Y and also upon the present vector observation δy; $P(M \mid a, t; N)$ denotes the conditional density of M, given $x(t) = a$ and N.

The derivation takes place in two parts. First, let $P(a, t \mid t)$ be given, take the observation δy, and compute $P(a, t \mid t + \Delta) - P(a, t \mid t)$, the change in the conditional density due to the last observation. This change is given by (14). The second part of the derivation assumes the change δx in x, and the Chapman-Kolmogorov equation is applied to include the effects of δx on the conditional density. Formal limits are then taken and the derivation is complete.

Derivation: Part 1. According to the notational convention

$$\begin{aligned} P(a, t \mid t + \Delta) &= P(a, t; Y, \delta y)/P(Y, \delta y) \\ &= \frac{P(\delta y \mid a, t; Y)P(a, t \mid Y)P(Y)}{P(Y) \int P(\delta y \mid a, t; Y)P(a, t \mid Y)\, da}. \end{aligned} \tag{7}$$

Since the distribution of δy is completely specified when a is given, (7) may be written as

$$(8) \qquad \frac{P(a, t \mid t)P(\delta y \mid a, t)}{\int P(\delta y \mid a, t)P(a, t \mid t)\, da}.$$

In fact, as discussed in [1], $P(a, t \mid t)$ is a Markov process in function space. Now, from (5),

$$(9) \qquad P(\delta y \mid a, t) \sim N[g\Delta, \Sigma\Delta],$$

where $N[g\Delta, \Sigma\Delta]$ denotes normal density with mean $g\Delta$ and covariance matrix $\Sigma\Delta$.

Substituting (9) into (8) yields

$$(10) \quad \frac{P(a, t \mid t) \exp\left[-\frac{1}{2\Delta}(\delta y - g(a)\Delta)'\Sigma^{-1}(\delta y - g(a)\Delta)\right]}{\int P(a, t \mid t) \exp\left[-\frac{1}{2\Delta}(\delta y - g(a)\Delta)'\Sigma^{-1}(\delta y - g(a)\Delta)\right] da},$$

where $da = \prod_1^q da_i$. Equation (10) may be further simplified by deleting the common term $\exp\left[-\frac{1}{2\Delta}\delta y'\Sigma^{-1}\delta y\right]$ from both numerator and denominator†. Thus,

$$(11) \quad \begin{aligned} R(\Delta, \delta y) &\overset{\Delta}{=} \frac{P(a, t \mid t + \Delta)}{P(a, t \mid t)} \\ &= \frac{\exp\left[\delta y'\Sigma^{-1}g(a) - \frac{1}{2}g'(a)\Sigma^{-1}g(a)\Delta\right]}{\int P(a, t \mid t) \exp\left[\delta y'\Sigma^{-1}g(a) - \frac{1}{2}g'(a)\Sigma^{-1}g(a)\Delta\right] da}. \end{aligned}$$

Assuming that the appropriate moments of $P(a, t \mid t)$ exist, (11) may be differentiated any number of times with respect to the infinitesimals Δ and δy_i. We wish to obtain an expansion of (11) which contains all terms of order Δ or less. Since $E\delta y \delta y' = \Sigma\Delta$, the expansion must be carried to the second degree in the components of δy, and to the first degree in Δ. It is easily shown that the remainder in the expansion has a mean value of smaller order than Δ and a mean square value of smaller order than Δ^2.

The differentiation of (11) is straightforward. Recalling that E refers to the expectation using $P(a, t \mid t)$, we have

$$\begin{aligned} R_\Delta(0, 0) &= -\tfrac{1}{2}[g'(a)\Sigma^{-1}g(a) - E(g'(a)\Sigma^{-1}g(a))], \\ R(0, 0) &= 1, \\ (12) \qquad R_{\delta y}(0, 0) &= \Sigma^{-1}g(a) - \Sigma^{-1}Eg(a), \end{aligned}$$

† If Σ depended upon x, this could not be done.

$$R_{\delta y,\delta y}(0,0) = (\Sigma^{-1}g(a))(\Sigma^{-1}g(a))' - 2\Sigma^{-1}g(a)(\Sigma^{-1}Eg(a))' + 2(\Sigma^{-1}Eg(a))(\Sigma^{-1}Eg(a))' - E(\Sigma^{-1}g(a)(\Sigma^{-1}g(a))'),$$

where $R_{\delta y}$ and $R_{\delta y,\delta y}$ are the gradient and Jacobian, respectively, of R with respect to δy. Thus,

$$\begin{aligned} &P(a, t \mid t + \Delta) = \\ (13)\quad &P(a, t \mid t)[1 + R_\Delta(0,0)\Delta + R'_{\delta y}(0,0)\delta y + \tfrac{1}{2}\delta y' R_{\delta y,\delta y}(0,0)\delta y] + r, \end{aligned}$$

where $Er \sim o(\Delta)$, $Er^2 \sim o(\Delta^2)$.

Although there is frequent occurrence of terms such as $E\delta y_i \delta y_j$ in probability theory, (13) is unusual in that these random terms are included without expectations. It would appear that these terms substantially complicate the result. It is quite remarkable that the term $\delta y_i \delta y_j$ may be replaced everywhere by its expectation without altering the result at all. The arguments for this are given in the Appendix: the replacement will be used hereafter in the text.† The simplification was not apparent in the earlier work. We have $E\delta y \delta y' = E[g(a)\Delta + \delta w][g(a)\Delta + \delta w]' = \Sigma\Delta + o(\Delta)$. Various terms in (12) may now be rewritten; e.g., replace $\delta y' \Sigma^{-1} g(a) \cdot (\Sigma^{-1}g(a))'\delta y = g(a)'\Sigma^{-1}\delta y \delta y' \Sigma^{-1} g(a)$ by $g'(a)\Sigma^{-1}g(a)\Delta + o(\Delta)$.

Now, adding and subtracting the terms $P(a, t \mid t)\ (Eg)'\Sigma^{-1}g\Delta$ and $P(a, t \mid t)\ (Eg)'\Sigma^{-1}(Eg)$, using the expectation substitutions for the second order terms, and rearranging terms yields

$$\begin{aligned} (14)\quad P(a, t \mid t + \Delta) - P(a, t \mid t) &\overset{\Delta}{=} \delta Q(a, t) \\ &= P(a, t \mid t) \cdot (\delta y - Eg\Delta)'\Sigma^{-1}(g - Eg) + r, \end{aligned}$$

where, again, $Er^2 \sim o(\Delta^2)$.

Completion of derivation. We are now prepared to use a modification of the Chapman-Kolmogorov [3] equation to complete our derivation by including the effects of $\delta x(t)$ on the conditional distribution. The method is a modification of the usual formal approach to the derivation of (2).

In general, by the definition of conditional probability,

$$P(a, t + \Delta \mid t + \Delta) = \int P(\alpha, t \mid t + \Delta) P(a, t + \Delta \mid \alpha, t;\ Y, \delta y)\ d\alpha.$$

If no observations are taken, it reduces to

$$\hat{P}(a, t + \Delta) = \int \hat{P}(\alpha, t) P(a, t + \Delta \mid \alpha, t)\ d\alpha.$$

† Although the second order random terms cannot be neglected since their expectation is of the order of Δ, their contribution is essentially deterministic. See Appendix for more details.

In our case, since dw is allowed to depend on dz but not on x, the distribution of $x(t + \Delta)$ is completely determined when δy and $x(t)$ are given. Thus

$$P(a, t + \Delta \,|\, t + \Delta) = \int P(\alpha, t \,|\, t + \Delta) P(a, t + \Delta \,|\, \alpha, t;\, \delta y)\, d\alpha. \tag{15}$$

To complete the procedure, multiply (15) by an arbitrary triply differentiable function $h(a)$, such that (19) holds and the integrals (16) exist. Thus from (15),

$$\begin{aligned} &\int h(a) P(a, t + \Delta \,|\, t + \Delta)\, da \\ &\quad = \iint h(\alpha + (a - \alpha)) P(\alpha, t \,|\, t + \Delta) P(a, t + \Delta \,|\, \alpha, t;\, \delta y)\, d\alpha\, da \\ &\quad = \iint \Big[h(\alpha) + h_a{}'(\alpha)(a - \alpha) \\ &\qquad\qquad + \frac{1}{2}(a - \alpha)' h_{a,a}(\alpha)\, (a - \alpha) + o((a - \alpha)'(a - \alpha)) \Big] \\ &\qquad\qquad \times \{P(\alpha, t \,|\, t + \Delta) P(a, t + \Delta \,|\, \alpha, t;\, \delta y)\, d\alpha\, da\}. \end{aligned} \tag{16}$$

Now, the density $P(a, t + \Delta \mid \alpha, t;\, \delta y)$ is normal. Since it is conditioned upon δy and $x(t)$, it is also conditioned upon δw. From standard theorems† on conditional normal variables [10],

$$\begin{aligned} E[a - \alpha \mid \delta y, x(t) = \alpha] &= E[f(\alpha)\Delta + F(\alpha)\delta z \mid \delta y = g(\alpha)\Delta + \delta w] \\ &\overset{\Delta}{=} \delta \hat{f} = f(\alpha)\Delta + (FC)\Sigma^{-1}(\delta y - g(\alpha)\Delta), \\ E[(a - \alpha)^2 \mid \delta y, x(t) = \alpha] &= E[(f(\alpha)\Delta + F(\alpha)\delta z)^2 \mid g(\alpha)\Delta + \delta w] \\ &\overset{\Delta}{=} \bar{V}\Delta = V\Delta - FC\Sigma^{-1}(FC)'\Delta + o(\Delta). \end{aligned}$$

Substituting these results into the last line of (16) yields

$$\int \Big[h(\alpha) + h_a{}'(\alpha)\,(\delta \hat{f}(\alpha)) + \frac{1}{2} \sum_{i,j} h_{a_i a_j}(\alpha) \bar{v}_{ij}(\alpha) \Delta \Big] P(\alpha, t \,|\, t + \Delta)\, d\alpha, \tag{17}$$

where $\bar{v}_{ij}$ is the (i, j)th entry of the matrix $\bar{V}$. Recall that $P(\alpha, t \mid t + \Delta) = P(\alpha, t) + \delta Q(\alpha, t)$.

† Given two normal vectors s, t, with $Es = \mu_s$, $Et = \mu_t$, $Est' = \Sigma_{12}$, $Ess' = \Sigma_{11}$, $Ett' = \Sigma_{22}$, we have $E[s \mid t] = \mu_s + \Sigma_{12}\Sigma_{22}^{-1}(t - \mu_t)$ and $\mathrm{Cov}[s \mid t] = \Sigma_{11} - \Sigma_{12}'\Sigma_{22}^{-1}\Sigma_{12}$. (See [10].)

Consider the term

$$\delta\hat{f}(\alpha)[P(\alpha, t \mid t) + \delta Q(\alpha, t)] =$$
$$[f(\alpha)\Delta + FC\Sigma^{-1}(\delta y - g(\alpha)\Delta)]$$
$$\cdot [1 + (\delta y - Eg(\alpha)\Delta)'\Sigma^{-1}(g(\alpha) - Eg(\alpha))]P(a, t \mid t).$$

Upon replacing $\delta y \delta y'$ by its expectation $\Sigma\Delta + o(\Delta)$ as discussed earlier, and rearranging, the term becomes* $\delta\bar{f}P \overset{\Delta}{=} [f\Delta + FC\Sigma^{-1}(\delta y - Eg\Delta) + o(\Delta)]P$. Upon replacing this in (17) and assuming (19), (17) may be partially integrated to yield (18).

$$(18) \qquad \int h(a)\left[P + \delta Q - \sum_i (\delta\bar{f}_i P)_{a_i} + \frac{\Delta}{2}\sum_{i,j} (\bar{v}_{ij}P)_{a_i a_j} + o(\Delta)\right].$$

$$(19) \qquad 0 = (\delta\bar{f}_i P)h \,\Big|_{a_i=-\infty}^{a_i=\infty} = (\bar{v}_{ij}P)h_{a_j} \,\Big|_{a_i=-\infty}^{a_i=\infty} = (\bar{v}_{ij}P)_{a_i} h \,\Big|_{a_i=-\infty}^{a_i=\infty}.$$

In (19), when $a_i = \pm\infty$, the a_j, $j \neq i$, are arbitrary. Equating (18) to the left hand side of (16) and recalling the arbitrariness of h yields, in the limit,†

$$\begin{aligned} dP &\overset{\Delta}{=} P(a, t + dt \mid t + dt) - P(a, t \mid t) \\ (20) \qquad &= dQ - \textstyle\sum_i (d\bar{f}_i P)_{a_i} + \frac{1}{2} \sum_{i,j} (\bar{v}_{ij}P)_{a_i a_j} dt, \\ dQ &= P(dy - Egdt)' \,\Sigma^{-1}(g - Eg). \end{aligned}$$

The equation (20) is the culmination of all our efforts. Observe that, as all the components of Σ tend to ∞ (as the value of the observations decreases), (20) tends to Kolmogorov's forward diffusion equation (in differential form). From a formal point of view, (20) may be divided through by dt and viewed as a differential equation with the observation dy/dt as a driving term or input.

It is easy to obtain a set of ordinary differential equations for the conditional moments of P. The method is given below.

3. Discussion of special cases and extensions.

3a. No dynamics. The simplest case is where $f = dz = 0$. Here x is an unknown vector. If some initial distribution $P(a,t_0)$ is assigned to x, then

$$(21) \qquad dP = P(dy - Egdt)'\Sigma^{-1}(g - Eg)$$

represents the conditional distribution.

* For brevity, $P = P(a, t)$, $\delta Q = \delta Q(a, t)$, $g(a) = g$ and $f(a) = f$ are used when no confusion will arise.

† As $\Delta \to 0$, the expectation of the $o(\Delta)$ in (17) is $o(\Delta)$ and its mean square value is $o(\Delta^2)$. Thus, we have (20) valid in the mean square sense, as discussed in the Appendix.

A special case of importance is where x may take on only finitely many values, $x^1, \cdots, x^s$. Since (21) must hold for each x^i, it reduces to a set of s ordinary differential equations with a simple analog computer representation, even for fairly general observation forms g.

$$Eg_i = \sum_j g_i(a^j)P(a^j, t \mid t).$$

3b. Linear dynamics. The case where $f(x) = Ax$, $F =$ constant, $g(x) = Gx$, and $P(a, 0)$ is Gaussian, where A and G are matrices, has been discussed in [4], where the ordinary differential equation for the conditional expectation of x was obtained. With our form, it is possible to compute all the moments of P in any case; in the linear case, with linear observations, it may be verified that our results specialize to those in [4]. This is, of course, the optimum filter for finite order Gaussian Markov processes.

3c. Filtering. The general problem here may be viewed as an optimum filtering problem, where $dx = fdt + Fdz$ represents the process, and dy is the nonlinear noisy observation. Then (20), or the equations for the moments, represent the form of the optimum filter, i.e., the simulation of (20) yields a running estimate of the conditional probability.

3d. Dependent observation noise. Up to now, the observation noise dw/dt has been white Gaussian. Assume $\beta = dw/dt$ is a correlated process and let it be represented as $d\beta = k(\beta)dt + d\epsilon$, where ϵ is a vector Brownian motion process with $E\delta\epsilon = 0$ and covariance $(\delta\epsilon) = \Sigma\Delta$. The observation is $b \overset{\Delta}{=} dy/dt = g(x) + \beta$. If the observation is considered to be

$$dy = dg + d\beta = (\dot{g} + k(\beta))dt + d\epsilon, \tag{22}$$

the previous theory may be applied: put $\dot{g} + k(\epsilon)$ whenever g appeared. Now, the distribution of β must also be estimated, and the x included the components of β. It appears to be typical of the estimation or filtering problem that, whenever observation noise is correlated, the noise as well as the quantity of interest must be estimated. The differentiation is not easy to simulate. If the observation is assumed to be $\beta + x + d\Psi/dt$, where Ψ is the Brownian motion, then by expanding the state vector x by adjoining β, the theory of the last section may be applied.

3e. Unknown system parameters or system order. Let γ be a constant parameter which either simply parametrizes f or determines the order of the system $dx = fdt + Fdz; f = f(x, \gamma)$. Let γ be given some initial distribution $P(\gamma, 0)$. Then, our results apply to the augmented system

$$dx = f(x, \gamma)dt + Fdz,$$

$$d\gamma = 0,$$

and we merely replace x by the vector $[x, \gamma]$ in the results. $P(a, \gamma, 0) = P(a, 0)P(\gamma, 0)$. This is a general solution to what has been called partial observability by some authors [5].

3f. Determination of the conditional moments. The moments

$$
\begin{aligned}
m_i &= \int a_i P(a, t \mid t)\, da, \\
m(j_1, \cdots, j_r, t) &= \int \prod_{i=1}^{r} (a_i - m_i)^{j_i} P(a, t \mid t)\, da, \qquad (23)\\
c(j_1, \cdots, j_r, t) &= \int \prod_{i=1}^{r} a_i^{j_i} P(a, t \mid t)\, da,
\end{aligned}
$$

satisfy ordinary differential equations, on the right hand side of which the observations appear linearly. The procedure is simple and we merely indicate it here.

We have

$$
(24) \quad dc(j_1, \cdots, j_r, t) = \int \prod_{i=1}^{r} a_i^{j_i} [P(a, t + dt \mid t + dt) - P(a, t \mid t)]\, da.
$$

Let $h(a) = \prod_{i=1}^{r} a_i^{j_i}$. Equating the left hand side of (16) and (17) yields

$$
\begin{aligned}
&\int h(a)[P(a, t + dt \mid t + dt) - P(a, t \mid t)]\, da \\
(25) \quad &= \int h(a) dQ\, da + \int \left[h_a'(a)\, d\bar{f}(a) + \frac{dt}{2} \sum_{i,j} h_{a_i a_j}(a) \bar{v}_{ij} \right] P(a, t \mid t)\, da.
\end{aligned}
$$

Upon performing the integration in (25), dc is obtained. This question is also discussed in [1].

3g. Applications to optimal stochastic control theory. The function $P(0, t)$ is a Markov process, and appears to be the most natural quantity which one may consider as the state variable of the differential system (1). To extend the form (1) to the optimal control formulation, write $dx = f(x, u)dt + F(x, u)dz$, where u is a control function which is to be determined so as to minimize some error criterion, say

$$
(26) \quad E \int_{t_0}^{T} k(x, u, t)\, dt = E \int_{t_0}^{T} \int P(a, t \mid t) k(a, u, t)\, da\, dt,
$$

where E is the expectation over all random variables. (See [1], [6], [7], [8].) Here the optimal control u^0 will be a functional of P.

It is possible to write a second order partial differential equation whose dependent variable is the minimum of (26) and whose independent variables are $P(a, t_0 \mid t_0)$ and t_0, and which yields many properties of u^0. This will not be done here. The equation is analogous to those appearing in [6], [7], [8].† The method of derivation is exactly that used in [1] for the scalar x and linear g case.

3h. Poisson z. The results may be extended to all dz for which the Chapman-Kolmogorov equation is valid; in particular, an equation for P may also be obtained when z is a Poisson process.

3i. The results have numerous applications to special problems in statistical communication theory; these will be considered elsewhere.

3j. Previous cases extended to case where $P(a,t_0)$ is concentrated at only finitely many points, and $x(t)$ is not necessarily generated by a differential equation. For the most general case, $P(a,t_0)$ is a sufficient statistic for control purposes; that is, the minimum of (26) can be written as a functional of $P(a, t_0 \mid t_0)$ for any t_0. When $F = 0$ and $P(a, t_0)$ is concentrated at only finitely many points, it is not usually convenient to take the point of view of §3g. Here, P is not differentiable with respect to a and (15) is a sum; $P(a, t + \Delta \mid \alpha, t)$ is either zero or is concentrated at only one point for any given α.

Although the formerly derived results are not valid for this case, an extremely simple extension is available—in fact, the extension is rigorously verifiable (it is essentially the case discussed in Appendix 2).

To view the results in a fairly general form, let us have a choice of n possible curves $x^i(t)$, $i = 1, \cdots, n$; the ith having conditional probability $p(i \mid t)$ at t. Each x^i could be the solution to the equation $\dot{x} = f(x)$ with a different initial condition, or with a different value of some parameter; or it could be an arbitrary signal function. The observation is $dy = g(a)dt + dw$, where a takes one of the values $x^i(t)$, $i = 1, \cdots, n$. We will write $g(i, t) = g(x^i(t))$.

The method is the following. Instead of keeping track of the arguments at which P is concentrated, as part of the procedure of generating P, we keep track of these arguments separately—and assume that the values of each $x^i(t)$ are available; thus, P is applied to the state i, $i = 1, \cdots, n$, which is not subject to dynamical changes. Carrying previous arguments over, we obtain

$$(27) \qquad dP(i \mid t) = P(i \mid t) \cdot (dy - Eg)' \Sigma^{-1} (g - Eg).$$

For this problem, (26) is rewritten as

$$(28) \qquad E \int_{t_0}^{T} \sum_i p(i \mid t) k(x^i, u, t)\, dt.$$

† Due to the presence of P, the equation contains functional derivatives as well as ordinary derivatives.

Equation (28) yields that the sufficient state variables for control purposes are all the $p(i \mid t)$ and their (effective) arguments $x^i(t)$ (occasionally some of the x^i can be derived from the others—and may be eliminated as state variables).

APPENDIX

The appendices contain several interesting facts and demonstrations relevant to our method of deriving the differential equations satisfied by certain stochastic processes, such as conditional probabilities. Appendix 1 contains some general remarks and in Appendix 2, the results are verified for some simple cases.

Appendix 1. We first discuss the meaning of the obtained stochastic equations by means of an example. Consider the scalar function $x = e^z$ where $z(t)$ is Brownian motion; $z(t) \sim N(0,\sigma^2 t)$. We are interested in a differential equation which represents x: since $z(t)$ is nowhere differentiable [3], the equation *cannot* be obtained in the usual formal manner. Consider

$$\delta x = e^{z+\delta z} - e^z = x(e^{\delta z} - 1) = x\left(\delta z + \frac{\delta z^2}{2} + \cdots\right). \tag{A.1}$$

Truncate the power series expansion and note that

$$E\left[\delta x - x\left(\delta z + \frac{\delta z^2}{2}\right)\right] = o(\Delta), \tag{A.2}$$

$$E\left[\delta x - x\left(\delta z + \frac{\delta z^2}{2}\right)\right]^2 = o(\Delta^2). \tag{A.3}$$

Thus, in the mean square sense, we have the differential equation

$$dx = x\left(dz + \frac{dz^2}{2}\right). \tag{A.4}$$

Note that, if the $dz^2/2$ term were omitted, (A.2) would be $O(\Delta)$ and (A.3) would be $O(\Delta^2)$; the errors would be of the order of dt, and the resulting solution would be meaningless.

Now, divide the time interval t into n equal sections and let $\Delta = t/n$. Let $\delta z_i = z((i+1)\Delta) - z(i\Delta)$. Thus a discrete approximation to (A.4) is

$$\delta x_i = x_i\left(\delta z_i + \frac{\delta z_i^2}{2}\right)$$

or

$$x_n = x_0 \prod_1^n \left(1 + \delta z_i + \frac{\delta z_i^2}{2}\right).$$

Now it is easily shown that

$$E[x_n - e^z] \to 0,\ E[x_n - e^z]^2 \to 0, \text{ as } \Delta \to 0,\ n \to \infty. \tag{A.5}$$

Thus, again in the mean square sense, (A.4) represents $x = e^z$. If the $dz^2/2$ terms were omitted, (A.5) would tend to some nonzero quantity. This holds in the general case also, since the truncation errors add linearly. Thus, the presence of the second order term dz^2 or $dy_i dy_j$ is justified.

There are some theorems in [3, pp. 286–291] which prove that, given a suitably regular continuous Markov process such as $x = e^z$, x has a representation of the form

$$\text{(A.6)} \qquad \begin{aligned} dx = E[x(t+dt) - x(t) \mid x(t)]dt \\ + E^{1/2}[(x(t+dt) - x(t))^2 \mid x(t)]du' \end{aligned}$$

where u is a Brownian motion process. The major problem appears to be the identification of the process u. Let $Edz^2 = \sigma^2 dt$. It may be shown here that $\sigma du = dz$, or

$$\text{(A.7)} \qquad dx = x\sigma^2 dt/2 + xdz.$$

In finite difference form $x_n = x_0 \prod_1^n (1 + \delta z_i + \sigma^2\Delta/2)$. It is verifiable, by direct computation, that

$$\text{(A.8)} \qquad E\left[\prod_1^n \left(1 + \delta z_i + \frac{\delta z_i^2}{2}\right) - \prod_1^n (1 + \delta z_i + \sigma^2\Delta/2)\right]^2 \to 0,$$

as $\Delta \to 0$, thus proving the validity of the replacement.

Now, we briefly discuss the nature and interpretation of $(dz)^i$, $i \geqq 2$. According to the derivation,

$$\text{(A.9)} \quad x(t) = \sum_1^n \delta x_i = \sum_1^n x_i \delta z_i + \sum_{i=1}^n x_i \frac{(\delta z_i)^2}{2} + \sum_{i=1}^n x_i \left(\frac{\delta z_i^3}{3!} + \cdots\right)$$

is an exact expression for $x(t)$. This suggests the integral

$$\text{(A.10)} \qquad \int (dz)^2,$$

which may be interpreted as the limit of Riemann sums. With this interpretation,

$$\int_0^t (dz)^2 = \sum (\delta z_i)^2.$$

In the limit as $\Delta \to 0$, $n\Delta = t$, $\sum(\delta z_i)^2$ tends to a constant, $\sigma^2 t$, with probability one and in mean square. With this definition of (A.10), the integral over any measurable t set may be defined and stochastic integrals of the form $\int x(dz)^2 = \int x(dt\,\sigma^2)$ considered. Similarly, for the higher terms $(dz)^i$, $i > 2$, whose integrals degenerate to zero with probability one and in mean square. Thus, the replacement of dz^2 by $\sigma^2 dt$ is again justified.

Appendix 2. Now, using the *limit of the Riemann sum* definition of the integrals we prove, in an indirect although instructive way, that the replacement of dy^2 by $dt\,\sigma^2$ is justified. We limit ourselves, only for simplicity, to a scalar case with no dynamics where x takes the values 0 or 1.†‡

Let $P(1,t) = P_1$. No generality is lost in letting $g(a) = a$. Here, $dP = dQ$ and, by rearranging (13), we obtain

$$dP = \frac{P}{2\sigma^2}\left[2(dy - m\,dt)(a - m) + \left(\frac{dy^2}{\sigma^2} - dt\right)\left((a - m)^2 - m_2\right)\right] + r, \tag{A.11}$$

where $Er \sim o(dt)$, $Er^2 \sim o(dt^2)$. Also $r(t) - Er(t)$ is an orthogonal process and orthogonal to any function of $P(a, t)$. Upon replacing dy^2 by its average value and neglecting r, we have

$$d\tilde{P} = \frac{\tilde{P}}{\sigma^2}\,[(dy - \tilde{m}\,dt)(a - \tilde{m})]. \tag{A.12}$$

Also

$$m = Ea = P_1\,,\ m_2 = E(a - Ea)^2 = (1 - P_1)P_1\,. \tag{A.13}$$

Thus, letting $\tilde{P}_1(1 - \tilde{P}_1) = k(\tilde{P}_1) = \tilde{k}$,

$$\begin{aligned} \sigma^2 dP_1 &= k[(dy - P_1 dt) + (dy^2/\sigma^2 - dt)(1 - 2P_1)/2] + r,\\ \sigma^2 d\tilde{P}_1 &= \tilde{k}(dy - \tilde{P}_1 dt) = \tilde{k}[(dy - P_1 dt) + (P_1 - \tilde{P}_1)dt], \end{aligned} \tag{A.14}$$

where $Er \sim o(dt)$, $Er^2 \sim o(dt^2)$.

Now, P_1 and $\tilde{P}_1$ are always in the interval [0, 1]. Letting $e_\tau = P(\tau) - \tilde{P}_1(\tau)$, we have the error

$$\begin{aligned} \sigma^2 e_t = \int_0^t (k - \tilde{k})(dy - P_1\,d\tau) + \int_0^t \tilde{k}e_\tau\,d\tau + \sigma^2 \int_0^t r\\ + \int_0^t k(dy^2/\sigma^2 - d\tau)(1 - 2P_1)/2. \end{aligned} \tag{A.15}$$

Note that $(dy - P_1 dt)$ and $(dy^2/\sigma^2 - dt)$ are orthogonal processes, and are orthogonal to any function of $P_1(\tau)$ or $\tilde{P}_1(\tau)$, $\tau \leqq t$. Using these facts,

$$\begin{aligned} \sigma^4 Ee_t^2 &= E\int_0^t (k - \tilde{k})^2\sigma^2\,d\tau + \int_0^t o(d\tau)\\ &+ E\int_0^t \tilde{k}e_\tau\,d\tau\left\{\int_0^t (k - \tilde{k})(dy - P_1\,d\tau) + \int_0^t k(dy^2/\sigma^2 - d\tau)(1 - 2P_1)/2\right\}. \end{aligned}$$

† Again, the reference quoted above implies some sort of replacement, but the Brownian motion in the differential equation is not identified. The technique here identifies all terms in terms of the observations and properties of the conditional densities.

‡ Generally, the replacement of second order terms by their average values is the easiest part to verify; it is more difficult to prove that our other limiting operations are valid.

Now, observing that k satisfies a Lipschitz condition for P in [0, 1] and using Schwartz's inequality on the last product of integrals yields

$$\text{(A.16)} \qquad Ee_t^2 \leqq K_1 \int_0^t Ee_\tau^2 \, d\tau + K_2 \left\{ \int_0^t Ee_\tau^2 \, d\tau \right\}^{1/2} + \int_0^\tau o(d\tau)$$

for some positive and finite K_1 and K_2. Thus $Ee_t^2 = 0$, since $P_1(0) = \tilde{P}_1(0)$, and the validity of (A.12) is proved.

In all cases checked, Doob's representation theorems (referred to in the text) yield equations of our form, where the observation noise process w is identified with the Brownian motion. For the problem of this Appendix, there are two families of stochastic processes. The first are the *family of actual sample functions* P_1, when $a = 1$ $(dy = dt + dw)$; the second when $a = 0$ $(dy = dw)$. Applying Doob's theorems to each of these yields the representation (A.12).

Acknowledgment. The author wishes to express his appreciation to Dr. W. M. Wonham for helpful and stimulating discussions on various aspects of the material discussed here.

REFERENCES

[1] H. J. KUSHNER, *On the dynamical equations of conditional probability density functions, with applications to optimal stochastic control theory*, J Math. Anal. Appl., 8(1964).

[2] R. L. STRATONOVICH, *Conditional Markov processes*, Theor. Probability Appl., 5(1962).

[3] J. L. DOOB, *Stochastic Processes*, Wiley, New York, 1953.

[4] R. E. KALMAN AND R. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83D(1961).

[5] J. J. FLORENTIN, *Partial observability and optimal control*, J. Electronics and Control, 13 (1962).

[6] ———, *Optimal control of continuous time, Markov, stochastic systems*, Ibid., 10 (1961).

[7] H. J. KUSHNER, *Optimal stochastic control*, IRE Trans. Automatic Control, AC-7(1962).

[8] M. W. WONHAM, *Stochastic problems in optimal control*, RIAS Report 63-14, (May 1963).

[9] H. J. KUSHNER, *Stochastic extremum problems, Part I, calculus; Part II, calculus of variations*, J. Math. Anal. Appl., to appear.

[10] T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1959.

J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1962

SOME TYPES OF OPTIMAL CONTROL OF STOCHASTIC SYSTEMS*

STUART E. DREYFUS†

1. Introduction. A stochastic system (i.e., a dynamic system involving random variables) which evolves according to a rule which also involves variables or parameters under external control is called a stochastic control system. If these variables or parameters are determined so that the system behaves as well as possible as measured by some well-defined criterion, one has achieved optimal control of the stochastic system.

Under varying assumptions concerning the information available to the controller, different optimal control policies result. In this paper we shall develop and illustrate several different control schemes and compare their behavior. In this way we intend to demonstrate that certain control philosophies that may appear superficially to be equivalent are really quite different. In the final section we derive asymptotic expressions for the cost of optimal control using several different schemes. This yields a quantitative measure of the vast superiority of feedback over open-loop control for a particular stochastic system.

2. A deterministic problem. Let us begin by considering a trivial three-stage discrete deterministic control problem. Given the directed network shown in Fig. 1, we wish to determine that path from point A to line B which has the minimal sum of the numbers written along the three arcs of the path.

Let us denote a decision to follow the diagonally-up arc from an intersection by U and the diagonally-down arc by D. By examining all eight possible paths from A to B, we discover that the path D-U-D (diagonally down, then up, then down) has sum-of-arc-numbers zero and is the unique optimal solution. We shall call such a designation of the solution, giving the sequence of control decisions to be followed from specific initial point to termination, the optimal *open-loop* control.

A second way of presenting the solution to this problem is to associate with each node of the figure a decision, either U or D, such that that decision is the initial one of the optimal path from the node to the terminal line. This set of decisions assigned to nodes is most efficiently determined recursively backwards from the terminal line [1]. We initially record the optimal decisions and minimal sum to termination (encircled) at the nodes

* Received by the editors July 22, 1963, and in final revised form April 21, 1964. This research is sponsored by the United States Air Force under Project RAND—contract No. AF 49(638)-700 monitored by the Directorate of Development Planning, Deputy Chief of Staff, Research and Development, Hq USAF.

† The RAND Corporation, 1700 Main Street, Santa Monica, California.

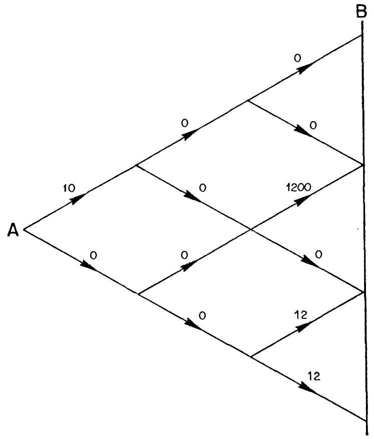


FIG. 1

along the line C in Fig. 2, and then use the circled numbers to determine the optimal decisions and sum along D and, finally, from A. The results are shown in Fig. 3. We shall call such a designation of the solution, giving the optimal decision associated with starting at each possible state of the system (i.e., at each node), the *feedback* optimal control.

The interpretation of Fig. 3 is that the optimal path starting at point A has sum zero and starts diagonally down. The node reached after making the downward move has a U written by it, indicating a decision to go diagonally up. This leads to a node with a down decision. Hence, D-U-D is the optimal path from A. Note that the feedback representation of the solution also yields the best path starting from other nodes not along the D-U-D path.

The important point is that for a specified initial point such as A, the open-loop and feedback solutions are equivalent for a deterministic process.

3. A stochastic problem. Let us now modify the above problem by introducing a stochastic aspect. We shall assume that the decision designated by U results in a probability of $\frac{3}{4}$ of moving diagonally up and $\frac{1}{4}$ of moving down. The alternative decision, D, has a $\frac{3}{4}$ chance of a diagonally downward move and a $\frac{1}{4}$ chance of an upward transition. We now have a stochastic control problem. We can still exert a controlling influence, but randomness determines the actual transformation of state.

As a criterion for comparing possible control schemes, let us attempt to minimize the expected sum along the path from A to line B.

To determine the best open-loop control policy, we consider all eight possible sequences of decisions and choose the one with minimal expected sum. For example, the decision sequence U-U-U has probability $\frac{27}{64}$ of

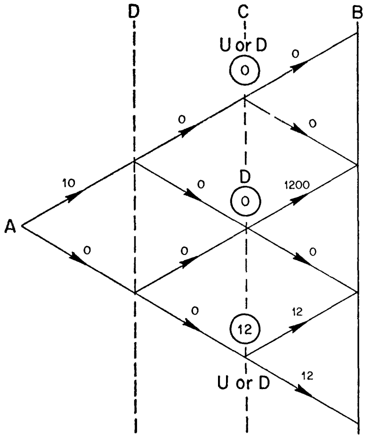


FIG. 2

actually yielding the path U-U-U with sum 10, $\frac{9}{64}$ probability of yielding the path D-U-U with sum 1200, etc. Multiplying the probabilities by the sums and adding, we get an expected sum E_{UUU} given by

$$E_{UUU} = \tfrac{27}{64}\cdot 10 + \tfrac{9}{64}\,(1200 + 1210 + 10) + \tfrac{3}{64}\,(10 + 0 + 12) + \tfrac{1}{64}\cdot 12 \cong 360.$$

It turns out that the sequence U-U-D has the minimal expected sum of approximately 120.

The best feedback control is computed recursively backwards just as in the deterministic example. Suppose that, for a given node, the expected sums starting at each of the two possible nodes to which one might go have been determined. Then the expected sum from the given node to the termination under decision U is obtained by multiplying the upward arc number plus the remaining expected sum associated with the node at the end of the up-arc by $\frac{3}{4}$ and adding $\frac{1}{4}$ times the corresponding downward numbers. Decision D is similarly evaluated reversing the $\frac{3}{4}$ and $\frac{1}{4}$, and the minimal expected sum is chosen. The minimizing decision and expected sum (encircled) are recorded at the node. This computation leads to Fig. 4. The expected sum using feedback control is $84\frac{1}{4}$ and the control policy is the set of letters associated with the nodes in Fig. 4.

At this point we would like to introduce a third control scheme. Let us use the optimal open-loop solution to yield our initial decision. Then, after a transition has occurred, let us observe the result and determine the best open-loop solution for the new two-stage problem. After implementing the initial control decision of this optimal open-loop solution, we again observe the state and use the optimal control decision for the remaining one-stage

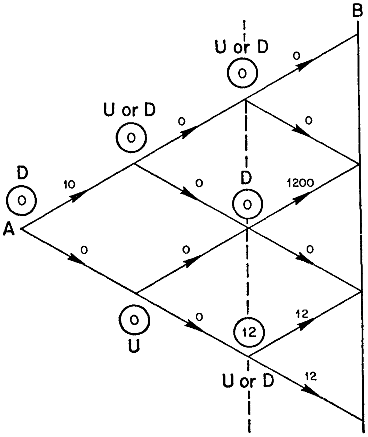


FIG. 3

problem. This scheme uses the optimal open-loop initial decision at each stage, but incorporates feedback in the observation of the actual state attained. We call this scheme *open-loop-optimal feedback* control.

This control scheme differs from either of the previous two. The initial optimal open-loop decision agrees with the feedback decision except for starting at node A. There, as has been shown, the optimal open-loop control dictates an upward decision. Therefore, the expected cost of the above scheme is $\frac{1}{4} \cdot 84 + \frac{3}{4} \cdot 85 = 84\frac{3}{4}$.

We can conclude from this example that

1) the pure open-loop scheme incorporating no use of subsequent information about actual transitions yields a large expected sum;

2) the pure feedback scheme where the state is assumed known when the decision is made yields the smallest possible expected sum for a stochastic problem;

3) the open-loop-optimal feedback scheme yields an intermediate expected sum. Although feedback is used, the fact that feedback is to be used is withheld from the computation determining the control decisions, which results in an inferior control scheme.

4. A continuous deterministic problem. Let us now consider briefly a standard continuous non-stochastic control problem. Given an initial time t_0 and final time T, we wish to use control $u(t)$, $t_0 \leqq t \leqq T$, so as to guide a particle, initially in state x_0, toward the origin $x = 0$. We attach a cost to using control and attempt to minimize the criterion function,

$$\int_{t_0}^{T} u^2(t)\, dt + x^2(T), \tag{4.1}$$

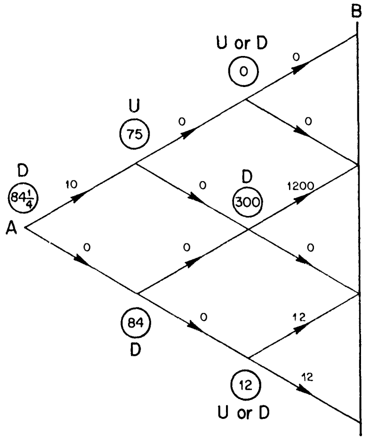


FIG. 4

where the first term represents the cost of control and the second term measures the deviation from the origin at the terminal time. Motion of the particle is given by the linear differential equation

$$\dot{x}(t) = ax(t) + bu(t). \tag{4.2}$$

This is a linear control problem with quadratic criterion and has been much analyzed. We consider it briefly here in order to acquaint the reader with the type of problem we shall consider subsequently and with the dynamic programming technique of solution.

The classical necessary conditions for an extremum of the above problem are given in terms of an adjoint variable or Lagrange multiplier λ which satisfies the equation

$$\dot{\lambda} = -a\lambda \tag{4.3}$$

and terminal condition

$$\lambda(T) = 2x(T). \tag{4.4}$$

The optimal control is given by the condition

$$2u + \lambda b = 0. \tag{4.5}$$

Solution of (4.3) with boundary condition (4.4) yields

$$\lambda(t) = 2x(T)e^{a(T-t)}, \tag{4.6}$$

and therefore

$$u(t) = -x(T)be^{a(T-t)}, \tag{4.7}$$

so $u(t)$ varies exponentially with time. The unknown terminal value of x, $x(T)$, can be expressed in terms of $x(t)$ by substituting the control rule (4.7) in (4.2) and solving. The resulting expression for $x(t)$ in terms of $x(T)$ can be inverted, and the control at time t is then given in terms of the state at time t by (4.7). Performing these steps we get

$$(4.8)\qquad x(t) = x(t_0)e^{a(t-t_0)} + \frac{x(T)b^2}{2a}e^{a(T-t)} - \frac{x(T)b^2}{2a}e^{a(T-2t_0+t)},$$

$$(4.9)\qquad x(t_0) = \left(1 - \frac{b^2}{2a} + \frac{b^2}{2a}e^{2a(T-t_0)}\right)e^{-a(T-t_0)}x(T),$$

$$(4.10)\qquad x(T) = \frac{e^{a(T-t)}x(t)}{1 - \frac{b^2}{2a} + \frac{b^2}{2a}e^{2a(T-t)}},$$

$$(4.11)\qquad u(t) = -\frac{be^{2a(T-t)}x(t)}{1 - \frac{b^2}{2a} + \frac{b^2}{2a}e^{2a(T-t)}}.$$

This is the feedback solution for control as a function of state. The optimal control is exponential in time, or, for a given time, it is a linear function of the state.

The dynamic programming solution of this problem proceeds as follows. Define an auxiliary function $f(x, t)$ as the minimal obtainable value of the criterion function (4.1) if we start the problem in state x at time t, $t_0 \leqq t \leqq T$. By the principle of optimality linking the initial decision with the remaining optimal decisions, we have

$$(4.12)\qquad f(x, t) = \min_{u(t)} [u^2(t)dt + f(x + (ax + bu)dt, t + dt)].$$

Expanding (4.12) in Taylor series, dividing by dt and letting dt approach 0, we get

$$(4.13)\qquad 0 = \min_u \left[u^2 + \frac{\partial f}{\partial x}(ax + bu) + \frac{\partial f}{\partial t}\right].$$

Differentiating with respect to u to minimize gives

$$(4.14)\qquad 2u + b\frac{\partial f}{\partial x} = 0,$$

and substituting u determined by (4.14) in (4.13), we obtain the nonlinear partial differential equation

$$(4.15)\qquad 0 = -\frac{b^2\left(\frac{\partial f}{\partial x}\right)^2}{4} + ax\frac{\partial f}{\partial x} + \frac{\partial f}{\partial t}.$$

Assuming $f(x, t)$ has the separable form $g(t)x^2$ and substituting in (4.15), we find that $g(t)$ satisfies the Riccati ordinary differential equation

$$-b^2g^2(t) + 2ag(t) + g'(t) = 0, \tag{4.16}$$

with

$$g(T) = 1. \tag{4.17}$$

Solution of this equation yields

$$g(t) = + \frac{e^{2a(T-t)}}{1 - \frac{b^2}{2a} + \frac{b^2}{2a} e^{2a(T-t)}}, \tag{4.18}$$

whence

$$f(x, t) = \frac{e^{2a(T-t)}x^2}{1 - \frac{b^2}{2a} + \frac{b^2}{2a} e^{2a(T-t)}}. \tag{4.19}$$

Substitution in (4.14) yields the control scheme

$$u(t) = -\frac{be^{2a(T-t)}}{1 - \frac{b^2}{2a} + \frac{b^2}{2a} e^{2a(T-t)}} x(t), \tag{4.20}$$

which agrees with (4.11). Again, as in §2, we see that for a deterministic problem the open-loop and feedback solutions are equivalent.

5. A continuous stochastic problem [2–5]. To construct a stochastic control problem, we attach a random variable to the equation defining the evolution of x. We write the discrete rule

$$x(t + \Delta t) = x(t) + [ax(t) + bu(t)]\Delta t + \xi(\Delta t), \tag{5.1}$$

where $\xi(\Delta t)$ is a stochastic process with, for all t,

(5.2) (1) $E(\xi(\Delta t)) = 0;$

(5.3) (2) $E(\xi^2(\Delta t)) = \sigma^2 \Delta t;$

(5.4) (3) $E(\xi^n(\Delta t)) = o(\Delta t), n > 2;$

(5.5) (4) $\xi(\Delta t_1), \cdots, \xi(\Delta t_n)$ are mutually independent for non-overlapping intervals $\Delta t_1, \cdots, \Delta t_n$,
where E is the expected value operator, σ^2 is a constant, and $x = o(\Delta t)$ means the limit as $\Delta t \to 0$ of $\frac{x}{\Delta t}$ is zero. In the conventional notation [6], $\xi(\Delta t)$ is written as the increment $\Delta z_t = z(t + \Delta t) - z(t)$, where $z(t)$ is

called a Brownian motion process. The limiting process as $\Delta t \to 0$ is the continuous control problem we shall consider. Our criterion function to be minimized is

$$E\left[\int_{t_0}^{T} u^2(t)\, dt + x^2(T)\right], \tag{5.6}$$

the expected cost of control plus terminal deviation.

The optimal open-loop control is deduced by considering all possible functions $u(t)$, $t_0 \leqq t \leqq T$, and choosing the one that minimizes the criterion (5.6). The cost of control integral is deterministic. Furthermore, if $x(T)$ is viewed, at the initial time t_0, as a random variable dependent upon $u(t)$, one notes that the variance $\sigma^2_{x(T)}$ of this random variable is independent of $u(t)$. Since the expected value of the square of a random variable is its mean squared plus its variance, we have

$$E(x^2(T)) = [E(x(T))]^2 + \sigma^2_{x(T)}, \tag{5.7}$$

so we wish to choose that $u(t)$ which minimizes

$$\int_{t_0}^{T} u^2\, dt + [E(x(T))]^2. \tag{5.8}$$

Due to the linearity of the equation of evolution (5.1), the expected value of $x(T)$ is the value of $x(T)$ that results from integrating (5.1) with forcing function $u(t)$ and with the stochastic process $\xi(\Delta t)$ replaced by its mean value at each time, zero. Hence, our problem reduces, for the special assumptions of linear equations and quadratic criterion, to precisely the deterministic problem that we solved in the previous section.

This observation leads to a fourth control scheme, called *certainty equivalent* control [7]. This scheme replaces the random variables in the stochastic problem by their expected values and solves the resulting deterministic control problem. Certainty equivalent control is seen to be equivalent to optimal open-loop control in the above example.

To obtain the open-loop-optimal feedback control for the above problem, we express the control as a function of state, as was done in (4.11), and use that control having observed the state transition. The actual realization of the control function then depends upon the realization of the stochastic process; one expects this scheme to perform better than the pure open-loop solution.

The pure feedback control law can be derived by dynamic programming. One defines $f(x, t)$ as the minimal value of (5.6), and writes

$$f(x, t) = \min_{u} \underset{\xi}{E}[u^2\Delta t + f(x + (ax + bu)\Delta t + \xi, t + \Delta t)]. \tag{5.9}$$

Hence, expanding in series and taking expectations using (5.2) through

(5.5),

$$0 = \min_u \left[u^2 + \frac{\partial f}{\partial x}(ax + bu) + \frac{1}{2}\sigma^2 \frac{\partial^2 f}{\partial x^2} + \frac{\partial f}{\partial t} \right]. \tag{5.10}$$

Therefore,

$$u = -\frac{b\dfrac{\partial f}{\partial x}}{2}, \tag{5.11}$$

and we must solve the equation

$$0 = -\frac{b^2\left(\dfrac{\partial f}{\partial x}\right)^2}{4} + ax\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 f}{\partial x^2} + \frac{\partial f}{\partial t}. \tag{5.12}$$

Letting

$$\begin{aligned} f(x, t) &= g(t)x^2 + h(t), \\ g(T) &= 1, \\ h(T) &= 0, \end{aligned} \tag{5.13}$$

we find that $g(t)$ satisfies the same equation, (4.16), as in the deterministic case. Since the optimal control only involves $g(t)$, we have the same control rule as in §4, but not the same expected cost, due to the $h(t)$ term reflecting the cost of the randomness. Hence, the optimal feedback control duplicates the open-loop-optimal feedback scheme.

These equivalences of various control schemes are unusual and are the result of our many assumptions of linearity and quadraticity. In the next section we shall modify the problem slightly and demonstrate the dissimilarity of the four different control philosophies we have distinguished.

6. Another continuous stochastic problem. We now modify the above problem slightly. We assume that the variance of $\xi(\Delta t)$ in (5.1) depends upon the control decision, with no randomness in the evolution of x if no control is exerted. This assumption reflects reality in many applications. We replace (5.3) by the equation

$$E(\xi^2(\Delta t)) = u^2\sigma^2\Delta t, \tag{6.1}$$

where σ^2 is a constant. We neglect the cost of control integral in the objective function (5.6), since the cost of control is now reflected in the uncertainty attendant upon the use of control. Our criterion function is now merely

$$E[x^2(T)]. \tag{6.2}$$

For simplicity, we take $a = 0$ in the equation of evolution (5.1), and use the continuous limit of

$$x(t + \Delta t) = x(t) + [bu(t)]\Delta t + \xi(\Delta t). \tag{6.3}$$

We first consider optimal open-loop control. The variance of the random variable $x(T)$ as viewed at time t_0 is

$$\int_{t_0}^{T} u^2(t)\sigma^2 \, dt, \tag{6.4}$$

and the criterion function equals

$$[E(x(T))]^2 + \int_{t_0}^{T} u^2\sigma^2 \, dt. \tag{6.5}$$

By the same reasoning as above, the expected value of $x(T)$ is the value yielded by replacing the stochastic process $\xi(\Delta t)$ at each time t by its mean, zero. We therefore have the same problem as in §4 and §5, except for a factor σ^2 in the criterion function and no ax term in the equation of motion. The adjoint variable $\lambda(t)$ is, in this case, a constant with terminal value $2E(x(T))$. The optimal control is given by

$$u(t) = -\frac{E(x(T))b}{\sigma^2}, \tag{6.6}$$

and is a constant function of time. Expressed in terms of state, we have

$$u(t) = -\frac{x(t)}{b\left(T - t + \dfrac{\sigma^2}{b^2}\right)}, \tag{6.7}$$

which, as before, is linear in the state at a given time. Using open-loop control, the expected terminal value of x, if we start at time t_0 in state $x(t_0)$, is

$$E[x(T)] = \frac{\sigma^2 x(t_0)}{b^2\left(T - t_0 + \dfrac{\sigma^2}{b^2}\right)}, \tag{6.8}$$

and the variance of the random variable $x(T)$ is given by

$$\sigma^2_{x(T)} = \frac{\sigma^2 x^2(t_0)(T - t_0)}{b^2\left(T - t_0 + \dfrac{\sigma^2}{b^2}\right)^2}. \tag{6.9}$$

Hence, the value of the criterion function is given by

$$E[x^2(T)] = [E(x(T))]^2 + \sigma^2_{x(T)} = \frac{\sigma^2 x^2(t_0)}{b^2\left(T - t_0 + \dfrac{\sigma^2}{b^2}\right)}. \tag{6.10}$$

We next analyze the open-loop-optimal feedback control scheme. This involves using the rule (6.7) for control as a function of state. The equation of motion becomes

$$x(t + \Delta t) = x(t) - \frac{x(t)}{T - t + \frac{\sigma^2}{b^2}} \Delta t + \xi(\Delta t). \tag{6.11}$$

If we define $f(x, t)$ as the expected value of $x^2(T)$ using the above rule, we have

$$f(x, t) = \underset{\xi}{E}\left[f\left(x - \frac{x\Delta t}{T - t + \frac{\sigma^2}{b^2}} + \xi, t + \Delta t \right) \right], \tag{6.12}$$

which, after series expansion, letting $\Delta t \to 0$, and taking the expectation, gives

$$0 = -\frac{x}{T - t + \frac{\sigma^2}{b^2}} \frac{\partial f}{\partial x} + \frac{x^2\sigma^2}{2b^2\left(T - t + \frac{\sigma^2}{b^2}\right)^2} \frac{\partial^2 f}{\partial x^2} + \frac{\partial f}{\partial t}. \tag{6.13}$$

Letting $f(x, t)$ have the form,

$$\begin{aligned} f(x, t) &= g(t)x^2, \\ g(T) &= 1, \end{aligned} \tag{6.14}$$

we obtain the linear homogeneous equation for $g(t)$,

$$g'(t) + \frac{1}{T - t + \frac{\sigma^2}{b^2}} \left[\frac{\sigma^2}{b^2\left(T - t + \frac{\sigma^2}{b^2}\right)} - 2 \right] g(t) = 0, \tag{6.15}$$

so that

$$f(x, t) = x^2 \exp\left\{ \int_t^T \frac{1}{T - \tau + \frac{\sigma^2}{b^2}} \left[\frac{\sigma^2}{b^2\left(T - \tau + \frac{\sigma^2}{b^2}\right)} - 2 \right] d\tau \right\} \tag{6.16}$$

$$= x^2 \exp\left\{ 1 - \frac{\sigma^2}{b^2\left(T - t + \frac{\sigma^2}{b^2}\right)} + 2 \log \frac{\sigma^2}{b^2} - 2 \log\left(T - t + \frac{\sigma^2}{b^2}\right) \right\}. \tag{6.17}$$

To evaluate the expected terminal x value, given that we start in state $x(t_0)$ at time t_0, we can solve equation (6.13) with solution of the form

$$\begin{aligned} f(x, t) &= g(t)x, \\ g(T) &= 1, \end{aligned} \tag{6.18}$$

obtaining

$$E[x(T)] = \frac{\sigma^2 x(t_0)}{b^2\left(T - t_0 + \frac{\sigma^2}{b^2}\right)}. \tag{6.19}$$

This result is the same as the pure open-loop result (6.8), which is explained by the linearity of the process.

Analysis of the feedback scheme begins with the definition of $f(x, t)$ as the value of the criterion if we start in state x at time t, $t_0 \leqq t \leqq T$, and use an optimal policy. By the principle of optimality, we have

$$f(x, t) = \min_u \mathop{E}_{\xi}[f(x + (bu)\Delta t + \xi, t + \Delta t)], \tag{6.20}$$

which yields

$$0 = \min_u \left[bu\frac{\partial f}{\partial x} + \frac{u^2\sigma^2}{2}\frac{\partial^2 f}{\partial x^2} + \frac{\partial f}{\partial t}\right]. \tag{6.21}$$

Hence, setting the derivative with respect to u equal to zero to minimize,

$$u = -\frac{b\frac{\partial f}{\partial x}}{\sigma^2\frac{\partial^2 f}{\partial x^2}}, \tag{6.22}$$

and, substituting (6.22) in (6.21),

$$0 = -\frac{b^2}{2\sigma^2}\frac{\left(\frac{\partial f}{\partial x}\right)^2}{\frac{\partial^2 f}{\partial x^2}} + \frac{\partial f}{\partial t}. \tag{6.23}$$

Setting

$$\begin{aligned} f(x, t) &= g(t)x^2, \\ g(T) &= 1, \end{aligned} \tag{6.24}$$

we get

$$0 = -\frac{b^2}{\sigma^2}g(t) + g'(t). \tag{6.25}$$

Solving for $g(t)$,

$$f(x, t) = e^{-(b^2/\sigma^2)(T-t)}x^2, \tag{6.26}$$

$$u = -\frac{bx}{\sigma^2}. \tag{6.27}$$

If we now define $h(x, t)$ to be the expected terminal x value starting in state x at time t and using control (6.27), we can characterize $h(x, t)$ by

$$h(x,t) = \underset{\xi}{E}\left[h\left(x - \frac{b^2x}{\sigma^2}\Delta t + \xi, t + \Delta t\right)\right], \tag{6.28}$$

where the boundary condition is now

$$h(x, T) = x. \tag{6.29}$$

Letting

$$\begin{aligned} h(x, t) &= g(t)x, \\ g(T) &= 1, \end{aligned} \tag{6.30}$$

we find

$$h(x, t) = e^{-(b^2/\sigma^2)(T-t)}x. \tag{6.31}$$

The final control philosophy we have mentioned above is certainty equivalent control, the optimal control for the deterministic system that results from replacing all random variables in the stochastic problem by their expected values. This yields the problem: choose $u(t)$ so that $x(T)$ given by

$$\begin{aligned} \dot{x}(t) &= bu(t), \\ x(t_0) &= x_0, \end{aligned} \tag{6.32}$$

minimizes the expression

$$x^2(T). \tag{6.33}$$

A little reflection shows that $x(T)$ can be made zero by any of an infinite class of controls, and the problem is therefore not meaningful.

We are now in a position to recapitulate our results. Foremost is the conclusion that the four different control schemes give four different optimal control rules. For open-loop control we have a rule given as a function of time and, naturally, dependent upon t_0, $x(t_0)$, and T. This rule, which never depends upon the realization of the stochastic process and which, in our particular example, is a constant function of time, is (by (6.6) and (6.8))

$$u(t) = -\frac{x(t_0)}{b\left(T - t_0 + \frac{\sigma^2}{b^2}\right)}. \tag{6.34}$$

The open-loop-optimal feedback control law is expressed as a function of current state and time and depends upon the realization of the stochastic process. It does not depend explicitly on the initial state or time. This law

(6.7) is

$$u(t) = -\frac{x(t)}{b\left(T - t + \frac{\sigma^2}{b^2}\right)}. \tag{6.35}$$

Note that this law is the same as (6.34) initially (for state $x(t_0)$ at time t_0) and that it duplicates (6.34) if and only if the stochastic process takes on its mean value, zero. The feedback control law depends on the current time and state, just as does the above scheme. However, due to the fact, stressed earlier, that the optimization mathematics is aware of the feedback nature of the control, we get a law different from (6.35), namely, (6.27),

$$u(t) = -\frac{bx(t)}{\sigma^2}, \tag{6.36}$$

which, in this particular case, does not happen to depend explicitly on the current time. The certainty equivalence concept, as noted earlier, is inappropriate here and yields no unique control law.

If we examine the asymptotic behavior of the criterion function for a long process ($T \rightarrow \infty$) starting at time zero in state x_0, we see that the expected value of $x^2(T)$ approaches zero in all cases. This is because for a long process very little control is exerted at any particular time, hence there is little randomness and we can steer assuredly toward the origin. The nature of the approach to zero as a function of the length of the process, T, is significant. For open-loop control the approach is inverse-linear, with, by (6.10),

$$E\,[x^2(T)] \sim \frac{\sigma^2 x_0^2}{b^2} T^{-1}. \tag{6.37}$$

For open-loop-optimal feedback control we have inverse-square convergence, with, by (6.17),

$$E[x^2(T)] \sim \frac{\sigma^4 e x_0^2}{b^2} T^{-2}. \tag{6.38}$$

Finally, the feedback control scheme yields negative-exponential convergence by (6.26):

$$E[x^2(T)] \sim e^{-(b^2/\sigma^2)T} x_0^2. \tag{6.39}$$

Both the open-loop and open-loop-optimal feedback schemes can be expected to reach the same terminal x value (see (6.8) and (6.19)), but due to its feedback nature, the latter scheme has less variance associated with it. The pure feedback control has an expected terminal value much closer to the origin (see (6.31)) since one can aim closer with the assurance that deviations resulting from the randomness caused by the greater control will be corrected later. Examining the control rules themselves for a fixed initial point, one finds that the pure feedback scheme calls for greater control. This can be explained by the fact that the feedback scheme can afford to aim closer to the origin in the assurance that overshooting due to randomness can be caught and corrected. While the open-loop-optimal feedback scheme will also catch and correct overshoot, the computation of the control rule is not cognizant of this fact and is, therefore, more conservative. Pure open-loop control, of course, will not compensate.

7. Conclusion. We see then that for any but the simplest stochastic problems, the various control philosophies that are equivalent for deterministic problems are quite dissimilar. Further, we have obtained some quantitative idea of the relative behavior and performance of several different optimal control schemes.

REFERENCES

[1] S. E. Dreyfus, *Dynamic programming*, Progress in Operations Research, vol. 1, R. L. Ackoff, ed., John Wiley, New York, 1961, Chap. 5.

[2] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, New Jersey, 1961.

[3] J. J. Florentin, *Optimal control of continuous time, Markov, stochastic systems*, J. Electronics Control, 10 (1961), pp. 473–488.

[4] H. J. Kushner, *Optimal stochastic control*, Correspondence, IRE Trans. on Automatic Control, October 1962, pp. 120–122.

[5] W. H. Fleming, *Some Markovian optimization problems*, J. Math. Mech., 12 (1963), pp. 131–140.

[6] J L. Doob, *Stochastic Processes*, John Wiley, New York, 1953, pp. 96-98.

[7] H. Theil, *A note on certainty equivalence in dynamic planning*, Econometrica, 25 (1957), pp. 346–349.

J.SIAM Control
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1962

ERRATA: A SOLUTION OF THE GODDARD PROBLEM*

BORIS GARFINKEL

Page 366. In (86), replace $\bar{\omega} > \omega_{\max}$ by $\bar{\omega} < \omega_{\max}$.

Page 366. In lines 5 and 6, replace the sentence beginning with the word "From" by the following. From (86) and Lemma 3, $\bar{v} \leqq v(x) < u(x)$, so that, by (83), $g_v < 0$ for all values of v between $\bar{v}$ and $v(x)$.

* This Journal, 1 (1963), pp. 349–368. Received by the editors April 13, 1964.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

ON THE EXISTENCE OF OPTIMAL FEEDBACK CONTROLS. II.*

T. F. BRIDGLAND, JR.†

1. Introduction. As has been noted recently [1], [2], two of the major problems of optimal control theory remain unsolved except in special cases. These problems are the ones of *existence* and of *synthesis*. Generally speaking, the problem of existence, given a control system and a performance criterion, involves the determination of conditions sufficient to ensure the existence of a control function which is optimal relative to the criterion. The synthesis problem requires the expression of the optimal control—granting its existence—as a function of the state of the control system. This function is called an optimal (feedback) control law.

In [3], a combined approach to the existence and synthesis problems is developed by means of a generalization of a technique originated by Carathéodory [4] and recently expounded in connection with optimal control by Kalman [5]. In order that we may develop the central purpose of the present paper, let us outline here those results of [3] which pertain to existence and synthesis.

Given a control system, represented mathematically by a vector differential equation

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

together with a specified set of control functions, $u(t)$, it is shown in [3] that if $L(t, x, u)$ is a functional possessing the property of determinacy, i.e., if there is a unique function, $\varphi_0(t, x)$, for which both

$$L(t, x, \varphi_0(t, x)) = 0$$

and

$$L(t, x, u) > 0, \qquad u \neq \varphi_0(t, x),$$

are satisfied, then $\varphi_0(t, x)$ is the unique optimal feedback control law for the control system in the sense that, along the trajectory, $\bar{x}(t; t_0, x_0)$, of the feedback system

$$\dot{x} = f(t, x, \varphi_0(t, x)),$$

we have

$$\int_{t_0}^{t^{\#}} L(\tau, \bar{x}(\tau; t_0, x_0), \varphi_0(\tau, \bar{x}(\tau; t_0, x_0)))\, d\tau = 0,$$

* Received by the editors March 20, 1964.

† Department of Mathematics, University of South Carolina, Columbia, South Carolina.

whereas, along any other trajectory, $x(t; t_0, x_0, u)$, of the control system, we have

$$\int_{t_0}^{t^\#} L(\tau, x(\tau; t_0, x_0, u), u(\tau))\, d\tau > 0.$$

In these expressions t_0 represents the *initial time*, x_0 the *initial state* of the control system, and $t^\#$ an appropriate *final time* depending on t_0, x_0, and u.

Now in any given problem, the likelihood that $L(t, x, u)$—the choice of which usually is determined by extramathematical factors—will satisfy the determinacy conditions is slight. However, as pointed out in the final remarks of [3], if a gauge function, $V(t, x)$, can be determined in such a way that $L^*(t, x, u) \equiv V^+(t, x; u) + L(t, x, u)$ has the determinacy property, where V^+ is a certain generalized total derivative of V, then the $\varphi_0(t, x)$ so determined is the unique optimal feedback control law relative to the criterion

$$\int_{t_0}^{t^\#} L(\tau, \bar{x}(\tau; t_0, x_0, u), u(\tau))\, d\tau.$$

Under the highly restrictive assumption that $V(t, x)$ is continuously differentiable with respect to both independent variables, Kalman [5] showed that $V(t, x)$ can be found as a solution to a Hamilton-Jacobi differential equation. It turns out that $V(t, x)$ is then given by

$$V(t, x) = \int_{t}^{t^\#} L(\tau, \bar{x}(\tau; t, x), \varphi_0(\tau, \bar{x}(\tau; t, x)))\, d\tau.$$

However, Pontryagin and his collaborators [1] have given several examples of basic problems of optimal control in which $V(t, x)$, as determined by the above, does *not* possess the strong differentiability properties required by Kalman's approach. Nonetheless, this representation of $V(t, x)$ has considerable appeal not only from the standpoint of the insight it conveys but also by virtue of the fact that the application of dynamic programming to optimal control problems rests upon the existence of such a representation for V.

Once the assumption of continuous differentiability for V is abandoned, it is still possible to consider a generalized Hamilton-Jacobi equation (the first of the determinacy conditions),

$$V^+(t, x; \varphi_0(t, x)) + L(t, x, \varphi_0(t, x)) = 0,$$

and it is with the construction of a solution, of the desired form, of such a generalized equation that this paper is concerned. Our construction requires the introduction of a vector $p(t, x)$ which corresponds to the Lagrange multiplier of classical methods. We shall show that if $p(t, x)$ satisfies a certain total differential equation—closely related to the ordinary differential equation of Pontryagin's maximum principle—as well as an appropriate transversality condition, then a unique optimal feedback control law, $\varphi_0(t, x)$, exists and $V(t, x)$ has the desired form. In addition, we discuss a method of finding $p(t, x)$. Our notation and terminology coincide with that of [3] and a reasonable familiarity with that paper is assumed.

2. Problem formulation. Let U comprise the totality of measurable functions on I which take values in a given subset $\Phi \subset R^m$. Consider the differential equation

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

where x, f are vectors in R^n. We assume the following properties for $f(t, x, \varphi)$:

(i) for each bounded subset D of R^n, $f(t, x, \varphi)$ is bounded on $I \times D \times \Phi$ and, for each $u \in U$, $f(t, x, u(t))$ is measurable in t for each x and continuous in x for each t;

(ii) the Jacobian matrix $f_x(t, x, \varphi)$ exists and is bounded on $I \times D \times \Phi$ for each bounded $D \subset R^n$.

The local existence and uniqueness of solutions of (1) is assured by (i), (ii); we assume further

(iii) each solution of (1) can be continued to all of I.

These three conditions ensure that (1) is of class A [3].

We assume the existence of a function $t^\#(t, x, u)$—the final time—on $I \times R^n \times U$ to $\bar{I}(t)$, satisfying

(iv) $t^\#(t, x(t; t_0, x_0, u), u) = t^\#(t_0, x_0, u)$, $\quad t_0 \leqq t < t^\#(t_0, x_0, u)$, and we define the set B, as in [3], by

$$(2) \qquad B = \{(t, x) \in I \times R^n \mid t^\#(t, x, u) > t \quad \text{for some} \quad u \in U\}.$$

The set of all $u \in U$ for which the defining property of B is satisfied will be denoted by $U(t, x)$.

Let $L(t, x, \varphi)$ be a function on $I \times R^n \times \Phi$ to R^1 for which

(v) $L(t, x, \varphi)$ is bounded on $I \times D \times \Phi$ for each bounded $D \subset R^n$ and, for each $u \in U$, $L(t, x, u(t))$ is measurable in t for each x and continuous in x for each t;

(vi) the gradient vector $L_x(t, x, \varphi)$ exists and is bounded on $I \times D \times \Phi$ for each bounded $D \subset R^n$.

We shall call a function on $I \times R^n$ to R^n a *gauge vector* if each of its components is a gauge function [3]. Now let us suppose there is a gauge vector $p(t, x)$ on $I \times R^n$ to R^n for which

(vii) there exists a function $\varphi_0(t, x)$ on $I \times R^n$ to Φ such that $\varphi_0(t, x(t))$ is measurable for every continuous $x(t)$;

(viii) in $I \times R^n$, the condition $\varphi \neq \varphi_0(t, x)$ implies[1]

$$L(t, x, \varphi) + p(t, x)\bullet f(t, x, \varphi) > L(t, x, \varphi_0(t, x)) + p(t, x)\bullet f(t, x, \varphi_0(t, x)).$$

Defining $\mathcal{L}(t, x)$, $\mathcal{F}(t, x)$ by

$$\mathcal{L}(t, x) \equiv L(t, x, \varphi_0(t, x)),$$

$$\mathcal{F}(t, x) \equiv f(t, x, \varphi_0(t, x)),$$

we assume that $\mathcal{L}$, $\mathcal{F}$ are measurable in t for each x, continuous in x for each t, and that for $(t, x_i) \in D$, $i = 1, 2$,

$$\text{(ix)} \qquad | \mathcal{L}(t, x_2) - \mathcal{L}(t, x_1)| \leqq \Lambda(D)\| x_2 - x_1 \|,$$

$$\text{(x)} \qquad \| \mathcal{F}(t, x_2) - \mathcal{F}(t, x_1)\| \leqq \Lambda(D)\| x_2 - x_1 \|$$

for each bounded $D \subset R^n$. It is a consequence of (x) and our previous assumptions on f that

$$\text{(3)} \qquad \dot{x} = \mathcal{F}(t, x)$$

is of class A.

Now define the set $\tilde{B}$, as in [3], by

$$\text{(4)} \qquad \tilde{B} = \{(t_0, x_0) \in B \mid \varphi_0(t, \bar{x}(t; t_0, x_0)) \in U(t_0, x_0)\},$$

where $\bar{x}(t; t_0, x_0)$ is the solution of (3). For $\tilde{B}$ we assume

(xi) $\tilde{B}$ is an $(n + 1)$-cell: $\tilde{B} = \{(t, x)| \, 0 \leqq t < T; a_i < x_i < b_i\}$.

We may define a functional $V(t, x)$ on $I \times R^n$ by

$$\text{(5)} \qquad \begin{aligned} V(t, x) &= \int_t^{\bar{t}(t,x)} \mathcal{L}(\tau, \bar{x}(\tau; t, x)) \, d\tau, && (t, x) \in \tilde{B}, \\ &= 0, && \text{elsewhere,} \end{aligned}$$

where $\bar{t}(t, x)$ is defined by

$$\bar{t}(t_0, x_0) \equiv t^{\#}(t_0, x_0, u_0(t; t_0, x_0)).$$

We assume

(xii) $\bar{t}(t, x)$ is a gauge function on $\tilde{B}$ having a nonnegative Y-derivate.

It is a consequence of (iii), (iv) that, if $(t_0, x_0) \in \tilde{B}$, then $(t, \bar{x}(t; t_0, x_0)) \in \tilde{B}$ for all $t \in [t_0, \bar{t}(t_0, x_0))$. Hence, from (5) there is obtained

$$\text{(6)} \quad V(t, \bar{x}(t; t_0, x_0)) = \int_t^{\bar{t}(t_0,x_0)} \mathcal{L}(\tau, \bar{x}(\tau; t_0, x_0)) \, d\tau, \quad 0 \leqq t < \bar{t}(t_0, x_0).$$

Now suppose that for fixed $(t_0, x_0) \in \tilde{B}$, $u(t)$ is an arbitrary control in $U(t_0, x_0)$; then by virtue of (iii), (xi) and the continuity of $x(t; t_0, x_0, u)$,

[1] The notation $p(t, x)\bullet f(t, x, \varphi)$ denotes the scalar product of p and f.

$(t, x(t; t_0, x_0, u)) \in \tilde{B}$ for all $t \in [t_0, T')$ for some T' such that $t_0 < T' \leqq T$. Hence, again there is obtained from (5)

$$V(t, x(t; t_0, x_0, u)) = \int_t^{\bar{t}(t, x(t; t_0, x_0, u))} \mathfrak{L}(\tau, \bar{x}(\tau; t, x(t; t_0, x_0, u)))\, d\tau, \tag{7}$$
$$t_0 \leqq t < T'.$$

Note. In future arguments we shall frequently use the abbreviated notation $x_u(t)$ in place of $x(t; t_0, x_0, u)$.

In the sequel, we shall show that, under appropriate conditions, $V(t, x)$ as defined in (5) is a gauge function on $\tilde{B}$. We then utilize the results of [3] to show that $\varphi_0(t, x)$ is the unique optimal feedback control law relative to the criterion

$$\int_{t_0}^{t^{\#}(t_0, x_0, u)} L(\tau, x_u(\tau), u(\tau))\, d\tau.$$

These results are contained in Theorem 1 below. Before stating this theorem, however, it will be convenient to establish a few lemmas.

3. Fundamental lemmas.

LEMMA 0. *Let $J = [a, b]$ be a closed interval in I and let $f(\tau, t)$ be a real-valued function defined on $J \times J$ which is integrable in τ for each t, continuous in t for each τ and which satifies*

$$| f(\tau, t_2) - f(\tau, t_1) | \leqq M(\tau) \, | t_2 - t_1 |,$$

where $M(\tau)$ is an integrable function of τ; then for the function $F(\tau, t)$ defined on $J \times J$ by

$$F(\tau, t) = \int_a^\tau f(\lambda, t)\, d\lambda,$$

it follows that the partial derivative, $F_\tau(\tau, t)$, satisfies $F_\tau(t, t) = f(t, t)$ almost everywhere on J.

Proof. For each $t \in J$, there is a set N_t for which $\mu_0(N_t) = 0$ such that

$$F_\tau(\tau, t) = f(\tau, t), \qquad \tau \in J - N_t.$$

Let ρ be the set of rationals in J; for the set $P = \bigcup_{t \in \rho} N_t$ we have $\mu_0(P) = 0$. Now let t be an arbitrary but fixed point in $J - P$ and let $\{t_n\}$ be a sequence of points in ρ having t as a limit. From the Lipschitz condition and the estimate

$$\left| h^{-1} \int_t^{t+h} f(\lambda, t)\, d\lambda - f(t, t) \right| \leqq \left| h^{-1} \int_t^{t+h} [f(\lambda, t) - f(\lambda, t_n)]\, d\lambda \right|$$
$$+ \left| h^{-1} \int_t^{t+h} f(\lambda, t_n)\, d\lambda - f(t, t_n) \right| + | f(t, t_n) - f(t, t) |,$$

we obtain

$$\left| h^{-1} \int_t^{t+h} f(\lambda, t)\, d\lambda - f(t, t) \right| \leqq | t_n - t | \left| h^{-1} \int_t^{t+h} M(\lambda)\, d\lambda \right|$$
$$+ \left| h^{-1} \int_t^{t+h} f(\lambda, t_n)\, d\lambda - f(t, t_n) \right| + | f(t, t_n) - f(t, t) |,$$

from which the conclusion of the lemma follows readily.

Our next result, which plays a fundamental role in the remainder of our arguments, is a generalization of the classical rule of Leibnitz for differentiation of an integral with respect to a parameter.

LEMMA 1. *Let J_0 and J_1 be closed intervals in I and let $f(\tau, t)$ be a real-valued function defined on $J_0 \times J_1$ which is measurable in τ for each t, continuous in t for each τ, and bounded by an integrable function of τ on $J_0 \times J_1$; let $\alpha(t)$, $\beta(t)$ be absolutely continuous, nondecreasing*[2] *functions on J_1 to J_0 for which $\alpha < \beta$. If, for almost every $t \in J_1$, $f_t(\tau, t)$ exists for almost all $\tau \in J_0$ and is bounded on $J_0 \times J_1$, then the function $v(t)$, defined by*

$$v(t) = \int_{\alpha(t)}^{\beta(t)} f(\tau, t)\, d\tau,$$

is absolutely continuous and, for almost all $t \in J_1$,

$$\dot{v}(t) = f(\beta(t), t)\dot{\beta}(t) - f(\alpha(t), t)\dot{\alpha}(t) + \int_{\alpha(t)}^{\beta(t)} f_t(\tau, t)\, d\tau.$$

Proof. The absolute continuity of $v(t)$ is readily verified; we omit the verification and show that the derivative on the right v' has the form indicated for $\dot{v}$. For $h > 0$ we find

$$(\gamma) \quad \begin{aligned} h^{-1}[v(t+h) - v(t)] &= h^{-1} \int_{\alpha(t)}^{\beta(t)} [f(\tau, t+h) - f(\tau, t)]\, d\tau \\ &+ h^{-1} \int_{\beta(t)}^{\beta(t+h)} [f(\tau, t+h) - f(\tau, t)]\, d\tau \\ &- h^{-1} \int_{\alpha(t)}^{\alpha(t+h)} [f(\tau, t+h) - f(\tau, t)]\, d\tau \\ &+ h^{-1} \left\{ \int_{\beta(t)}^{\beta(t+h)} f(\tau, t)\, d\tau - \int_{\alpha(t)}^{\alpha(t+h)} f(\tau, t)\, d\tau \right\}. \end{aligned}$$

The first term on the right of (γ) tends to $\int_{\alpha(t)}^{\beta(t)} f_t(\tau, t)\, d\tau$ as $h \to 0$ (cf. [6, p. 217]). The second and third terms on the right of (γ) tend to zero

[2] A sufficient condition for this is that the upper right-hand derivates of α, β be nonnegative almost everywhere [6, p. 207].

with h by virtue of the continuity of α, β and the boundedness of $f_t(\tau, t)$. By [6, p. 211], the final terms in (γ) may be replaced by

$$(\delta) \qquad h^{-1} \int_t^{t+h} g(\tau, t)\, d\tau,$$

where the function $g(\tau, t)$, defined by

$$g(\tau, t) \equiv f(\beta(\tau), t)\dot{\beta}(\tau) - f(\alpha(\tau), t)\dot{\alpha}(\tau),$$

satisfies the hypotheses of Lemma 0 on $J_1 \times J_1$. Hence, the consequent of that Lemma implies that the expression in (δ) tends to $g(t,t)$ as $h \to 0$ for almost all $t \in J_1$.

LEMMA 2. *For fixed* $\tau > 0$, *the solution*, $\bar{x}(\tau; t, x)$, *of* $dx/d\tau = \mathfrak{F}(\tau, x)$ *is a gauge vector on the set* $S = \{(t, x) \mid 0 \leqq t < \tau, x \in D\}$, *where D is any closed sphere in* R^n.

Proof. We have

$$\bar{x}(\tau; t, x) = x + \int_t^\tau \mathfrak{F}(\lambda, \bar{x}(\lambda; t, x))\, d\lambda;$$

hence, we may obtain the estimate

$$\begin{aligned} \| \bar{x}(\tau; t_2, x_2) - \bar{x}(\tau; t_1, x_1) \| \leqq & \{ \| x_2 - x_1 \| \\ & + | t_2 - t_1 | \sup_{[t_1, t_2]} \| \mathfrak{F}(\lambda, \bar{x}(\lambda; t_2, x_2)) \| \} \\ & + \left| \int_{t_1}^{\tau} \Lambda(D) \| \bar{x}(\lambda; t_2, x_2) - \bar{x}(\lambda; t_1, x_1) \| \, d\lambda \right| . \end{aligned}$$

From this estimate, there follow by virtue of the Bellman-Gronwall lemma the inequalites

$$(8) \qquad \begin{aligned} & \| \bar{x}(\tau; t + h, x + k) - \bar{x}(\tau; t, x) \| \\ & \leqq \{ \| k \| + | h | \sup_{[t, t+h]} \| \mathfrak{F}(\lambda, \bar{x}(\lambda; t + h, x + k)) \| \} \\ & \qquad \cdot \exp \Lambda(D) | \tau - t |; \end{aligned}$$

$$(9) \qquad \| \bar{x}(\tau; t, x_2) - \bar{x}(\tau; t, x_1) \| \leqq \| x_2 - x_1 \| \exp \Lambda(D) \, | \tau - t |;$$

$$(10) \qquad \begin{aligned} & \| \bar{x}(\tau; t_n + \delta_n, x_n) - \bar{x}(\tau; t_n, x_n) \| \\ & \leqq \delta_n \sup_{[t_n, t_n + \delta_n]} \| \mathfrak{F}(\lambda, \bar{x}(\lambda; t_n + \delta_n, x_n)) \| \exp \Lambda(D) | \tau - t_n |. \end{aligned}$$

By virtue of (i), $\sup \| \mathfrak{F} \|$ may be replaced in these inequalities by $\sup_{I \times D \times \Phi} \| f(t, x, \varphi) \|$, and $| \tau - t |$ by a bound of appropriate magnitude. The assertion of the lemma then follows by virtue of the definition of gauge function [3], continuity being given by (8), the local Lipschitz property by (9) and uniform absolute continuity (acu) by (10).

LEMMA 3. *For fixed* $\tau > 0$, $\mathfrak{F}(\tau, \bar{x}(\tau; t, x))$ *is a gauge vector and* $\mathfrak{L}(\tau, \bar{x}(\tau; t, x))$ *a gauge function on the set* S *of Lemma* 2.

Proof. The assertion is a direct consequence of (ix), (x), and an argument similar to that of Lemma 2.

LEMMA 4. *The function* $V(t, x)$ *defined in* (5) *is a gauge function on* $\tilde{B}$.

Proof. For (t_1, x_1), $(t_2, x_2) \in \tilde{B}$, the following estimate is readily obtained from (5):

$$| V(t_2, x_2) - V(t_1, x_1) | \leqq \{ \| x_2 - x_1 \| + | t_2 - t_1 | \sup_{[t_1, t_2]} | \mathfrak{L}(\lambda, \bar{x}(\lambda; t_2, x_2)) | \}$$

$$\cdot \Lambda(D) \int_{t_1}^{\bar{t}(t_1, x_1)} \exp (\Lambda(D) \, | \tau - t_1 |) \, d\tau$$

$$+ \, | t_2 - t_1 | \sup_{[t_1, t_2]} | \mathfrak{L}(\lambda, \bar{x}(\lambda; t_2, x_2)) |$$

$$+ \, | \bar{t}(t_2, x_2) - \bar{t}(t_1, x_1) | \sup_{[\bar{t}(t_1, x_1), \bar{t}(t_2, x_2)]} | \mathfrak{L}((\tau, \bar{x}(\tau; t_2, x_2)) |.$$

By an argument similar to that for Lemma 2, the conclusion follows from this estimate in conjunction with (v) and (xii).

Now let us define a function $q(\tau, t)$ by

$$(11) \qquad q(\tau, t) \equiv \frac{\partial}{\partial t} \bar{x}(\tau; t, x_u(t));$$

of course, q also depends on (t_0, x_0), but this will be fixed in any particular argument. The existence of $q(\tau, t)$ for almost all $t \in [t_0, T')$ and each $\tau \in [t_0, T)$ is guaranteed by Lemma 2 and [3, Lemma 4]. Since, for each $t \in [t_0, T')$ and each $\tau \in [t_0, T)$,

$$(12) \qquad \bar{x}(\tau; t, x_u(t)) = x_u(t) + \int_t^\tau \mathfrak{F}(\lambda, \bar{x}(\lambda; t, x_u(t))) \, d\lambda,$$

an estimate, similar to (8), based on (12) shows that $q(\tau, t)$ is bounded uniformly on $[t_0, T) \times [t_0, T')$. As a consequence of this boundedness and (x), we conclude by a similar estimate that $\frac{\partial}{\partial t} \mathfrak{F}(\tau, \bar{x}(\tau; t, x_u(t)))$ is bounded uniformly on $[t_0, T) \times [t_0, T')$; that this latter derivative exists is a consequence of Lemma 3 and [3, Lemma 4]. An application of Lemma 1 permits us to write

$$(13) \qquad \begin{aligned} q(\tau, t) &= f(t, x_u(t), u(t)) \\ &\qquad - \mathfrak{F}(t, x_u(t)) + \int_t^\tau \frac{\partial}{\partial t} \mathfrak{F}(\lambda, \bar{x}(\lambda; t, x_u(t))) \, d\lambda \end{aligned}$$

for almost all $t \in [t_0, T')$ and each $\tau \in [t, T)$. A partial differentiation of (13) with respect to τ then permits the statement of the next lemma.

LEMMA 5. *For almost every $t \in [t_0, T')$, $q(\tau, t)$ satisfies*

$$\frac{\partial}{\partial \tau} q(\tau, t) = \frac{\partial}{\partial t} \mathcal{F}(\tau, \bar{x}(\tau; t, x_u(t))),$$

$$q(t, t) = f(t, x_u(t), u(t)) - \mathcal{F}(t, x_u(t)),$$

almost everywhere on $[t, T)$.

LEMMA 6. (a) *For each $(t_0, x_0) \in \tilde{B}$,*

$$\frac{d}{dt} V(t, \bar{x}(t; t_0, x_0)) = -\mathcal{L}(t, \bar{x}(t; t_0, x_0))$$

almost everywhere on $[t_0, \bar{t}(t_0, x_0))$;

(b) *For each $(t_0, x_0) \in \tilde{B}$,*

$$\frac{d}{dt} V(t, x_u(t)) = -\mathcal{L}(t, x_u(t)) + \mathcal{L}(\bar{t}(t, x_u(t)), \bar{x}(\bar{t}(t, x_u(t)); t, x_u(t)))$$
$$\cdot \frac{d}{dt} \bar{t}(t, x_u(t)) + \int_t^{\bar{t}(t, x_u(t))} \frac{\partial}{\partial t} \mathcal{L}(\tau, \bar{x}(\tau; t, x_u(t)))\, d\tau$$

almost everywhere on $[t_0, T')$.

The proof of (a) consists merely of differentiating (6), whereas (b) is a consequence of application to (7) of Lemma 1, together with (xii) and Lemma 3. In the latter case, the proof is similar to that for Lemma 5 and for this reason we omit it.

Consider the functional $k(t, x, \varphi)$ defined by

$$(14) \qquad \begin{aligned} k(t, x, \varphi) \equiv L(t, x, \varphi) + p(t, x) \cdot f(t, x, \varphi) \\ - \mathcal{L}(t, x) - p(t, x) \cdot \mathcal{F}(t, x); \end{aligned}$$

then for all $(t, x) \in I \times R^n$, $k(t, x, \varphi)$ has a minimum on Φ at $\varphi_0(t, x)$. Let $(t_0, x_0) \in \tilde{B}$ and, for arbitrary fixed $u \in U(t_0, x_0)$, let σ satisfy $t_0 < \sigma < T'$. Define $\xi \equiv x_u(\sigma)$ and $\varphi^*(t, \sigma, \xi) \equiv \varphi_0(t, \bar{x}(t; \sigma, \xi))$. Then for fixed $\tau \in (\sigma, T)$, the function $k(\tau, \bar{x}(\tau; t, x_u(t)), \varphi^*(\tau, \sigma, \xi))$ has a minimum at $t = \sigma$; hence,

$$(15) \qquad \frac{\partial}{\partial t} k(\tau, \bar{x}(\tau; t, x_u(t)), \varphi^*(\tau, \sigma, \xi)) \,|_{t=\sigma} = 0$$

for each σ for which this derivative exists. Thus we obtain

LEMMA 7. *For each* $(t_0, x_0) \in \tilde{B}$ *and each* $\tau \in (t, T)$,

$$\frac{\partial}{\partial t} \mathcal{L}(\tau, \bar{x}(\tau; t, x_u(t))) = L_x(\tau, \bar{x}(\tau; t, x_u(t)), \varphi_u(\tau, t)) \cdot q(\tau, t)$$

$$(16) \qquad + p(\tau, \bar{x}(\tau; t, x_u(t))) \cdot [f_x(\tau, \bar{x}(\tau; t, x_u(t)), \varphi_u(\tau, t)) q(\tau, t) - \frac{\partial}{\partial t} \mathcal{F}(\tau, \bar{x}(\tau; t, x_u(t)))]$$

for almost all $t \in [t_0, T')$, *where* $\varphi_u(\tau, t) \equiv \varphi_0(\tau, \bar{x}(\tau; t, x_u(t)))$.

LEMMA 8. *If* $p(t, x)$ *satisfies*[3]

$$(17) \quad p^{+}(t, x; \varphi_0(t, x)) + f_x^{T}(t, x, \varphi_0(t, x)) p(t, x) + L_x(t, x, \varphi_0(t, x)) = 0,$$

then the right hand side of (16) *has the value*

$$-\frac{\partial}{\partial \tau} [p(\tau, \bar{x}(\tau; t, x_u(t))) \cdot q(\tau, t)]$$

for almost all $\tau \in [t, T)$.

Proof. For $(t_0, x_0) \in \tilde{B}$, (17) implies

$$\frac{\partial}{\partial \tau} p(\tau, \bar{x}(\tau; t, x_u(t))) + f_x^{T}(\tau, \bar{x}(\tau; t, x_u(t)), \varphi_u(\tau, t)) p(\tau, \bar{x}(\tau; t, x_u(t))) + L_x(\tau, \bar{x}(\tau; t, x_u(t)), \varphi_u(\tau, t)) = 0$$

for almost all $\tau \in [t, T)$. If the scalar product (on the right) of this equation with the vector $q(\tau, t)$ be formed, the identity $(f_x^{T} p) \cdot q \equiv p \cdot (f_x q)$ noted and Lemma 5 invoked, the conclusion follows readily.

4. Sufficient conditions for optimal feedback control. From Lemma 6(a), it follows by virtue of [3, Lemma 3] that

$$(18) \quad V^{+}(t, \bar{x}(t; t_0, x_0); \varphi_0(t, \bar{x}(t; t_0, x_0))) + \mathcal{L}(t, \bar{x}(t; t_0, x_0)) = 0$$

almost everywhere on $[t_0, \bar{t}(t_0, x_0))$ for each $(t_0, x_0) \in \tilde{B}$. In a similar way, replacing $(\partial/\partial t)\mathcal{L}(\tau, \bar{x}(\tau; t, x_u(t)))$ in Lemma 6(b) by

$$-(\partial/\partial \tau)[p(\tau, \bar{x}(\tau; t, x_u(t))) \cdot q(\tau, t)]$$

—as is justified by Lemmas 7, 8 provided $p(t, x)$ satisfies (17)—we may write, by virtue of (viii) and [3, Lemma 3],

$$(19) \quad \begin{aligned} &V^{+}(t, x_u(t); u(t)) + L(t, x_u(t), u(t)) \\ &\qquad > \left\{ \mathcal{L}(\bar{t}, \bar{x}(\bar{t}; t, x_u(t))) \frac{d\bar{t}}{dt} - p(\bar{t}, \bar{x}(\bar{t}; t, x_u(t))) \cdot q(\bar{t}, t) \right\} \end{aligned}$$

[3] The superscript "T" denotes the transposed matrix.

almost everywhere on $[t_0, T')$ for each $(t_0, x_0) \in \tilde{B}$. In (19), we have written simply $\bar{t}$ for $\bar{t}(t, x_u(t))$.

Applying [3, Lemma 2] to (18), (19) leads to the conclusion that, for almost all $(t, x) \in \tilde{B}$, the following statements hold:

$$(20) \qquad V^+(t, x; \varphi_0(t, x)) + \mathfrak{L}(t, x) = 0;$$

$$V^+(t, x; u(t)) + L(t, x, u(t))$$
$$(19a) \qquad > \{\mathfrak{L}(\bar{t}(t, x), \bar{x}(\bar{t}(t, x); t, x))\bar{t}^+(t, x; u(t))$$
$$- p(\bar{t}(t, x), \bar{x}(\bar{t}(t, x); t, x)) \cdot q(\bar{t}(t, x), t)\}$$

when $u(t) \neq \varphi_0(t, x)$. If the *transversality condition*

$$\text{(xiii)} \quad \mathfrak{L}(\bar{t}(t, x), \bar{x}(\bar{t}(t, x); t, x))\bar{t}^+(t, x; u(t))$$
$$-p(\bar{t}(t, x), \bar{x}(\bar{t}(t, x); t, x)) \cdot [\bar{x}^+(\tau; t, x; u(t))]_{\tau=\bar{t}(t,x)} \geqq 0$$
$$\text{for almost all } (t, x) \in \tilde{B} \text{ when } u(t) \neq \varphi_0(t, x),$$

be assumed, then (19a) becomes

$$(21) \qquad V^+(t, x; u(t)) + L(t, x, u(t)) > 0 \qquad \text{when} \quad u(t) \neq \varphi_0(t, x).$$

Under this condition then, (20) and (21) together are equivalent to the statement that $L^*(t, x, u(t)) \equiv V^+(t, x; u(t)) + L(t, x, u(t))$ is determinate on B.

Now let us define the generators Π, R by

$$(22) \qquad \Pi(t; t_0, x_0, u) \equiv V(t, x_u(t)) + \int_{t_0}^{t} L(\tau, x_u(\tau), u(\tau))\, d\tau;$$

$$(23) \qquad R(t; t_0, x_0, u) \equiv \Pi(t; t_0, x_0, u) - V(t_0, x_0).$$

Inasmuch as the generator Π and the function L^* satisfy the conditions of [3, Theorem 3], it follows from that theorem that $R(t; t_0, x_0, u)$ satisfies the hypotheses of [3, Theorem 1]. Thus we may conclude from the latter theorem:

$$(24) \qquad \int_{t_0}^{\bar{t}(t_0,x_0)} \mathfrak{L}(\tau, \bar{x}(\tau; t_0, x_0))\, d\tau + V(\bar{t}(t_0, x_0), \bar{x}(\bar{t}(t_0, x_0); t_0, x_0))$$
$$= V(t_0, x_0);$$

$$(25) \qquad \int_{t_0}^{t^{\#}(t_0,x_0,u)} L(\tau, x_u(\tau), u(\tau))\, d\tau + V(t^{\#}(t_0, x_0, u), x_u(t^{\#}(t_0, x_0, u)))$$
$$> V(t_0, x_0)$$

for each t_0 and almost all x_0 such that $(t_0, x_0) \in \tilde{B}$. Of course, from (24) and (5) we find that

$$(26) \qquad V(\bar{t}(t_0, x_0), \bar{x}(\bar{t}(t_0, x_0); t_0, x_0)) = 0;$$

however, (26) can be deduced directly from (5) by continuity, so that (24) contains nothing new.

If the (apparently artificial) assumption

(xiv) $u(t) \neq u_0(t; t_0, x_0)$ implies $(t^{\#}(t_0, x_0, u), x_u(t^{\#}(t_0, x_0, u))) \notin \tilde{B}$

is made, then it is a consequence of (25) and (5) that

$$(27) \qquad \int_{t_0}^{t^{\#}(t_0, x_0, u)} L(\tau, x_u(\tau), u(\tau))\, d\tau > V(t_0, x_0)$$

and this is all that is needed [3, (4)] to justify the assertion that, for $(t_0, x_0) \in \tilde{B}$, $\varphi_0(t, x)$ is the unique optimal feedback control law relative to the criterion $Q(t_0, x_0, u)$ defined by

$$(28) \qquad Q(t_0, x_0, u) \equiv \int_{t_0}^{t^{\#}(t_0, x_0, u)} L(\tau, x_u(\tau), u(\tau))\, d\tau.$$

Remark. The assumption (xiv) may be justified in the following way. Certainly if the consequent of the assumption holds, then (27) ensues. Suppose, however, that for some $(t_0, x_0) \in \tilde{B}$ and some $u \in U(t_0, x_0)$, $u \neq u_0$, that $(t^{\#}(t_0, x_0, u), x_u(t^{\#}(t_0, x_0, u))) \in \tilde{B}$. Then for the "extended control" $u^*(t)$ defined by

$$u^*(t) = \begin{cases} u(t), & t_0 \leqq t < t^{\#}(t_0, x_0, u), \\ u_1(t), & t^{\#}(t_0, x_0, u) \leqq t, \end{cases}$$

where u_1 is optimal from $(t^{\#}(t_0, x_0, u), x_u(t^{\#}(t_0, x_0, u)))$, it is a readily verifiable consequence of (25) and (7) that

$$\int_{t_0}^{t^{\#}(t_0, x_0, u^*)} L(\tau, x_{u^*}(\tau), u^*(\tau))\, d\tau > V(t_0, x_0).$$

Thus $\varphi_0(t, x)$ is still optimal relative to such "extended controls". However, rather than become involved in the obviously messy complications associated with the concept of extended controls, assumption (xiv) is made. This discussion shows that (xiv) is thus a convenience rather than an essential.

Let us now summarize the foregoing results.

THEOREM 1. *If* (i), . . . , (xiv) *hold, if* $p(t, x)$ *satisfies the differential equation* (17), *then for each* t_0 *and almost all* x_0 *such that* $(t_0, x_0) \in \tilde{B}$, *the control*

$$u_0(t; t_0, x_0) \equiv \varphi_0(t, \bar{x}(t; t_0, x_0))$$

is the unique optimal control in $U(t_0, x_0)$ *relative to the criterion* $Q(t_0, x_0, u)$ *defined by* (28), *and the minimum value of this criterion is* $V(t_0, x_0)$.

The assumptions (viii), (xiii), (17) are the "key" hypotheses of Theorem 1 and, as such, may be said to constitute a "minimum principle" for feedback controls. The similarity of this "minimum principle" to the "maximum principle" of Pontryagin is obvious.

5. Determination of $p(t, x)$. Our discussion of the determination of the vector $p(t, x)$ will be formal, since it involves the solution of a generalized nonlinear total differential equation. The existence of solutions to such equations is moot although in some applications it may be obvious.

Let us suppose, for each $(t, x) \in I \times R^n$ and each fixed $p \in R^n$, that the functional

$$L(t, x, \varphi) + p \cdot f(t, x, \varphi) \tag{29}$$

has a unique absolute minimum over Φ at $\bar{\varphi}_0(t, x, p)$. Now suppose that a gauge vector $p(t, x)$ can be found as a solution of the differential equation

$$\begin{aligned} p^{+}(t, x; \bar{\varphi}_0(t, x, p)) + f_x^{T}(t, x, \bar{\varphi}_0(t, x, p))p \\ + L_x(t, x, \bar{\varphi}_0(t, x, p)) = 0. \end{aligned} \tag{30}$$

If this can be done, then it is clear that by taking

$$\varphi_0(t, x) \equiv \bar{\varphi}_0(t, x, p(t, x)),$$

there follows

$$\mathfrak{L}(t, x) \equiv L(t, x, \bar{\varphi}_0(t, x, p(t, x))),$$

$$\mathfrak{F}(t, x) \equiv f(t, x, \bar{\varphi}_0(t, x, p(t, x))),$$

and (30) reduces to (17). There remains only to verify that $p(t, x)$ satisfies the transversality condition (xiii); actually, the latter condition serves as a boundary condition for (30).

6. Discussion. While it might seem at first blush that the requirement (characteristic of the Carathéodory technique) that $\varphi_0(t, x)$ be single-valued is too severe, a little thought convinces one that uniqueness of an optimal control is essential to the solution of the synthesis problem. Indeed, one wonders if the requirement of such uniqueness should not extend even to the "open-loop" situation, i.e., the case of programmed controls. For if, from a given "phase" (t_0, x_0), there is more than one optimal control relative to a given criterion, one is faced—in practice—with a choice of only one of these optimal controls and in such a choice there is implicitly involved another criterion by means of which the aptness of the choice may be judged. One cannot avoid the conclusion that if the original criterion had been sufficiently "strong", the necessity for such choice would be obviated.

Theorem 1 having been established under the assumption (xi), it is an immediate corollary that the theorem remains true with $\tilde{B}$ assumed to be of the form

$$\tilde{B} = \{(t, x) \mid 0 \leqq t < T, a_i \leqq x_i < b_i\},$$

a half-open $(n + 1)$-cell. But by virtue of a well-known representation of an arbitrary nonvoid open set [6, p. 18], the validity of the theorem may now be asserted for any nonvoid bounded open $\tilde{B}$. Actually, we could as well have used, in place of (xi), the assumption that $\tilde{B}$ is an appropriately formed set of positive measure. Then, in place of the continuity argument used to establish (7), we could have used [3, Lemma 2].

REFERENCES

[1] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[2] J. K. Hale and J. P. LaSalle, *Differential equations: linearity vs. nonlinearity*, SIAM Review, 5 (1963), pp. 249–272.

[3] T. F. Bridgland, Jr., *On the existence of optimal feedback controls*, this Journal, 1, pp. 261–274.

[4] C. Carathéodory, *Variationsrechnung und partielle Differentialgleichungen erster Ordnung*, Leipzig, 1935.

[5] R. E. Kalman, *The theory of optimal control and the calculus of variations*, RIAS Tech. Rep. 61–3, 1961.

[6] E. J. McShane, *Integration*, Princeton University Press, Princeton, 1947.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

THE BANG-BANG PRINCIPLE FOR LINEAR CONTROL SYSTEMS*

L. M. SONNEBORN AND F. S. VAN VLECK†

1. Introduction and statement of results. The bang-bang principle has been treated mathematically by many people starting with Bushaw [1] and continuing to Neustadt [2]. A rather complete bibliography of these and other results may be found in [2]. For linear systems the most general results are those given in [2]. The purpose of this paper is to extend these results as far as is possible in one direction. For some particular systems, it may be possible to use a smaller restraint set, but unless additional hypotheses are imposed on the system our results are the best possible.

We consider the real linear differential system

$$\dot{x} = A(t)x + B(t)u + f(t), \tag{1.1}$$

where $A(t)$ is an $n \times n$ matrix, $B(t)$ is an $n \times m$ matrix, and $f(t)$ is an n-vector, each measurable on E^1 and integrable (absolutely) on each compact interval. For each measurable (and integrable) function u, called a *control function*, on a compact interval $t_0 \leqq t \leqq t_1$, the solution of (1.1) initiating at x_0 is

$$x(t) = X(t)x_0 + X(t) \int_{t_0}^{t} X^{-1}(s)[B(s)u(s) + f(s)]\, ds, \tag{1.2}$$

where $X(t)$ is the fundamental solution of the homogeneous system $\dot{x} = A(t)x$ for which $X(t_0) = I$. We consider control functions $u(t)$ on $t_0 \leqq t \leqq t_1$ which lie in a nonempty, bounded restraint set $U \subset E^m$; that is, $u(t) \in U$ for each t, $t_0 \leqq t \leqq t_1$.

With these fixed data $\{(1.1), x_0, t_0, U\}$, the set of all endpoints $x(t_1)$ defines the *set of attainability* $K_U(t_1) \subset E^n$. For any compact $V \subset E^m$, V_0 will denote the set of extremal points of $H(V)$, the convex hull of V, and will be called *the set of extreme points of* V.

In this terminology our main results are the following theorem and its corollaries:

THEOREM 1. *If $U \subset E^m$ is compact and convex, then $K_{U_0}(t_1)$ and $K_U(t_1)$ are compact and convex. Further,*

$$K_{U_0}(t_1) = K_U(t_1). \tag{1.3}$$

COROLLARY 1. *If $V \subset E^m$ is a compact set with convex hull $H(V)$ and extremal point set V_0 and W is any set such that $V_0 \subset W \subset H(V)$, then*

$$K_W(t_1) = K_{H(V)}(t_1) = K_{V_0}(t_1). \tag{1.4}$$

* Received by the editors March 23, 1964.

† Department of Mathematics, The University of Kansas, Lawrence, Kansas.

To see that Corollary 1 follows from Theorem 1, note that $A \subset B$ implies that $K_A(t_1) \subset K_B(t_1)$ and hence

$$K_{V_0}(t_1) \subset K_W(t_1) \subset K_{H(V)}(t_1).$$

But on the other hand, $K_{V_0}(t_1) = K_{H(V)}(t_1)$ by Theorem 1, so that all three sets are equal.

Corollary 2. *If V_0 is the set of vertices of a compact polytope V, then*

$$K_{V_0}(t_1) = K_V(t_1).$$

It is well known that $K_U(t_1)$ is compact and convex whenever U is; the new feature of Theorem 1 is the assertion that $K_{U_0}(t_1)$ is compact and convex and that $K_{U_0}(t_1)$ actually equals $K_U(t_1)$. Theorem 1 asserts that anything that can be done by a control having values in U can be done by a control ranging over only the set of extreme points of U. Corollary 2 includes the results of LaSalle [3] and Pontryagin [4]; LaSalle's bang-bang principle for the system (1.1) restricted U to be a parallelepiped while Pontryagin's results were for a polytope. Theorem 1 also extends the result of Neustadt [2] in so far as linear systems are concerned. Neustadt's result as applied to linear systems can be stated as follows.

Theorem 2. *If $V \subset E^m$ is compact, then $K_V(t_1)$ is compact and convex. Further, if $H(V)$ is the convex hull of V, then*

$$K_V(t_1) = K_{H(V)}(t_1).$$

To see that Theorem 1 implies Theorem 2, suppose we have a compact set $V \subset E^m$. Then $V_0 \subset V \subset H(V)$ and $V_0 = [H(V)]_0$. Hence by Corollary 1, Theorem 2 follows. On the other hand, Theorem 2 does not necessarily imply Theorem 1 if $m > 2$. This is evident if one recalls that, for $m > 2$, the set of extreme points of a compact convex set need not be compact.

2. Proofs. If f is a function with domain E and $A \subset E$, f_A will denote the restriction of f to A. U is a fixed compact convex subset of E^m, and $H = \{x \in E^{m+1} \mid \sum_{i=1}^{m+1} x_i = 1; x_i \geqq 0, i = 1, 2, \cdots, m+1\}$, the standard m-simplex of E^{m+1}.

Definition. *A subset A of a Euclidean space is an analytic set if, and only if, there exists a closed set $A(n_1, n_2, \cdots, n_k)$ for each finite sequence $(n_1, n_2, \cdots, n_k)$ of positive integers such that*

$$A = \bigcup_S \bigcap_{k=1}^{\infty} A(n_1, n_2, \cdots, n_k),$$

where S is the set of all infinite sequences of positive integers.

Lemma 1. *If E is a bounded (Lebesgue) measurable subset of E^m and $u: E \to E^n$ is measurable and $A \subset E^n$ is an analytic set, then $u^{-1}(A)$ is measurable.*

Proof. If $\epsilon > 0$, there is a compact subset $C \subset E$ such that $E - C$ has measure less than ϵ and u_C is continuous. Since $u_C(C)$ is compact, $u_C(C) \cap A$ is analytic. Hence $u_C^{-1}(u_C(C) \cap A) = u_C^{-1}(A)$ which is analytic and therefore (Lusin [5, p. 152]) measurable. Since ϵ was arbitrary, $u^{-1}(A)$ is measurable.

LEMMA 2. *If A is a G_δ (countable intersection of open sets) in a complete metric space (X, d), then there is a metric d^* for A such that (A, d^*) is a complete metric space with the same topology as (A, d) and $d^*(x, y) \geqq d(x, y)$ for all $x, y \in A$.*

See Hausdorff [6, p. 244].

Lemma 3, which follows, is an extension of a lemma due to Filippov [7]. It and its proof are due to N. Aronszajn (private communication) whose generous help is deeply appreciated by the authors.

LEMMA 3. *If $V \subset E^m$ is a G_δ and $\phi: V \times [0, 1] \rightarrow E^n$ is continuous and $y:[0, 1] \rightarrow E^n$ is measurable and satisfies*

$$y(t) \in \phi(V, t) \quad \textit{for all} \quad t \in [0, 1],$$

then there is a measurable function $v:[0, 1] \rightarrow V$ such that

$$y(t) = \phi(v(t), t) \quad \textit{for all} \quad t \in [0, 1].$$

Proof. Define $\phi': V \times [0, 1] \rightarrow E^n \times [0, 1]$ by $\phi'(u, t) = (\phi(u, t), t)$. Clearly ϕ' is continuous. If $\epsilon > 0$ and $u \in V$, then, due to the compactness of $[0, 1]$, there is a neighborhood $N(u)$ of u such that $d(\phi'(N(u), t)) < \epsilon$ for all $t \in [0, 1]$. ($d(A)$ is the diameter of the bounded set A.) We also can require that $d^*(N(u))$, the diameter of $N(u)$ in the complete metric topology for V given by Lemma 2, also be $<\epsilon$. Thus, since V is separable, we inductively (on k) define for each finite sequence $(n_1, n_2, \cdots, n_k)$ of positive integers closed sets $V(n_1, \cdots, n_k)$ such that

(i) $V(\emptyset) = V$,

(ii) $d^*(V(n_1, \cdots, n_k)) < 2^{-k}$ for $k > 0$,

(iii) $V(n_1, \cdots, n_{k-1}) = \bigcup_{n_k} V(n_1, \cdots, n_{k-1}, n_k)$ for $k > 0$, and

(iv) $d(\phi'(V(n_1, \cdots, n_k), t)) < 2^{-k}$.

Next we let $y':[0, 1] \rightarrow E^n \times [0, 1]$ be given by $y'(t) = (y(t), t)$, and let $I(\emptyset) = [0,1]$. We now prove that there are measurable sets $I(n_1, \cdots, n_k)$ corresponding to the above sets $V(n_1, \cdots, n_k)$ such that

(v) $I(\emptyset) = [0, 1]$,

(vi) $y'(t) \in \phi'(V(n_1, \cdots, n_k), t)$ for each $t \in I(n_1, \cdots, n_k)$, and

(vii) $I(n_1, \cdots, n_{k-1}) = \bigcup_{n_k} I(n_1, \cdots, n_{k-1}, n_k)$ for all $k > 0$,

where, clearly, if such $I(n_1, \cdots, n_k)$ exist the last union may be made into a disjoint union by the usual process.

By hypothesis, $I(\emptyset)$ satisfies (iv). We assume, then, that $I(n_1, \cdots, n_{k-1})$

exist satisfying all the above. Then

$$y'(t) \in \phi'(V(n_1, \cdots, n_{k-1}), t) \text{ for } t \in I(n_1, \cdots, n_{k-1}),$$

and, therefore,

$$\phi'(V(n_1, \cdots, n_{k-1}), I(n_1, \cdots, n_{k-1}))$$
$$= \bigcup_{n_k} \phi'(V(n_1, \cdots, n_k), I(n_1, \cdots, n_{k-1})) \subset \bigcup_{n_k} \phi'(V(n_1, \cdots, n_k), I(\emptyset)),$$

where each of these last sets is analytic as a continuous image of a G_δ (cf. Sierpinski [8, p. 219]). Let

$$I(n_1, \cdots, n_k) = y'^{-1}(\phi'(V(n_1, \cdots, n_k), I(\emptyset)) \cap I(n_1, \cdots, n_{k-1}).$$

By Lemma 1, the sets $I(n_1, \cdots, n_k)$ are measurable. They clearly have property (vii). Now if $t \in I(n_1, \cdots, n_k)$, $y' = (y(t), t) \in (\phi(V, t), t) = \phi'(V, t)$ and $y'(t) \in \phi'(V(n_1, \cdots, n_k), I(\emptyset))$. Thus $y'(t) \in \phi'(V(n_1, \cdots, n_k), t)$, and the constructed sets satisfy (vi) also.

Note that for each $k \geqq 0$, $I(\emptyset) = [0, 1] = \bigcup I(n_1, \cdots, n_k)$. We construct a sequence of "step functions" $v_k:[0, 1] \to V$ by choosing a point $v(n_1, \cdots, n_k) \in V(n_1, \cdots, n_k)$ for each $(n_1, \cdots, n_k)$ and setting $v_k(t) = v(n_1, \cdots, n_k)$ for each $t \in I(n_1, \cdots, n_k)$. The measurability of $I(n_1, \cdots, n_k)$ guarantees the measurability of v_k. Because of condition (ii), the sequence $\{v_k(t)\}$ is a Cauchy sequence in the d^* metric and hence also in the Euclidean metric. Since d^* is complete, $\lim_{k\to\infty} v_k(t) = v(t) \in V$ in d^* and hence in the Euclidean metric. Hence $v:[0, 1] \to V$ is measurable. From the continuity of ϕ', we get

$$\lim_{k\to\infty} \phi'(v_k(t), t) = \phi'(v(t), t).$$

On the other hand, $y'(t) \in \phi'(V(n_1, \cdots, n_k), t)$ for $t \in I(n_1, \cdots, n_k)$ and $d(\phi'(V(n_1, \cdots, n_k), t)) < 2^{-k}$ so that

$$d(y'(t), \phi'(v_k(t), t)) < 2^{-k},$$

whence

$$(\phi(v(t), t), t) = \lim_{k\to\infty} \phi'(v_k(t), t) = y'(t) = (y(t), t)$$

so that, finally, $y(t) = \phi(v(t), t)$. This concludes the proof of Lemma 3.

The next lemma is well known but for the sake of completeness we include a brief proof.

LEMMA 4. *Let V_0 be the set of extreme points of a compact convex subset V of a normed linear space. V_0 is a G_δ.*

Proof. Let $A_n = \{x \in V \mid \text{there exist } y, z \in V \text{ and } \lambda \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right] \text{ such that } x = \lambda y + (1 - \lambda)z\}$. Since V and the interval $I_n = \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$ are compact, the set A_n is closed. Therefore $A \equiv \bigcup_{n=1}^{\infty} A_n$ is an F_σ which contains all points of V except the extreme points. Thus $V_0 = V - A$ is a G_δ.

The following lemma is used to represent a control function; it was suggested by a similar construction due to H. Hermes [9].

LEMMA 5. *Let $f:[0, 1] \rightarrow U$ be measurable. Then f admits a representation of the form*

$$f(t) = \sum_{i=1}^{m+1} \alpha^i(t) u_i(t) \tag{2.1}$$

where the real-valued functions α^i are measurable, $\alpha(t) = (\alpha^1(t), \cdots, \alpha^{m+1}(t)) \in H$ for each $t \in [0, 1]$ and $u_i:[0, 1] \rightarrow U_0$ is measurable, $i = 1, 2, \cdots, m + 1$.

Proof. For each $t \in [0, 1]$ there exist a point $\alpha(t) = (\alpha^1(t), \cdots, \alpha^{m+1}(t)) \in H$ and points $u_i(t) \in U_0$ such that

$$f(t) = \sum_{i=1}^{m+1} \alpha^i(t) u_i(t).$$

It remains to show that the functions on the right-hand side can be selected to be measurable. In order to do that, note that since U_0 is a G_δ the set

$$G = H \times \underbrace{U_0 \times \cdots \times U_0}_{m + 1}$$

is a G_δ and define a function $F:G \times [0, 1] \rightarrow E^m$ by

$$F(\alpha^1, \alpha^2, \cdots, \alpha^{m+1}, u_1, u_2, \cdots, u_{m+1}, t) = \sum_{i=1}^{m+1} \alpha^i u_i .$$

Then F is continuous on $G \times [0, 1]$ and for each $t \in [0, 1], f(t) \in F(G, t)$. Thus by the principal lemma, Lemma 3, there exist measurable functions $\alpha(t) \in H$ and $u_i(t) \in U_0$ such that for all t,

$$f(t) = F(\alpha(t), u_1(t), \cdots, u_{m+1}(t), t) = \sum_{i=1}^{m+1} \alpha^i(t) u_i(t).$$

This completes the proof of Lemma 5.

The next theorem is due to Dvoretzky, Wald and Wolfowitz [10, p. 68] and is an extension of Liapounov's Theorem [11, 12]. This together with the preceding lemma enables us to obtain a bang-bang control which accomplishes the same end as a given control.

THEOREM 3. *Let the classes of functions β and $\bar{\beta}$ be defined as follows where H_q is the standard q-simplex in E^{q+1}:*

$\beta = \{\alpha \mid \alpha:[0, 1] \rightarrow H_q$, *$\alpha$ measurable*$\}$

$\bar{\beta} = \{\bar{\alpha} \mid \bar{\alpha} \in \beta$ *and for each $t \in [0, 1]$ exactly one of the components of $\bar{\alpha}$ equals one*$\}$.

Let $\mu_1, \cdots, \mu_p$ be finite measures and let

$$A_{pq} = \left\{ \left(\int_0^1 \alpha^1 \, d\mu_1, \cdots, \int_0^1 \alpha^1 \, d\mu_p, \int_0^1 \alpha^2 \, d\mu_1, \cdots, \int_0^1 \alpha^2 \, d\mu_p, \cdots, \int_0^1 \alpha^q \, d\mu_p \right) \Big| \, \alpha \in \beta \right\}$$

and

$$\bar{A}_{pq} = \left\{ \left(\int_0^1 \bar{\alpha}^1 \, d\mu_1, \cdots, \int_0^1 \bar{\alpha}^1 \, d\mu_p, \cdots, \int_0^1 \bar{\alpha}^q \, d\mu_p \right) \Big| \, \bar{\alpha} \in \bar{\beta} \right\}.$$

If $\mu_1, \cdots, \mu_p$ are atomless measures, then $A_{pq} = \bar{A}_{pq}$.

THEOREM 4. *If Y is a measurable and (absolutely) integrable $n \times m$ matrix defined on* [0, 1],

$$A = \left\{ \int_0^1 Y(t)u(t) \, dt \mid u \text{ measurable}, u(t) \in U \right\},$$

and

$$A_0 = \left\{ \int_0^1 Y(t) \, u_0(t) \, dt \mid u_0 \text{ measurable}, u_0(t) \in U_0 \right\},$$

then

$$A = A_0 .$$

Proof. By Lemma 5, if f is measurable with $f(t) \in U$, then f may be written in the form (2.1) where α^i and u_i are as given in the lemma. Thus

$$\int_0^1 Y(t)f(t) \, dt = \int_0^1 Y(t) \left[\sum_{i=1}^{m+1} \alpha^i(t)u_i(t) \right] dt = \sum_{i=1}^{m+1} \int_0^1 \alpha^i(t)v_i(t) \, dt,$$

where $v_i(t) = Y(t)u_i(t)$.

For each measurable subset $E \subset [0, 1]$, define

$$\mu_i^j(E) = \int_E v_i^j(t) \, dt$$

where v_i^j is the jth component of the n-vector v_i, $i = 1, 2, \cdots, m + 1$; $j = 1, 2, \cdots, n$. Each of the measures μ_i^j is atomless. Next consider the $[(m + 1)^2 \cdot n]$-dimensional vector w_α, $\alpha \in \beta$, defined by

$$w_\alpha = \left(\int_0^1 \alpha^1 \, d\mu_1^1, \cdots, \int_0^1 \alpha^1 \, d\mu_1^n, \int_0^1 \alpha^1 \, d\mu_2^1, \cdots, \int_0^1 \alpha^1 \, d\mu_{m+1}^n, \int_0^1 \alpha^2 \, d\mu_1^1, \cdots, \int_0^1 \alpha^{m+1} \, d\mu_{m+1}^n \right).$$

By Theorem 3, there exists a measurable $\bar{\alpha}$ with $\bar{\alpha}^i(t) = 0$ or 1 and $\sum_{i=1}^{m+1} \bar{\alpha}^i(t) = 1$, such that $w_\alpha = w_{\bar{\alpha}}$.
Therefore

$$\left(\int_0^1 \bar{\alpha}^1\, d\mu_1^{\;1}, \cdots, \int_0^1 \bar{\alpha}^1\, d\mu_1^{\;n}\right) = \left(\int_0^1 \alpha^1\, d\mu_1^{\;1}, \cdots, \int_0^1 \alpha^1\, d\mu_1^{\;n}\right),$$

$$\left(\int_0^1 \bar{\alpha}^2\, d\mu_2^{\;1}, \cdots, \int_0^1 \bar{\alpha}^2\, d\mu_2^{\;n}\right) = \left(\int_0^1 \alpha^2\, d\mu_2^{\;1}, \cdots, \int_0^1 \alpha^2\, d\mu_2^{\;n}\right),$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$\left(\int_0^1 \bar{\alpha}^{m+1}\, d\mu_{m+1}^1, \cdots, \int_0^1 \bar{\alpha}^{m+1}\, d\mu_{m+1}^n\right) = \left(\int_0^1 \alpha^{m+1}\, d\mu_{m+1}^1, \cdots, \int_0^1 \alpha^{m+1}\, d\mu_{m+1}^n\right),$$

and hence

$$\sum_{i=0}^{m+1} \int_0^1 \alpha^i\, v_i\, dt = \sum_{i=0}^{m+1} \int_0^1 \bar{\alpha}^i\, v_i\, dt.$$

Therefore

$$\int_0^1 Y(t)f(t)\, dt = \sum_{i=1}^{m+1} \int_0^1 \alpha^i\, v_i\, dt = \sum_{i=1}^{m+1} \int_0^1 \bar{\alpha}^i\, v_i\, dt.$$

If we let

$$I_i = \{t \in [0, 1] \mid \bar{\alpha}^i(t) = 1\}, \qquad i = 1, 2, \cdots, m + 1,$$

then each set I_i is measurable, $\bigcup_{i=1}^{m+1} I_i = [0, 1]$, and $I_i \cap I_j = \delta_{ij} I_i$. Next define

$$f_0(t) = u_i(t) \quad \text{for} \quad t \in I_i, \qquad i = 1, 2, \cdots, m + 1.$$

Then f_0 is measurable, $f_0(t) \in U_0$, and

$$\int_0^1 Y(t)f_0(t) = \sum_{i=1}^{m+1} \int_{I_i} Y(t)u_i(t)\, dt$$
$$= \sum_{i=1}^{m+1} \int_0^1 \bar{\alpha}^i(t)\, v_i(t)\, dt = \int_0^1 Y(t)f(t)\, dt.$$

Thus any element of A is an element of A_0. Since $A_0 \subset A$, Theorem 4 is established.

To finish the proof of Theorem 1, note that without loss of generality we may assume $t_0 = 0$ and $t_1 = 1$ so that

$$x(1) = c + C \int_0^1 Y(t)u(t)\, dt$$

where $c = X(1)x_0 + X(1)\int_0^1 X^{-1}(t)f(t)\,dt$, C is the nonsingular matrix $X(1)$, and $Y(t) = X^{-1}(t)B(t)$. Thus

$$K_U(t_1) = K_U(1) = c + CA$$

and

$$K_{U_0}(t_1) = K_{U_0}(1) = c + CA_0 .$$

But, by Theorem 4, $A = A_0$, so Theorem 1 is established.

3. Remarks.

1. If U is a given compact set, one might hope that a set $S \subset U$ has the bang-bang property, namely $K_S(t) = K_{H(U)}(t)$, if and only if $U_0 \subset S$. Unfortunately this is not true. For example, consider the (one dimensional) control system $\dot{x} = x$ with any restraint set U. Since the set of attainability (in time t) is always the one point $e^{(t-t_0)}x_0$, we do not have a converse to Theorem 1.

2. To see that Theorem 1 is, in general, *best possible*, consider the (three dimensional) control system $\dot{x} = u$ when the restraint set U is any bounded set whose extreme point set U_0 is not compact. Any subset of U not containing all points of U_0 obviously fails to have the bang-bang property.

3. As Neustadt [2, p. 115] has remarked, the convexity of the set of attainability is of importance if it is desired to compute an optimal control by means of the Pontryagin maximum principle. Moreover, the fact that the optimal control can always be chosen to be a bang-bang control should be of use to design engineers.

4. The existence of optimal controls, assuming the system is controllable and the cost is given by $C(u) = \int_{t_0}^{t_1} [a(t)x(t) + b(t)u(t) + f_0(t)]\,dt$, is assured by the same argument as given by Neustadt [2, pp. 115–116] since we have shown that the set of attainability depends at most on the extreme points of a compact restraint set. (We are assuming that the target set, which is closed for each t, moves in an upper semi-continuous manner.)

REFERENCES

[1] D. W. Bushaw, *Optimal discontinuous forcing terms*, Contributions to the Theory of Nonlinear Oscillations, Vol. IV, Princeton University Press, Princeton, 1958, pp. 29–52.

[2] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[3] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, Vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.

[4] L. S. Pontryagin, *Optimal control processes*, Uspehi Mat. Nauk, 14, 1 (85) (1959), pp. 3–20.

[5] N. Lusin, *Leçons sur les Ensembles Analytiques*, Gauthier-Villars, Paris, 1930.

[6] F. Hausdorff, *Set Theory*, Chelsea, New York, 1957.

[7] A. F. Filippov, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ. Ser. Mat. Meh. Astr. Fiz. Him, 2 (1959), pp. 25–32. English translation in this Journal, 1 (1962), pp. 76–84.

[8] W. Sierpinski, *General Topology*, University of Toronto Press, Toronto, 1952.

[9] H. Hermes, *A note on the range of a vector measure; application to the theory of optimal control*, J. Math. Anal. Appl., 8 (1964), pp. 78–83.

[10] A. Dvoretzky, A. Wald and J. Wolfowitz, *Relations among certain ranges of vector measures*, Pacific J. Math., 1 (1951), pp. 59–74.

[11] A. Liapunov, *Sur les fonctions-vecteurs complétement additives*, Bull. Acad. Sci. URSS. Ser. Math. (Izv. Akad. Nauk SSSR), 4 (1940), pp. 465–478.

[12] P. R. Halmos, *The range of a vector measure*, Bull. Amer. Math. Soc., 54 (1948), pp. 416–421.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

STABILITY CRITERIA FOR FEEDBACK SYSTEMS WITH A TIME LAG*

ALLAN M. KRALL†

1. Introduction. In the back of many earlier texts on control systems, topics such as time lag systems, sample data systems and nonlinear systems appear. At the present time, books are appearing in these various fields with one notable exception—time lag systems. The principal reason for this omission is simply that not enough is known about them, especially when tests for stability are considered. This paper proposes to fill in some of the gaps in this area.

We will consider a linear feedback system in which the open loop transfer function has a time lag τ. In a great many instances the open loop transfer function may be represented by $Ke^{-\tau s}h(s)/g(s)$ where $g(s)$ and $h(s)$ are relatively prime polynomials in s, $g(s) = s^n + a_1s^{n-1} + \cdots$, $h(s) = s^m + b_1s^{m-1} + \cdots$. If the open loop output is multiplied[1] by $e^{i\theta}$ and added to the input to form the "error", the closed loop transfer function is of the form

$$\frac{Ke^{-s\tau}h(s)}{g(s) - Ke^{i\theta}e^{-s\tau}h(s)}.$$

Stability problems may then be resolved by studying the zeros of the characteristic equation

$$F(z) = g(z) - Ke^{i\theta}e^{-\tau z}h(z) = 0,$$

the system being stable if all the zeros of $F(z)$ have negative or zero real parts.

Unfortunately, unlike systems with no time lag, $F(z)$ has infinitely many zeros, and sometimes an infinite number with arbitrarily large positive real part. Since slight variations in the coefficients of $F(z)$ only vary the zeros locally, it is necessary to know when such situations occur.

THEOREM 1.1. *Let* $F(z) = g(z) - Ke^{i\theta}e^{-\tau z}h(z)$, *where* $g(z) = z^n + a_1z^{n-1} + \cdots$, *and* $h(z) = z^m + b_1z^{m-1} + \cdots$, $\tau > 0$, $K \geqq 0$ and $\theta \geqq 0$ *are real constants*, a_i *and* b_i *are complex constants.*

I. *If* $n > m$: *the number of zeros of* $F(z)$ *with positive real part* (*or lying in any right halfplane*) *is finite*; *if* $K \neq 0$, $F(z)$ *has an infinite number of zeros with arbitrarily large negative real parts.*

* Received by the editors January 30, 1964, and in revised form May 1, 1964.

† 333 McAllister Hall, Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania.

[1] By letting $\theta = 0$, we have positive feedback; $\theta = \pi$, negative feedback. Thus we can consider both simultaneously.

II. *If $n = m$: when $K \neq 0$, $F(z)$ has an infinite number of zeros given by*

$$z = \frac{1}{\tau} (\log_e K + i(\theta + 2k\pi)) + o(1), \tag{1.1}$$

where $k = 0, \pm 1, \pm 2, \cdots$, and only a finite number of other zeros. If $K < 1$, $F(z)$ has only a finite number of zeros with positive real parts. If $K > 1$, $F(z)$ has only a finite number of zeros with negative real part.

III. *If $n < m$: the number of zeros of $F(z)$ with negative real part (or lying in any left halfplane) is finite; if $K \neq 0$, $F(z)$ has an infinite number of zeros with arbitrarily large positive real parts.*

Proof. See [7].

THEOREM 1.2. *With the notation of Theorem 1.1, if $n > m$ or $n = m$, $K < 1$, for fixed K, all of the zeros of $F(z)$ with positive real part lie within a circle of radius $\rho = M + 1$, where*

$$M = \sup [\{| a_i | + | K |\cdot| b_i |\}_1^m \quad \text{and} \quad \{| a_i |\}_{m+1}^n] \quad \text{if } n > m,$$

$$M = \sup \left[\left\{ \frac{| a_i | + | K |\cdot| b_i |}{1 - K} \right\}_1^n \right] \qquad \text{if } n = m.$$

Proof. If $| z | > \rho$ and $\operatorname{Re}(z) \geqq 0$, then when $n > m$,

$$\begin{aligned} | F(z) | &\geqq | z |^n - \sum_{i=1}^n | a_i |\cdot| z |^{n-i} - | K | \sum_{i=1}^m | b_i |\cdot| z |^{n-i} \\ &\geqq | z |^n - M \sum_{i=1}^n | z |^{n-i} \\ &= | z |^n - M \left(\frac{| z |^n - 1}{| z | - 1} \right) \\ &= \frac{| z |^n[| z | - (1 + M)] + M}{| z | - 1} \\ &> 0. \end{aligned}$$

When $n = m$,

$$\begin{aligned} | F(z) | &\geqq | 1 - K |\cdot| z |^n - \sum_{i=1}^n [| a_i | + K | b_i |] | z |^{n-i} \\ &\geqq | 1 - K | \left[| z |^n - M \sum_{i=1}^n | z |^{n-i} \right] \\ &> 0. \end{aligned}$$

Note that the radius of this circle is independent of τ as long as $\tau \geqq 0$.

Thus in trying to find which values of K lead to stable systems, we need to consider only the cases $n > m$ and $n = m$, $K < 1$. We now extend the Nyquist criterion and the root-locus method to cover these situations.

2. Nyquist criterion. If $n > m$ or $n = m$, $K < 1$, let R be any number greater than the number ρ of Theorem 1.2. R is large enough so that a circle of radius R centered at the origin contains all of the zeros of $g(z)$. Let C_R be a semicircular contour varying along the imaginary axis from $-R$ to R and then from $(0, R)$ to $(0, -R)$ along half of the previously mentioned circle in a clockwise manner, avoiding zeros of $g(z)$ on the imaginary axis by arbitrarily small semicircles centered at those zeros.

THEOREM 2.1. *The number of times $Ke^{-\tau z}h(z)/g(z)$ passes through $e^{-i\theta}$ as z varies around C_R is equal to the number of imaginary zeros of $F(z)$. If $F(z)$ has no imaginary zeros, let P be the number of zeros of $g(z)$ with positive real parts, Z be the number of zeros of $F(z)$ with positive real parts, N be the number of counterclockwise encirclements of $e^{-i\theta}$ by $Ke^{-\tau z}h(z)/g(z)$ as z varies around C_R. Then $Z = P - N$.*

Proof. (See [4].) It is well known that the number of counterclockwise encirclements of the origin by a meromorphic function as z varies in a counterclockwise manner around a contour is equal to the number of zeros minus the number of poles of the function contained within the contour. Now as z varies around C_R, the number of encirclements of $e^{-i\theta}$ by $Ke^{-\tau z}h(z)/g(z)$ is the same as the number of encirclements of the origin by $e^{-i\theta} - Ke^{-\tau z}h(z)/g(z)$. This is the same as the number of encirclements of the origin by $1 - Ke^{i\theta}e^{-\tau z}h(z)/g(z)$, which is the same as the number of encirclements of the origin by $F(z)/g(z)$. Since C_R is a clockwise contour, we see $Z = P - N$.

COROLLARY 2.2. *A necessary and sufficient condition that $F(z)$ have no zeros with positive real parts is that $N = P$.*

Note that in constructing the path of $Ke^{-\tau z}h(z)/g(z)$ as z varies along the imaginary axis, the magnitude of $Ke^{-\tau z}h(z)/g(z)$ is the same as when $\tau = 0$. Only the argument is changed by an amount $-\tau\omega$ when $z = i\omega$. Further note that M and N circles may be used the same as when $\tau = 0$. (See [6, pp. 141–144].)

There is an alternate method which may be used which involves only the Nyquist contour with $\tau = 0$. This procedure was first used by A. A. Sokolov and N. N. Miasnikov (see [9, p. 421]) who were considering the Mikhailov criterion—the Soviet equivalent of the Nyquist criterion.

THEOREM 2.3. *Let $N(\tau)$ be the number of counterclockwise encirclements of $e^{-i\theta}$ by $Ke^{-\tau z}h(z)/g(z)$ as z varies over C_R. Then if the path of $Kh(z)/g(z)$ does not intersect the unit circle as z varies over C_R, $N(\tau) = N(0)$ for all $\tau \geqq 0$.*

Proof. If $N(\tau) \neq N(0)$ for some $\tau \neq 0$, then, since the Nyquist contour is continuous in τ, there must be a τ_0, $0 < \tau_0 < \tau$, for which the Nyquist contour passes through $e^{-i\theta}$. Thus there is an ω such that $Ke^{-i\tau_0\omega}h(i\omega)/g(i\omega) = e^{-i\theta}$ and $| Kh(i\omega)/g(i\omega) | = 1$, which is impossible.

If the path of $Kh(z)/g(z)$ does intersect the unit circle, let $e^{i\alpha_1}$, $e^{i\alpha_2}$, $\cdots$,

$e^{i\alpha_j}, \cdots$ be the points of intersection. For each $e^{i\alpha_j}$, let $i\omega_j$ be a point on the imaginary axis such that $Kh(i\omega_j)/g(i\omega_j) = e^{i\alpha_j}$. Then if $Ke^{-\tau z}h(z)/g(z)$ is to pass through $e^{-i\theta}$ we must have

$$\frac{Ke^{-i\tau\omega_j}h(i\omega_j)}{g(i\omega_j)} = e^{i(\alpha_j - \tau\omega_j)} = e^{-i\theta}$$

or

$$\tau = \frac{1}{\omega_j}(\theta + \alpha_j + 2k\pi),$$

where $k = 0, \pm 1, \pm 2, \cdots$. Let these nonnegative values of τ for all j, k be arranged in an increasing sequence $\tau_1, \tau_2, \cdots, \tau_n, \cdots$. We then have

THEOREM 2.4. *If t_1 and t_2 are in the same open interval (τ_i, τ_{i+1}), then $N(t_1) = N(t_2)$.*

Thus tests for stability may be made by considering the ordinary Nyquist diagrams with $\tau = 0$.

3. The root-locus method. Although the Nyquist criterion is left relatively unchanged for systems with a delay, the root locus diagrams are radically altered. This is to be expected, since the characteristic equation contains an infinite number of zeros. We will see, however, that only a small part of the root-locus diagram is important, and with the aid of some construction rules, that part may be easily found.

It will be convenient to distinguish between various parts of the root locus. Hence the following.

DEFINITION. *The root-locus of $F(z)$ is the set of all points z such that z is a zero of $h(z)$, or for which there is a real number K, $-\infty < K < \infty$, such that $F(z) = 0$.*

The positive root-locus of $F(z)$ is the set of all points z such that z is a zero of $h(z)$, or for which there is a real number K, $0 \leqq K < \infty$, such that $F(z) = 0$. The zeros of $h(z)$ are included in the root-locus since they are limit points of the zeros of $F(z)$ for all of the appropriate choices of K, i.e., they are zeros of $F(z)$ when $K = \infty$.

The negative root-locus can be similarly defined although we will not need to consider it. It is easy to see that the negative root-locus for θ is the positive root-locus for $\pi + \theta$.

THEOREM 3.1. *Let z be a point in the complex plane. The following statements are equivalent.*

(i) *z is on the root-locus of $F(z)$.*

(ii)

$$\cos(\theta - \tau y)\,\mathrm{Im}\,(h(z)\overline{g(z)}) + \sin(\theta - \tau y)\,\mathrm{Re}\,(h(z)\overline{g(z)}) = 0. \tag{3 1}$$

(See [5].)

Proof. Suppose z is on the root-locus. If $g(z) \neq 0$ then for some $K \neq 0$, $Ke^{i\theta}e^{-\tau z}/g(z) = 1$. Thus

$$\frac{h(z)}{g(z)} = K^{-1}\, e^{\tau x}\, [\cos\,(\theta - \tau y) - i \sin\,(\theta - \tau y)].$$

$$h(z)\overline{g(z)} = K^{-1}\, e^{\tau x}\, |\, g(z)\,|^2\, [\cos\,(\theta - \tau y) - i \sin\,(\theta - \tau y)].$$

Since K, τ, x are real,

$$\begin{aligned} \text{Re}\,(h(z)\overline{g(z)}) &= K^{-1}e^{\tau x}\,|\, g(z)\,|^2 \cos\,(\theta - \tau y), \\ \text{Im}\,(h(z)\overline{g(z)}) &= -K^{-1}e^{\tau x}\,|\, g(z)\,|^2 \sin\,(\theta - \tau y). \end{aligned} \tag{3.2}$$

Multiplying the first by $\sin\,(\theta - \tau y)$, the second by $\cos\,(\theta - \tau y)$ and adding, achieves (3.1). So (i) implies (ii).

Conversely, if (3.1) is satisfied, then $\text{Im}\,(e^{i\theta}e^{-\tau z}h(z)\overline{g(z)}) = 0$. So $e^{i\theta}e^{-\tau z}h(z)\overline{g(z)} = R(z)$, where $R(z)$ is real. If $R(z) = 0$, then either $h(z) = 0$ or $g(z) = 0$ and z is on the root-locus. If $R(z) \neq 0$, let $K = |\, g(z)\,|^2/R(z)$. If $K = 0$, then $g(z) = 0$ and z is on the root-locus. If $K \neq 0$, then $Ke^{i\theta}e^{-\tau z}h(z)/g(z) = 1$ and $F(z) = 0$. So (ii) implies (i).

Note that K can be found by

$$K = e^{\tau x}\,|\, g(z)\,|^2 \cos\,(\theta - \tau y)/\text{Re}\,(h(z)\overline{g(z)}),$$

or by

$$K = -e^{\tau x}\,|\, g(z)\,|^2 \sin\,(\theta - \tau y)/\text{Im}\,(h(z)\overline{g(z)}).$$

THEOREM 3.2. *The multiple points of the root-locus are isolated and satisfy*

$$h(z)[g'(z) + \tau g(z)] - g(z)h'(z) = 0. \tag{3.3}$$

Proof. If z is a multiple zero of $F(z)$ for some value of K, then $F(z) = 0$ and $F'(z) = 0$. Eliminating $Ke^{i\theta}e^{-\tau z}$ from these two equations results in (3.3). Since (3.3) is a polynomial of degree at most $n + m$, there can be only a finite number of isolated multiple roots of the root-locus.

THEOREM 3.3. *The points on the root-locus of $F(z)$ for specific K are continuous functions of K.*

This follows directly from Hurwitz' Theorem (see [10, p. 119]). To "make" functions, different branches of the root-locus may be identified at multiple points first according to argument and then according to magnitude for values of K first slightly less than and then slightly larger than that value of K giving a multiple point.

THEOREM 3.4. *With the exception of multiple points, the points on the root-locus of $F(z)$ for specific K are differentiable functions of K.*

Proof. (See [3].) Let z_0 be a simple zero of $F(z)$ when $K = K_0$. We need to show that

$$\lim_{K \to K_0} \frac{z - z_0}{K - K_0}$$

exists, where z is a point of the root-locus and $\lim_{K \to K_0} z = z_0$. We have

$$0 = g(z) - Ke^{i\theta}e^{-\tau z}h(z),$$

$$0 = g(z) - K_0e^{i\theta}e^{-\tau z}h(z) \quad - \quad (K \quad - \quad K_0)e^{i\theta}e^{-\tau z}h(z),$$

$$0 = (z - z_0)W(z) - (K - K_0)e^{i\theta}e^{-\tau z}h(z),$$

where $W(z_0) \neq 0$ and $W(z_0) = dF/dz \mid z = z_0$, $K = K_0$. From this we find that

$$\lim_{K \to K_0} \frac{z - z_0}{K - K_0} = \frac{e^{i\theta}\, e^{-\tau z_0}\, h(z_0)}{g'(z_0) - K_0e^{i\theta}\, e^{-\tau z_0}\, (h'(z_0) - \tau h(z_0))}. \tag{3.4}$$

THEOREM 3.5. *If $h(z)$ and $g(z)$ have real coefficients,* (3.1) *becomes*

$$\begin{aligned} &\cos\,(\theta - \tau y) \sum_{k=0}^{\infty} \frac{(-1)^k y^{2k+1}}{(2k+1)!} \sum_{i=0}^{2k+1} \binom{2k+1}{i} \\ (3.5) \quad &\cdot (-1)^{2k+1-i} h^{(i)}(x) g^{(2k+1-i)}(x) + \sin\,(\theta - \tau y) \sum_{k=0}^{\infty} \frac{(-1)^k y^{2k}}{(2k)!} \\ &\cdot \sum_{i=0}^{2k} \binom{2k}{i} (-1)^{2k-i} h^{(i)}(x) g^{(2k-1)}(x) = 0. \end{aligned}$$

Proof. (See [5].) This follows from expanding $h(z)$ and $g(z)$ in MacLaurin expansions about x and solving for the real and imaginary parts of $h(z)\overline{g(z)}$.

THEOREM 3.6. *The root-locus contains the entire real line* ($y = 0$) *if and only if $\theta = 0$ or $\theta = \pi$ when $h(z)$ and $g(z)$ are real polynomials.*

Proof. If the x-axis is contained in the root-locus, then $y = 0$ is a solution of (3.5). Thus $\sin \theta = 0$, and $\theta = 0$ or $\theta = \pi$. The converse is trivial.

4. The positive root-locus.

THEOREM 4.1. *As x becomes arbitrarily large to the right, the positive root-locus of $F(z)$ approaches*

$$y = \frac{1}{\tau}\,(\theta + 2k\pi), \tag{4.1}$$

where $k = 0, \pm 1, \pm 2, \cdots$, in the right halfplane asymptotically. Further, $K \to \infty$ as $x \to \infty$.

As x becomes arbitrarily large to the left, the positive root-locus of $F(z)$ approaches

$$y = \frac{1}{\tau}\,(\theta - (n - m)\pi + 2k\pi), \tag{4.2}$$

where $k = 0, \pm 1, \pm 2, \cdots$, *in the left halfplane asymptotically. Further* $K \to 0$ *as* $x \to -\infty$.

Proof. Let $u(z) = e^{i\theta}e^{-\tau z}h(z)/g(z)$ and consider only those values of z greater in absolute value than each of the zeros of $g(z)$ and $h(z)$. For those values of z, the positive root-locus of $F(z)$ consists of all points where $u(z)$ is real and $u(z) > 0$, i.e., $\arg u(z) = 2k\pi$ for some integer k.

Now $\arg u(z) = \theta - \tau y + \arg h(z) - \arg g(z)$. For bounded y, as $x \to \infty$, $\arg h(z) \to 0$ and $\arg g(z) \to 0$. Thus for bounded y, as $x \to \infty$, $\arg u(z) = \theta - \tau y + o(1)$.

Choose any $\epsilon > 0$ and then any $y = (1/\tau)\,(\theta + 2k\pi) - (1/\tau)\epsilon$, where k is any integer. If $z = x + iy$, $\arg u(z) = 2k\pi + \epsilon + o(1)$. By choosing $x > x_0$ so that $|\,o(1)\,| < \epsilon/2$, we see $\arg u(z) = 2k\pi + \gamma$ where γ is between $\epsilon/2$ and $3\epsilon/2$. Similarly, if $x > x_1$ and $y = (1/\tau)\,(\theta + 2k\pi) + (1/\tau)\epsilon$, $\arg u(z) = 2k\pi - \delta$, where δ is between $\epsilon/2$ and $3\epsilon/2$. Choose x so that $x > x_0$ and $x > x_1$. Consider a straight line between $z_0 = x + i(1/\tau) \cdot(\theta + 2k\pi) + i(1/\tau)\epsilon$ and $z_1 = x + i(1/\tau)\,(\theta + 2k\pi) - i(1/\tau)\epsilon$. Since $\arg u(z)$ is continuous in z, at some point between z_0 and z_1, $\arg u(z) = 2k\pi$ and $u(z) > 0$.

Note that as $x \to \infty$, $|\arg h(z) - \arg g(z)\,| \to 0$, so that ϵ may be chosen arbitrarily small. Further note that for each z approaching the asymptotes, K is given by

$$K = \frac{e^{-i\theta}e^{\tau z}g(z)}{h(z)} = e^{\tau x}x^{n-m}(1 + o(1)),$$

as $x \to \infty$, so that $K \to \infty$ as $x \to \infty$.

The second part of the theorem follows by replacing z by $-z$.

Note that those values of K for which the root-locus crosses the imaginary axis increase as the root-locus becomes farther away from the origin. This means that for fixed K most of the zeros of $F(z)$ lie in the left halfplane, and also that it takes a larger value of K to force more to cross the imaginary axis. Thus only a finite part of the complex plane near the origin needs to be considered.

We need to represent $g(z)$ and $h(z)$ in factored form. Let

$$g(z) = \prod_j (z - p_j)^{\alpha_j}, \qquad h(z) = \prod_j (z - z_j)^{\beta_j},$$

where $\sum_j \alpha_j = n$ and $\sum_j \beta_j = m$. We also need the following.

DEFINITION. *The angle of departure* (*arrival*) *of the root-locus of* $F(z)$ *at* z_0 *is the angle made at* z_0 *by the tangent to the root-locus for increasing* (*decreasing*) K.

THEOREM 4.2. *As K approaches* 0, α_j *distinct branches of the positive root-locus of* $F(z)$ *approach each zero*, p_j, *of* $g(z)$. *As K approaches* ∞, β_j *distinct branches of the positive root-locus of* $F(z)$ *approach each zero*, z_j, *of* $h(z)$.

As in (3.3) this follows directly from Hurwitz' Theorem (see [10, p. 119]).

THEOREM 4.3. *If p_j is a zero of $g(z)$ of order α_j, then the positive root-locus of $F(z)$ departs from p_j making angles*

$$\phi_j = \frac{1}{\alpha_j}\Big(\sum_i \beta_i \arg(p_j - z_i) - \sum_{i \neq j} \alpha_i \arg(p_j - p_i) + \theta - \tau y_j - 2k\pi\Big), \tag{4.3}$$

where $k = 0, 1, \cdots, \alpha_j - 1$ and $y_j = \operatorname{Im} p_j$.

If z_j is a zero of $h(z)$ of order β_j, then the positive root-locus of $F(z)$ arrives at z_j making angles

$$\theta_j = \frac{1}{\beta_j}\Big(\sum_i \alpha_i \arg(z_j - p_i) - \sum_{i \neq j} \beta_i \arg(z_j - z_i) - \theta + \tau y_j + 2k\pi\Big), \tag{4.4}$$

where $k = 0, 1, \cdots, \beta_j - 1$ and $y_j = \operatorname{Im} z_j$.

Proof. (See [4].) Consider one of the branches of the positive root-locus which departs from p_j. Choose K close to 0 and let z be on that branch for that value of K. Then we have

$$\frac{Ke^{i\theta}e^{-\tau z}\prod_i (z - z_i)^{\beta_i}}{\prod_i (z - p_i)^{\alpha_i}} = 1.$$

Taking arguments,

$$\sum \beta_i \arg(z - z_i) - \sum \alpha_i \arg(z - p_i) + \theta - \tau y = 2k\pi.$$

Solving for those terms involving p_j (or z_j) and letting K approach 0 (or ∞) completes the proof.

THEOREM 4.4. *Let z_0 be any point on the real axis, $h(z)$ and $g(z)$ have real coefficients, $\{z_i\}_1^r$ and $\{p_i\}_1^s$ be the real zeros of $h(z)$ and $g(z)$ greater than z_0 and $\theta = 0$ ($\theta = \pi$). Then z_0 is contained in the positive root-locus of $F(z)$ if and only if $\sum_{i=1}^r \beta_i + \sum_{i=1}^s \alpha_i$ is even (odd).*

Proof. (See [4].) Consider the case where $\theta = 0$. Since $g(z)$ and $h(z)$ have real coefficients, zeros of $g(z)$ and $h(z)$, if complex, occur in conjugate pairs. Along the real axis, if z_1 and $\bar{z}_1$ are complex conjugates, $\arg(z_0 - z_1) + \arg(z_0 - \bar{z}_1) = 0$.

Now, as in the proof of Theorem 4.1, consider

$$u(z_0) = \frac{e^{-\tau z_0}h(z_0)}{g(z_0)};$$

$$\arg u(z_0) = -\tau y_0 + \sum_i \beta_i \arg(z_0 - z_i) - \sum_i \alpha_i \arg(z_0 - p_i).$$

On the real axis $y_0 = 0$ and the arguments from complex zeros drop out. Thus

$$\arg u(z_0) = \sum_i \beta_i \arg (z_0 - z_i) - \sum_i \alpha_i \arg (z_0 - p_i),$$

where the sums are taken over real zeros.

If $\sum_{i=1}^{r} \beta_i + \sum_{i=1}^{s} \alpha_i$ is even, then $\sum_{i=1}^{r} \beta_i - \sum_{i=1}^{s} \alpha_i$ is also even, $u(z) > 0$, and z_0 is on the positive root-locus. If $\sum_{i=1}^{r} \beta_i + \sum_{i=1}^{s} \alpha_i$ is odd, then $\sum_{i=1}^{r} \beta_i - \sum_{i=1}^{s} \alpha_i$ is odd, $u(z) < 0$ and z_0 is not on the positive root-locus.

The case where $\theta = \pi$ is similar.

THEOREM 4.5. *If the coefficients of $g(z)$ and $h(z)$ are real, $\theta = 0$ or π, and $F(z)$ has a zero of order m at $z = a$ on the real axis for $K = K_0$, $0 < K_0 < \infty$, then the positive root-locus arrives at $z = a$ making angles*

$$\theta_k = \frac{2k\pi}{m}, \qquad k = 0, 1, \cdots, m - 1,$$

and departs from $z = a$ making angles

$$\phi_k = \frac{(2k + 1)\pi}{m}, \qquad k = 0, 1, \cdots, m - 1;$$

or arrives at $z = a$ making angles

$$\theta_k = \frac{(2k + 1)\pi}{m}, \qquad k = 0, 1, \cdots, m - 1,$$

and departs from $z = a$ making angles

$$\phi_k = \frac{2k\pi}{m}, \qquad k = 0, 1, \cdots, m - 1.$$

Proof. (See [4].) Since the coefficients of $g(z)$ and $h(z)$ are real for $\theta = 0, \pi$, if zeros of $g(z) - Ke^{i\theta}e^{-\tau z}h(z)$ leave or arrive at the real axis, they do so in conjugate pairs as K varies from 0 to K_0. Thus the evenness or oddness of the number of zeros of $g(z)$ plus zeros of $h(z)$ to the right of $z = a$ is the same as that of $g(z) - K_0e^{i\theta}e^{-\tau z}h(z)$ and $h(z)$.

Write

$$F(z) = g(z) - K_0e^{i\theta}e^{-\tau z}h(z) - (K - K_0)e^{i\theta}e^{-\tau z}h(z) = 0,$$

where z is on one of the branches of positive root-locus near a. Let $g(z) - K_0e^{i\theta}e^{-\tau z}h(z) = (z - a)^m G(z)$, where $G(a) \neq 0$. $G(z)$ is real on the real axis. Then

$$(z - a)^m G(z) - (K - K_0)e^{i\theta}e^{-\tau z}h(z) = 0;$$

dividing by the second term and taking arguments,

$$m \arg (z - a) + \arg G(z) - \arg (K - K_0) - \theta + \tau y - \arg h(z) = 2k\pi,$$

where k is an integer. Thus

$$\arg (z - a) = \frac{1}{m} (-\arg G(z) + \arg h(z) + \arg (K - K_0) + \theta - \tau y + 2k\pi).$$

Now $-\arg G(z) + \arg h(z) + \theta - \tau y + 2k\pi$ approaches either an even or odd multiple of π as z approaches a since $G(z)$ and $h(z)$ are nonzero and real on the real axis and $y = 0$. Note that $\arg (K - K_0)$ is either 0 or π depending upon whether $K > K_0$ or $K < K_0$. Letting K approach K_0 completes the proof.

The most frequent occurrence is when there is a double zero of $F(z)$ on the real axis for some value of K. In this case, $\theta_0 = \pi/2$, $\theta_1 = 3\pi/2$ and $\varphi_0 = 0$, $\phi_1 = \pi$; or $\theta_0 = 0$, $\theta_1 = \pi$ and $\phi_0 = \pi/2$, $\phi_1 = 3\pi/2$.

From the preceding theorems it would appear that the root-locus for time lag systems is similar to those with no time lag. This similarity, however, is superficial. The root-locus diagrams become radically altered as simple examples such as $z - Ke^{-\tau z} = 0$ and $z^2 - Ke^{-\tau z} = 0$ (studied extensively by E. M. Wright) as well as the asymptotic theorem will testify.

A rather easy procedure has been found by Yaohan Chu [2] for constructing time lag root-locus diagrams. It consists of first constructing a diagram where there is no time lag and using this diagram to construct the time lag diagram. We refer the readers to his paper rather than reproduce it here.

Acknowledgment. The author would like to express his thanks to the referee who is responsible for many corrections and improvements in this paper.

REFERENCES

[1] RICHARD BELLMAN AND KENNETH L. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.

[2] YAOHAN CHU, *Feedback control systems with dead-time lag or distributed lag by root-locus method*, Trans. Amer. Inst. Elec. Engrs., 71 (1952), pp. 291–296.

[3] KONRAD KNOPP, *Theory of Functions*, vol. 2, Dover, New York, 1947.

[4] ALLAN M. KRALL, *An extension and proof of the root-locus method*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 644–653.

[5] ———, *A closed expression for the root-locus method*, Ibid., 11 (1963), pp. 700–704.

[6] FLOYD E. NIXON, *Principles of Automatic Controls*, Prentice Hall, Englewood Cliffs, New Jersey, 1953.

[7] EDMUND PINNEY, *Ordinary Difference-Differential Equations*, University of California Press, 1958.

[8] L. S. PONTRJAGIN, *On the zeros of some elementary transcendental functions*, Amer. Math. Soc. Transl., Ser. 2, 1 (1955), pp. 95–110.

[9] E. P. POPOV, *The Dynamics of Automatic Control Systems*, Addison-Wesley, Reading, Massachusetts, 1962.

[10] E. C. TITCHMARSH, *The Theory of Functions*, 2nd ed., Oxford University Press, Oxford, 1939.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

SOME CONDITIONS FOR THE STABILITY OF NONLINEAR TIME-DEPENDENT DIFFERENTIAL EQUATIONS*

H. H. ROSENBROCK†

1. Introduction. In an earlier paper [1] a method was given for investigating the stability of a nonlinear system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t); \qquad f(0, t) = 0. \tag{1}$$

The equation (1) was replaced by

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x}, t)\mathbf{x}, \tag{2}$$

and conditions on the elements of **A** were found which ensured stability.

This method will be applied here to the nth order differential equation

$$x^{(n)} = f(x, \dot{x}, \cdots, x^{(n-1)}, t), \tag{3}$$

which will be replaced by

$$\begin{aligned} x^{(n)} + a_n x^{(n-1)} + \cdots + a_2\dot{x} + a_1 x &= 0, \\ a_i = a_i(x, \dot{x}, \cdots, x^{(n-1)}, t)&. \end{aligned} \tag{4}$$

Conditions on the a_i will be found which ensure uniform asymptotic stability of the point $\mathbf{x} = 0$.

If $\lambda_1, \lambda_2, \cdots, \lambda_n$ are the roots of the equation

$$\lambda^n + a_n\lambda^{n-1} + \cdots + a_2\lambda + a_1 = 0, \tag{5}$$

then knowledge of the $\lambda_i(x, \dot{x}, \cdots, x^{(n-1)}, t)$ is equivalent to knowledge of the a_i. Consequently conditions on the a_i which ensure stability can be replaced by conditions on the λ_i, and it turns out to be convenient to do this. Conditions are given under which the solution $\mathbf{x} = 0$ is uniformly asymptotically stable.

2. Derivation of results. For convenience, a special form of the result proved earlier [1] will first be derived. Equation (4) is written in the form

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x}, t)\mathbf{x}, \tag{6}$$

* Received by the editors February 12, 1964, and in revised form April 20, 1964.

† Cambridge University, Cambridge, England. Now at Electronic Systems Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts.

where

$$\text{(7)} \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_n \end{bmatrix},$$

and the element x_1 of $\mathbf{x}$ replaces x in (4). This way of representing (3) is not generally unique, because a term such as x_1x_2 in the equation for $\dot{x}_n$ can be regarded as a_1x_1 with $a_1 = x_2$ or as a_2x_2 with $a_2 = x_1$. The results developed later may therefore be applied to any one of the equations (6) which can be derived from a given equation (3).

The matrix $\mathbf{A}$ in (7) may be represented by a point in an n-dimensional Euclidean space E_n having coordinates a_i. When this is done, the restrictions on $\mathbf{A}$ which ensure the asymptotic stability of $\mathbf{x} = 0$ will be expressed by the condition that $\mathbf{A}$ remains always in some region G of E_n. The way in which the region G may be obtained will appear later.

In an open region R of the space $\{\mathbf{x}\}$, equation (6) is supposed to satisfy conditions which ensure the existence and uniqueness of solutions starting from any $\mathbf{x} \in R$ at $t \geqq t_0$. Also, for all $\mathbf{x} \in R$ and for all $t \geqq t_0$ we suppose that $\mathbf{A} \in G$.

The stability conditions are developed by considering a closed, convex, bounded region H in R, which has $\mathbf{x} = 0$ as an interior point and is such that each point in the boundary of H is in at least one of a given set of hyperplanes (these will be $n + 1$ or more in number).

Let $\mathbf{n}$ be the outward unit normal to one face of H. Then we can prove the following result (see Appendix 1).

THEOREM 1. *Let the following conditions be fulfilled.*

(i) *For all* $\mathbf{x} \in R$ *and all* $t \geqq t_0$, $\mathbf{A} \in G$.

(ii) *For each face of* H *at every vertex* $\mathbf{u}$, *and for all* $\mathbf{A} \in G$,

$$\text{(8)} \qquad \mathbf{n}'\mathbf{A}\mathbf{u} \leqq -\epsilon < 0.$$

Then $\mathbf{x} = 0$ *is uniformly, asymptotically stable. All solutions of* (6) *starting at* $t_1 \geqq t_0$ *from some point* $\mathbf{x}_1 \in H$, *where* H *is in* R, *remain in* H *and tend to* $\mathbf{x} = 0$, *uniformly in* t_1, *as* $t \to \infty$. *If* R *is the whole space* $\{\mathbf{x}\}$, *the uniform asymptotic stability of* $\mathbf{x} = 0$ *is global.*

In the above statement of the method the region R was supposed to be given and the region H to be sought within R. If the closed region H is given, however, we need only verify that $\mathbf{A} \in G$ for $\mathbf{x} \in H$, $t \geqq t_0$. Similarly the existence and uniqueness of solutions need only be demonstrated for initial points $\mathbf{x}$ and times t, where $\mathbf{x} \in H$, $t \geqq t_0$; the boundary points of H will usually require special attention when this is done.

The above results will now be applied to the closed region H which is the smallest convex set containing the $2n$ points $\mathbf{v}$ in the space $\{\mathbf{x}\}$:

$$(9)\qquad \begin{aligned} \mathbf{v}_i' &= \beta_i(1, \alpha_i, \alpha_i^2, \cdots, \alpha_i^{n-1}), & i &= 1, 2, \cdots, n, \\ \mathbf{v}_{n+i}' &= -\beta_i(1, \alpha_i, \alpha_i^2, \cdots, \alpha_i^{n-1}), & i &= 1, 2, \cdots, n, \end{aligned}$$

where the α_i are all real and distinct and the β_i are positive. The bounding hyperplanes each pass through n of the points $\mathbf{v}$, of which no two may have the same value of i (see Appendix 2). For convenience call these n points $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$, where $\mathbf{u}_i$ may be either $\mathbf{v}_i$ or $\mathbf{v}_{n+i}$, and write

$$(10)\qquad \mathbf{u}_i = b_i \begin{pmatrix} 1 \\ \alpha_i \\ \alpha_i^2 \\ \vdots \\ \alpha_i^{n-1} \end{pmatrix},$$

so that the possible negative sign is absorbed in b_i.

Let $\mathbf{x}$ be a point in the hyperplane S defined by $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$. Joining $\mathbf{u}_r$ to $\mathbf{x}$ and to each of the other $\mathbf{u}_i$, and expressing the fact that these lines are coplanar, we obtain

$$(11)\qquad \begin{aligned} |\mathbf{u}_1 - \mathbf{u}_r, \mathbf{u}_2 - \mathbf{u}_r, \cdots, \mathbf{u}_{r-1} - \mathbf{u}_r,& \\ \mathbf{x} - \mathbf{u}_r, \mathbf{u}_{r+1} - \mathbf{u}_r, \cdots, \mathbf{u}_n - \mathbf{u}_r| &= 0, \end{aligned}$$

where the vectors shown are the columns of the determinant. Equation (11) gives

$$(12)\qquad \begin{aligned} &|\mathbf{u}_1 - \mathbf{u}_r, \mathbf{u}_2 - \mathbf{u}_r, \cdots, \mathbf{u}_{r-1} - \mathbf{u}_r, \mathbf{x}, \mathbf{u}_{r+1} - \mathbf{u}_r, \cdots, \mathbf{u}_n - \mathbf{u}_r| \\ &= |\mathbf{u}_1 - \mathbf{u}_r, \mathbf{u}_2 - \mathbf{u}_r, \cdots, \mathbf{u}_{r-1} - \mathbf{u}_r, \mathbf{u}_r, \mathbf{u}_{r+1} - \mathbf{u}_r, \cdots, \mathbf{u}_n - \mathbf{u}_r| \\ &= |\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n|, \end{aligned}$$

which on expanding has the form

$$(13)\qquad \mathbf{m}'\mathbf{x} = p,$$

where $\mathbf{m}$ is a certain vector. This equation expresses the fact that the projection of $\mathbf{x}$ on the vector $\mathbf{m}$ is constant.

The condition expressed by (8) can now be written

$$(14)\qquad \frac{1}{p}\mathbf{m}'\mathbf{A}\mathbf{u}_r \leqq -\epsilon < 0,$$

since it follows from (12) and (13) that $\frac{1}{p}\mathbf{m}$ is an outward normal of fixed length. Then from (7) and (10),

$$\mathbf{A}\mathbf{u}_r = \begin{pmatrix} b_r \alpha_r \\ b_r \alpha_r^2 \\ \cdots \\ b_r \alpha_r^{n-1} \\ b_r \alpha_r^n - b_r(\alpha_r^n + a_n \alpha_r^{n-1} + \cdots + a_2 \alpha_r + a_1) \end{pmatrix} \tag{15}$$

$$= \alpha_r \left(\mathbf{u}_r - \frac{b_r}{\alpha_r} \mathbf{\Phi}_r \right), \tag{16}$$

where

$$\mathbf{\Phi}_r = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \phi_r \end{pmatrix}, \tag{17}$$

and

$$\phi_r = \alpha_r^n + a_n \alpha_r^{n-1} + \cdots + a_2 \alpha_r + a_1 \tag{18}$$

$$= \prod_{1 \leqq i \leqq n} (\alpha_r - \lambda_i), \tag{19}$$

and the λ_i are defined by (5).

On using (16) and (14) we obtain (compare (12))

$$\alpha_r \left| \mathbf{u}_1 - \mathbf{u}_r, \mathbf{u}_2 - \mathbf{u}_r, \cdots, \mathbf{u}_{r-1} - \mathbf{u}_r, \mathbf{u}_r - \frac{b_r}{\alpha_r} \mathbf{\Phi}_r, \mathbf{u}_{r+1} - \mathbf{u}_r, \cdots, \mathbf{u}_n - \mathbf{u}_r \right| \div |\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n| \leqq -\epsilon < 0. \tag{20}$$

Adding the rth column in the numerator determinant to each other column gives

$$\alpha_r \left| \mathbf{u}_1 - \frac{b_r}{\alpha_r} \mathbf{\Phi}_r, \mathbf{u}_2 - \frac{b_r}{\alpha_r} \mathbf{\Phi}_r, \cdots, \mathbf{u}_n - \frac{b_r}{\alpha_r} \mathbf{\Phi}_r \right| \div |\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n| \leqq -\epsilon < 0. \tag{21}$$

Then using (17) we obtain

$$\alpha_r \left\{ 1 - \frac{b_r}{\alpha_r} \begin{vmatrix} b_1 & b_2 & \cdots & b_n \\ b_1 \alpha_1 & b_2 \alpha_2 & \cdots & b_n \alpha_n \\ \cdots & \cdots & \cdots & \cdots \\ b_1 \alpha_1^{n-2} & b_2 \alpha_2^{n-2} & \cdots & b_n \alpha_n^{n-1} \\ \phi_r & \phi_r & \cdots & \phi_r \end{vmatrix} \div b_1 b_2 \cdots b_n \Delta \right\} \leqq -\epsilon < 0, \tag{22}$$

where Δ is the Vandermonde determinant [2]

$$(23)\qquad \Delta = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdots & \alpha_n^{n-1} \end{vmatrix} = \prod_{1 \leqq i < j \leqq n} (\alpha_j - \alpha_i).$$

By using the result analogous to (23) for each cofactor of ϕ_r we obtain from (22),

$$(24)\qquad \alpha_r \left\{ 1 - \frac{b_r \phi_r}{\alpha_r} \sum_{1 \leqq k \leqq n} (-1)^{n-k} \frac{1}{b_k} \prod_{\substack{1 \leqq i < j \leqq n \\ i,j \neq k}} (\alpha_j - \alpha_i) \div \Delta \right\} \leqq -\epsilon < 0,$$

$$(25)\qquad \alpha_r \left\{ 1 - \frac{b_r \phi_r}{\alpha_r} \sum_{1 \leqq k \leqq n} \frac{(-1)^{n-k}}{b_k \prod_{1 \leqq i < k} (\alpha_k - \alpha_i) \prod_{k < j \leqq n} (\alpha_j - \alpha_k)} \right\} \leqq -\epsilon < 0.$$

Then using (19) and remembering that b_i may be chosen with either sign, we see that (25) is equivalent to

$$(26)\qquad \alpha_r - \frac{\prod_{1 \leqq i \leqq n} (\alpha_r - \lambda_i)}{\prod_{\substack{1 \leqq i \leqq n \\ i \neq r}} (\alpha_r - \alpha_i)} + \sum_{\substack{1 \leqq k \leqq n \\ k \neq r}} \left| \frac{\beta_r \prod_{1 \leqq i \leqq n} (\alpha_r - \lambda_i)}{\beta_k \prod_{\substack{1 \leqq i \leqq n \\ i \neq k}} (\alpha_k - \alpha_i)} \right| \leqq -\epsilon < 0.$$

For simplicity of application it is convenient to introduce arbitrary positive constants θ_r defined in terms of the (arbitrary) constants β_r by

$$(27)\qquad \theta_r = \left| \beta_r \prod_{\substack{0 \leqq i \leqq n \\ i \neq r}} (\alpha_r - \alpha_i) \right|^{-1}, \qquad r = 1, 2, \cdots, n,$$

and to write

$$(28)\qquad \gamma_r = \frac{\prod_{1 \leqq i \leqq n} (\alpha_r - \lambda_i)}{\prod_{\substack{0 \leqq i \leqq n \\ i \neq r}} (\alpha_r - \alpha_i)}, \qquad r = 1, 2, \cdots, n,$$

where for symmetry the constant $\alpha_0 = 0$ has been introduced. Then (26) becomes

$$(29)\qquad \alpha_r \left\{ 1 - \gamma_r + (\operatorname{sgn} \alpha_r) \, | \gamma_r | \frac{1}{\theta_r} \sum_{\substack{1 \leqq k \leqq n \\ k \neq r}} \theta_k \right\} \leqq -\epsilon < 0.$$

The result embodied in (29) contains the stability condition which was sought. If (29) is satisfied for some α_r, θ_r, $r = 1, 2, \cdots, n$, this ensures that condition (8) is fulfilled. Then if some condition $\mathbf{A} \in G$ is sufficient to satisfy (29), Theorem 1 can be applied subject to this condition. Alternatively, since the a_i are determined by the λ_i, we may prove the result subject to $\boldsymbol{\lambda} \in F$, where $\boldsymbol{\lambda} = (\lambda_i)$and F is some region of the space $\{\boldsymbol{\lambda}\}$. The result which has been proved can therefore be stated in the following way.

THEOREM 2. *Let the differential equation* (2) *have a unique solution in a region R. Let H be the closed set in R which is the smallest convex set containing the points* (9), *and let there exist a region F and a constant ϵ such that* (29) *is satisfied for all $\boldsymbol{\lambda} \in F$ with some real, distinct α_r and positive θ_r and for $r = 1, 2, \cdots, n$. Then all solutions of* (2) *starting at $t_1 \geqq t_0$ from some $\mathbf{x}_1 \in H$ remain always in H and tend uniformly, asymptotically to $\mathbf{x} = 0$ provided that $\boldsymbol{\lambda} \in F$ for all $\mathbf{x} \in H$ and all $t \geqq t_0$.*

3. Examples. Consider first the second-order differential equation

$$\ddot{x} + a_2(x, \dot{x}, t)\dot{x} + a_1(x, \dot{x}, t)x = 0. \tag{30}$$

Suppose that λ_1, λ_2 are real, and let the region F be defined by

$$\lambda_2 \leqq \alpha_2 \leqq \lambda_1 \leqq \alpha_1 < \alpha_0 = 0, \tag{31}$$

where α_1 and α_2 are distinct negative constants. Then for $\boldsymbol{\lambda} \in F$,

$$\begin{aligned} \gamma_1 &= \frac{(\alpha_1 - \lambda_1)(\alpha_1 - \lambda_2)}{\alpha_1(\alpha_1 - \alpha_2)} \leqq 0, \\ \gamma_2 &= \frac{(\alpha_2 - \lambda_1)(\alpha_2 - \lambda_2)}{\alpha_2(\alpha_2 - \alpha_1)} \leqq 0, \end{aligned} \tag{32}$$

which shows that (29) is satisfied with $\theta_1 = \theta_2 = 1$. Thus if (31) is satisfied for all $\mathbf{x}$ in the appropriate region H and for all $t \geqq t_0$, all solutions starting in H at $t_1 \geqq t_0$ remain in H and tend uniformly, asymptotically to $\mathbf{x} = 0$. This result is slightly stronger than one obtained previously [3].

Now consider the nth order equation (4), with $n > 2$, and suppose again that each λ_i is real and that F is defined by

$$0 \leqq \frac{\alpha_r - \lambda_r}{\alpha_r - \alpha_{r-1}} \leqq \theta_r, \qquad r = 1, 2, \cdots, n, \tag{33}$$

where

$$\alpha_n < \alpha_{n-1} < \cdots < \alpha_1 < \alpha_0 = 0, \tag{34}$$

$$\sum_{1 \leqq k \leqq n} \theta_k \leqq 1 - \eta < 1. \tag{35}$$

Then for $\boldsymbol{\lambda} \in F$,

$$\begin{aligned} |\gamma_r| = \Bigg| &\frac{\alpha_r - \lambda_1}{\alpha_r} \cdot \frac{\alpha_r - \lambda_2}{\alpha_r - \alpha_1} \cdot \cdots \cdot \frac{\alpha_r - \lambda_r}{\alpha_r - \alpha_{r-1}} \\ &\cdot \frac{\alpha_r - \lambda_{r+1}}{\alpha_r - \alpha_{r+1}} \cdot \cdots \cdot \frac{\alpha_r - \lambda_n}{\alpha_r - \alpha_n} \Bigg|, \end{aligned} \tag{36}$$

$$|\gamma_r| \leqq \frac{\alpha_r - \lambda_r}{\alpha_r - \alpha_{r-1}} \leqq \theta_r, \tag{37}$$

so that

$$1 - \gamma_r - |\gamma_r| \frac{1}{\theta_r} \sum_{\substack{1 \leqq k \leqq n \\ k \neq r}} \theta_k \geqq 1 - \sum_{1 \leqq k \leqq n} \theta_k \geqq \eta > 0. \tag{38}$$

It follows from (34) that (29) is satisfied with $\epsilon = -\alpha_1 \eta$. If (33) is satisfied for all $\mathbf{x}$ in the appropriate H, and for all $t \geqq t_0$, all solutions starting in H at $t_1 \geqq t_0$ remain in H and tend uniformly, asymptotically to $\mathbf{x} = 0$.

Finally, consider (30) again, but allow the λ_i to be complex. Let F be defined by

$$\begin{aligned}
&\operatorname{Re} \lambda_1 \leqq -m - \eta < 0,\\
&\operatorname{Re} \lambda_2 \leqq -m - \eta < 0,\\
&(\operatorname{Re} \lambda_1 + m)^2 + (\operatorname{Im} \lambda_1)^2 \leqq (m - \eta)^2,\\
&(\operatorname{Re} \lambda_2 + m)^2 + (\operatorname{Im} \lambda_2)^2 \leqq (m - \eta)^2,\\
&\left.\begin{aligned}\operatorname{Im} \lambda_1 &= -\operatorname{Im} \lambda_2\\ \operatorname{Re} \lambda_1 &= \operatorname{Re} \lambda_2\end{aligned}\right\} \quad \text{if} \quad \operatorname{Im} \lambda_1 = 0 \quad \text{or} \quad \operatorname{Im} \lambda_2 = 0,
\end{aligned} \tag{39}$$

where

$$0 < \eta < \frac{m}{4}. \tag{40}$$

The geometrical implication of conditions (39) is illustrated in Fig. 1.

In (29) put

$$\begin{aligned}
\alpha_1 &= -m, \quad \alpha_2 = -m^2/\eta,\\
\theta_1 &= \eta/m, \quad \theta_2 = 1 - \eta/m,
\end{aligned} \tag{41}$$

and suppose first that λ_1 and λ_2 are complex,

$$\lambda_1 = -\sigma + i\omega, \quad \lambda_2 = -\sigma - i\omega. \tag{42}$$

The left-hand side of (29) then becomes, for $r = 1$,

$$\alpha_1 \left\{1 - \gamma_1 - |\gamma_1| \frac{\theta_2}{\theta_1}\right\} \leqq \alpha_1 \left\{1 - |\gamma_1| \left(\frac{\theta_1 + \theta_2}{\theta_1}\right)\right\} \tag{43}$$

$$= -m \left\{1 - \frac{(m - \sigma)^2 + \omega^2}{m^2 - m\eta}\right\}. \tag{44}$$

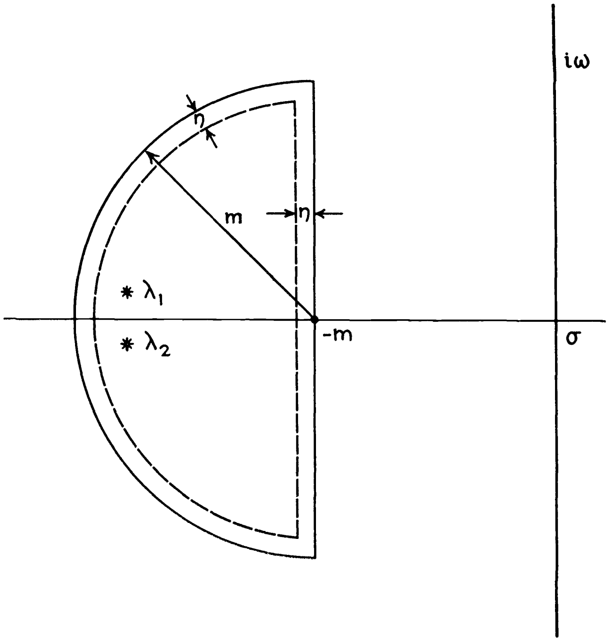


Fig. 1

By (39) and (40),

$$(45)\quad \begin{aligned}(m-\sigma)^2+\omega^2 \leqq (m-\eta)^2 &= m^2-m\eta-(m\eta-\eta^2)\\ &< m^2-m\eta-\frac{m}{2}\eta,\end{aligned}$$

so that

$$(46)\quad -m\left\{1-\frac{(m-\sigma)^2+\omega^2}{m^2-m\eta}\right\} < -m\left\{\frac{m\eta/2}{m^2-m\eta}\right\} \leqq -\frac{\eta}{2}.$$

For $r = 2$ the left-hand side of (29) becomes

$$(47)\quad \alpha_2\left\{1-\gamma_2-|\gamma_2|\frac{\theta_1}{\theta_2}\right\} \leqq \alpha_2\left\{1-|\gamma_2|\left(\frac{\theta_1+\theta_2}{\theta_2}\right)\right\}$$

$$(48)\quad = -\frac{m^2}{\eta}\left\{1-\frac{\left(\frac{m^2}{\eta}-\sigma\right)^2+\omega^2}{\left(\frac{m^2}{\eta}-m\right)^2}\right\}.$$

By (39) and (40),

$$\begin{aligned}\left(\frac{m^2}{\eta}-\sigma\right)^2+\omega^2 &\leqq \left(\frac{m^2}{\eta}-m-\eta\right)^2+(m-\eta)^2 \\ &= \left(\frac{m^2}{\eta}-m\right)^2-m^2+2\eta^2,\end{aligned} \tag{49}$$

and so

$$-\frac{m^2}{\eta}\left\{1-\frac{\left(\frac{m^2}{\eta}-\sigma\right)^2+\omega^2}{\left(\frac{m^2}{\eta}-m\right)^2}\right\}=\frac{-\eta(m^2-2\eta^2)}{(m-\eta)^2}<-\frac{\eta}{2}. \tag{50}$$

When λ_1 and λ_2 are real, it is easy to see that $|\gamma_1|$ and $|\gamma_2|$ are no greater than the values which they take in (43) and (47) respectively. Hence for $\boldsymbol{\lambda} \in F$, where F is defined by (39), the condition (29) is fulfilled for $r = 1, 2$ with $\epsilon = \eta/2$. If (39) is satisfied for all $\mathbf{x}$ in the appropriate H and for all $t \geqq t_0$, then all solutions starting in H at $t \geqq t_0$ remain in H and tend uniformly, asymptotically to $\mathbf{x} = 0$.

Appendix 1. Proof of Theorem 1. We first show that condition (8) implies that $\mathbf{n}'\mathbf{A}\mathbf{x} \leqq -\epsilon < 0$ whenever $\mathbf{A} \in G$ and $\mathbf{x}$ is a point in the face of H having $\mathbf{n}$ as its unit outward normal. For each such point $\mathbf{x}$ can be written

$$\mathbf{x} = \sum_i \alpha_i \mathbf{u}_i, \tag{51}$$

where $\mathbf{u}_i$ are the vertices of H lying in the face considered, every $\alpha_i \geqq 0$, and $\sum_i \alpha_i = 1$. Then for any given $\mathbf{A} \in G$,

$$\mathbf{n}'\mathbf{A}\mathbf{x} = \sum_i \alpha_i \mathbf{n}'\mathbf{A}\mathbf{u}_i \leqq -\epsilon \sum_i \alpha_i = -\epsilon. \tag{52}$$

Choose a vertex $\mathbf{u}_\alpha$ of H and let r_0 be the distance from $\mathbf{u}_\alpha$ to the origin. Define the sets $U(r)$, $0 \leqq r \leqq r_0$, by the property that $\mathbf{y}$ belongs to $U(r)$ if and only if $\mathbf{y} = r\mathbf{x}/r_0$, where $\mathbf{x}$ belongs to the boundary of H. The sets $U(r)$ are clearly boundaries of sets which have the same properties as H and for any $\mathbf{y}$ in $U(r)$, $\mathbf{n}'\mathbf{A}\mathbf{y} = \mathbf{n}'\mathbf{A}\mathbf{x}(r/r_0) \leqq -\epsilon r/r_0$. If we now let $V(\mathbf{y}) = r$ for $\mathbf{y}$ in $U(r)$, it is clear that V is a Lyapunov function and all of the conclusions of Theorem 1 are valid.

Appendix 2. Properties of H. The region H is defined by the points $\mathbf{v}$ in an n-dimensional space. We have to show that the bounding hyperplanes are found by selecting all sets of n points $\mathbf{v}$ having different values of i. This can be proved in the following way.

(i) Consider a hyperplane S passing through n of the points $\mathbf{v}$, no two of which have the same value of i. Then S does not include the origin. For if it did the determinant formed from the n chosen points $\mathbf{v}$ would be zero. This determinant [2] is

$$\pm\beta_1\beta_2 \cdots \beta_n \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_n^2 \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdots & \alpha_n^{n-1} \end{vmatrix} \tag{53}$$

$$= \pm\beta_1\beta_2 \cdots \beta_n \prod_{1 \leq i < j \leq n} (\alpha_j - \alpha_i),$$

which is nonzero when the α_i are distinct.

(ii) The n points $\mathbf{v}$ which are not in S lie on the same side of S as the origin. For the straight line joining each vertex in S to the origin, when continued, passes through a $\mathbf{v}$ not in S.

(iii) No hyperplane containing n points $\mathbf{v}$, of which two have the same value $i = p$, is a bounding hyperplane. For such a hyperplane includes the origin, yet does not contain two points $\mathbf{v}$ having the same value $i = q$. These last points lie on opposite sides of the hyperplane.

(iv) From (i) and (ii) it follows that each hyperplane such as S is a bounding hyperplane. By (iii) there are no others.

REFERENCES

[1] H. H. Rosenbrock, *A method of investigating stability*, Proc. I. F. A. C. Congress, Basle, 1963.

[2] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960, p. 186.

[3] H. H. Rosenbrock, *On the stability of a second-order differential equation*, J. London Math. Soc., 39 (1964), pp. 77–80.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

ON EXPONENTIAL STABILITY OF LINEAR DIFFERENTIAL SYSTEMS*

NAM P. BHATIA†

1. Introduction.

In this note we examine conditions on the linear differential system

$$\dot{x} = A(t)x, \tag{1.1}$$

(where the dot denotes the derivative with respect to t) which guarantee the existence of a quadratic form as a Liapunov function. We also give a proof of the stability theorem of Perron (see [5, pp. 142–152]) for the system

$$\dot{x} = A(t)x + f(t, x), \qquad f(t, 0) = 0,\ t \geqq 0, \tag{1.2}$$

removing thereby the restriction of boundedness of the elements $a_{ik}(t)$ of the matrix $A(t)$.

Throughout this note x denotes an n-vector in R^n, the real n-dimensional Euclidean space and $A(t)$ is an $n \times n$ real matrix whose elements $a_{ik}(t)$ are defined and continuous on $I = \{t:0 \leqq t < +\infty\}$. No assumption as to boundedness of these elements is made. $\| x \|$ stands for the euclidean norm of x. Thus $\| x \|^2 = x'x$ ($'$ denotes transpose).

It seems relevant to quote the existing results which motivated this note, with perhaps a few comments. For this we need the following definitions.

Definition 1.1. *The solution $x = 0$ of (1.1) is said to be exponentially stable, if there exist positive constants α and a such that for any solution $x(t)$ of (1.1), $x(t_0) = x_0$, the inequality*

$$\| x(t) \| \leqq \alpha \| x_0 \| \exp [-a(t - t_0)], \qquad t \geqq t_0, \tag{1.3}$$

holds.

Let $B(t)$ be a symmetric matrix with elements $b_{ik}(t) = b_{ki}(t)$ defined and continuous on I.

Definition 1.2. *The quadratic form $x'B(t)x$ is said to be positive definite if there exists a positive constant b such that*

$$x'B(t)x \geqq bx'x, \qquad t \geqq 0. \tag{1.4}$$

Definition 1.3. *The quadratic form $x'B(t)x$ will be said to have the property P if it is positive definite and if the elements $b_{ik}(t)$ of $B(t)$ are uniformly bounded on I.*

* Received by the editors June 8, 1964, and in revised form July 1, 1964.

† Department of Mathematics, Western Reserve University, Cleveland, Ohio.

Thus $x'B(t)x$ has property P if and only if there exist positive constants b_1 and b_2 such that

$$(1.5) \qquad b_1x'x \leqq x'B(t)x \leqq b_2x'x, \qquad t \geqq 0.$$

Necessary and sufficient conditions for $x'B(t)x$ to have property P are reproduced here from [2]: The quadratic form $x'B(t)x$ has property P if and only if the matrix $B(t)$ has uniformly bounded coefficients and the inequalities

$$B_k(t) > 0, \qquad k = 1, 2, \cdots, n-1;$$
$$B_n(t) = \det B(t) \geqq \delta$$

hold for $t \geqq 0$, where δ is an arbitrary but fixed positive number and $B_k(t)$ stands for the principal minor of the matrix $B(t)$ of order k.

If $V = x'B(t)x$, then we shall set

$$(1.6) \qquad V^*_{(1.1)} = x'\left[\frac{dB(t)}{dt} + A'(t)B(t) + B(t)A(t)\right]x,$$

assuming that the $b_{ik}(t)$ have continuous partial derivatives on I.

It is an elementary excercise to show [1, 4] that the existence of a quadratic form V having property P such that $-V^*_{(1.1)}$ is positive definite guarantees exponential stability of the solution $x = 0$ of (1.1). The converse of this theorem has been proved under certain restrictive conditions on $A(t)$ [1, 6, 7, 8]. Malkin [6] (also reproduced in [1]) showed that:

THEOREM 1.1 (Malkin). *If the solution $x = 0$ of (1.1) is exponentially stable and if the elements $a_{ik}(t)$ of $A(t)$ are uniformly bounded on I, then corresponding to each quadratic form $x'C(t)x$ with property P one can give a quadratic form $V = x'B(t)x$ possessing property P such that $V^*_{(1.1)} = -x'C(t)x$. And in fact the following formula determines V.*

$$(1.7) \qquad V = x'B(t)\,x = \int_t^\infty [X(\tau)X^{-1}(t)x]'C(\tau)[X(\tau)X^{-1}(t)x]\,d\tau,$$

where $X(t)$ is any fundamental matrix solution of (1.1).

Roseau [7, 8] improved upon Malkin's result and he proved:

THEOREM 1.2 (Roseau). *If the solution $x = 0$ of the system (1.1) is exponentially stable and if the matrix $A(t)$ satisfies the condition*

$$(1.8) \qquad R(s, t) = \int_t^s A(\tau)X(\tau)X^{-1}(t)\,d\tau \rightarrow 0 \quad as \quad (s - t) \rightarrow 0$$

uniformly on $s \geqq t \geqq 0$, then for every quadratic form $x'C(t)x$ having property P one can give a quadratic form $V = x'B(t)x$ having property P such that $V^*_{(1.1)} = -x'C(t)x$. In fact formula (1.7) holds.

In §2 we give an example to show that the existence of a quadratic form V such that both V and $-V^*_{(1.1)}$ have property P does not imply (1.8) and give necessary and sufficient conditions for the existence of such a quadratic form. In this connection we introduce the definition of exponential decay of solutions of the system (1.1). It turns out that the exponential decay of solutions of (1.1) implies a certain property of the trace of $A(t)$. This yields another necessary and sufficient condition for the existence of a quadratic form V such that V and $-V^*_{(1.1)}$ both have property P.

In §3 we introduce a more general notion, "the generalized exponential decay" (g.e.d.) of solutions of (1.1). This was motivated by Hale's definition of exponential stability [3]. We give necessary and sufficient conditions in terms of the existence of quadratic forms for this case.

In §4 we give a proof by the Liapunov method of Perron's theorem on stability of (1.2), without, however, the restriction of boundedness of the elements of $A(t)$ as in the classical result [5]. In this we use an idea of Yoshizawa [9]. Roseau [8] has already proved this result by another method.

2. Exponential decay. Consider the scalar differential equation

$$\dot{r} = (2t \cos t^2 - 1)r. \tag{2.1}$$

Its general solution $r(t)$, $r(t_0) = r_0$, is

$$r(t) = r_0 \exp [\sin t^2 - \sin t_0^2 - t + t_0]. \tag{2.2}$$

Notice that

$$| r(t) | \leqq | r_0 |e^2 \exp [-(t - t_0)], \qquad t \geqq t_0 ,$$

so that we have exponential stability. Malkin's formula (1.7) gives the Liapunov function

$$V = r^2 \int_t^\infty \exp [2(\sin \tau^2 - \sin t^2) - 2(\tau - t)] \, d\tau,$$

for which $V^* = -r^2$ and V satisfies

$$r^2 \frac{e^{-4}}{2} \leqq V \leqq r^2 \frac{e^4}{2},$$

so that both V and $-V^*$ have property P. Notice however that the coefficient $(2t \cos t^2 - 1)$ is neither bounded nor does Roseau's condition (1.8) hold. For in this case

$$R(s, t) = \int_t^s (2\tau \cos \tau^2 - 1) \exp [\sin \tau^2 - \sin t^2 - (\tau - t)] \, d\tau$$

$$= \exp [\sin s^2 - \sin t^2 - (s - t)] - 1.$$

Setting $s = t + 1/t$, we notice that $s - t \to 0$ as $t \to \infty$, but $R(t + 1/t, t)$ does not approach 0. Notice however that (2.2) satisfies

$$(2.3) \quad e^{-2} | r_0 | \exp [-(t - t_0)] \leqq | r(t) | \leqq e^{2} | r_0 | \exp [-(t - t_0)], \quad t \geqq t_0 .$$

DEFINITION 2.1. *The solutions of the system* (1.1) *are said to decay exponentially if there exist positive constants* a, α, b, β *such that every solution* $x(t)$, $x(t_0) = x_0$, *of* (1.1) *satisfies the inequalities*

$$(2.4) \quad \begin{aligned} &\| x_0 \| \beta \exp [-b(t - t_0)] \\ &\qquad \leqq \| x(t) \| \leqq \| x_0 \| \alpha \exp [-a(t - t_0)], \quad t \geqq t_0 . \end{aligned}$$

This leads us to the following theorem.

THEOREM 2.1. *The solutions of* (1.1) *decay exponentially if and only if there exists a quadratic form* $V = x'B(t)x$ *such that* V *and* $-V^*_{(1.1)}$ *both have property* P.

Proof. Let the solutions of (1.1) decay exponentially. Let $x'C(t)x$ be any quadratic form having property P. Set (following Malkin)

$$V = x' B(t)x = \int_t^\infty [X(\tau)X^{-1}(t)x]'C(\tau) [X(\tau) X^{-1}(t)x] \, d\tau.$$

Then $V^*_{(1.1)} = -x'C(t)x$ and V has property P. To see that V has property P, notice that there exist positive constants c_1 and c_2 such that

$$(2.5) \qquad c_1 x'x \leqq x'C(t)x \leqq c_2 x'x.$$

Then

$$c_1 \int_t^\infty \| X(\tau)X^{-1}(t)x \|^2 \, d\tau \leqq V \leqq c_2 \int_t^\infty \| X(\tau)X^{-1}(t) \, x \|^2 \, d\tau.$$

We recall now that any solution $x(t)$, $x(t_0) = x_0$, has the form $x(t) = X(t)X^{-1}(t_0)x_0$. This together with (2.4) implies

$$\begin{aligned} \| x \|^2 \beta^2 \int_t^\infty \exp [-2b(\tau - t)] \, d\tau &\leqq \int_t^\infty \| X(\tau)X^{-1}(t) \, x \|^2 \, d\tau \\ &\leqq \| x \|^2 \alpha^2 \int_t^\infty \exp [-2a(\tau - t)] \, d\tau, \end{aligned}$$

i.e.,

$$\| x \|^2 \frac{\beta^2}{2b} \leqq \int_t^\infty \| X(\tau)X^{-1}(t)x \|^2 \, d\tau \leqq \| x \|^2 \frac{\alpha^2}{2a},$$

and thus

$$c_1 \frac{\beta^2}{2b} x'x \leqq V = x'B(t)x \leqq \frac{c_2 \alpha^2}{2a} x'x.$$

Suppose now that there is a quadratic form $x'B(t)x = V$ such that V and $-V^*_{(1.1)}$ both have property P. Let c_1, c_2, b_1, b_2 be positive constants such that

$$b_1x'x \leqq V \leqq b_2x'x \quad \text{and} \quad -c_2x'x \leqq V^*_{(1.1)} \leqq -c_1x'x.$$

This implies that

$$-\frac{c_2}{b_1} V \leqq V^*_{(1.1)} \leqq -\frac{c_1}{b_2} V.$$

Along any solution $x(t)$, $x(t_0) = x_0$, set $V(t) = x'(t)B(t)x(t)$. Then this last inequality implies

$$-\frac{c_2}{b_1} V(t) \leqq \frac{dV(t)}{dt} \leqq -\frac{c_1}{b_2} V(t), \qquad t \geqq 0.$$

This yields on integration as $V(t) > 0$,

$$V(t_0) \exp\left[-\frac{c_2}{b_1}(t - t_0)\right] \leqq V(t) \leqq V(t_0) \exp\left[-\frac{c_1}{b_2}(t - t_0)\right]$$

which in turn implies (2.4) because of the property P of V. Thereby $\alpha = 1/\beta = \sqrt{b_2/b_1}$, $a = c_1/2b_2$, $b = c_2/2b_1$. This proves the theorem completely.

A similar proof as above can be constructed to prove the following.

THEOREM 2.2. *The solutions of* (1.1) *decay exponentially if and only if there exists a positive definite form V of order m with uniformly bounded coefficients such that $-V^*_{(1.1)}$ (it is also a form of order m) is positive definite and has uniformly bounded coefficients.*

This result improves Malkin's Theorem 24.5 in [4] in that the restriction of boundedness of elements of $A(t)$ is removed.

We now prove the following results.

THEOREM 2.3. *If the solutions of* (1.1) *decay exponentially then*

$$(2.6) \qquad k \leqq \int_t^\infty \left[\exp\left(2\int_t^\tau \operatorname{Tr} A(s)\, ds\right)\right] d\tau \leqq K, \qquad t \geqq 0,$$

for some positive constants k, K. ($\operatorname{Tr} A(t) = \sum_{i=1}^n a_{ii}(t)$.)

THEOREM 2.4. *If the solution $x = 0$ of* (1.1) *is exponentially stable and if there is a positive constant k such that*

$$(2.7) \qquad \int_t^\infty \exp\left(2\int_t^\tau \operatorname{Tr} A(s)\, ds\right) d\tau \geqq k, \qquad t \geqq 0,$$

then the solutions of (1.1) *decay exponentially.*

For the proof of these two theorems we need the following lemma.

LEMMA 2.1. *If the solution $x = 0$ of* (1.1) *is exponentially stable and $X(t)$ denotes any fundamental matrix solution of* (1.1), *then there exist posi-*

tive constants a_1 and a_2 such that

$$a_1x'x \leqq x'Z'(\tau, t)Z(\tau, t)x \leqq a_2x'x, \qquad \tau \geqq t \geqq 0. \tag{2.8}$$

Here $Z(\tau, t) = \operatorname{adj} X(\tau)X^{-1}(t)$.

Proof. Let $Y(\tau, t) = X(\tau)X^{-1}(t)$. Then $Z(\tau, t) = (\det Y(\tau, t))Y^{-1}(\tau, t)$. Since $Y(\tau, t)$ has uniformly bounded elements for $\tau \geqq t \geqq 0$ the same holds for the elements of $Z(\tau, t)$. Now the matrix $Z'(\tau, t)Z(\tau, t)$ is symmetric and has elements which are continuous functions of t, τ. Also $Z'(t, t)Z(t, t) = E$, the identity matrix, and $\det (Z'(\tau, t)Z(\tau, t)) = 1$ for $\tau \geqq t \geqq 0$. This implies, using the argument of Theorem 1 in [2], the existence of positive constants a_1 and a_2 such that (2.8) holds.

Proof of Theorem 2.3. Exponential decay of solutions of (1.1) implies that the quadratic form

$$x'\left[\int_t^\infty [X(\tau)X^{-1}(t)]'\,[X(\tau)\,X^{-1}(t)]\,d\tau\right]x \tag{2.9}$$

has property P. This implies, because of (2.8), that the quadratic form

$$x'\left[\int_t^\infty [X(\tau)X^{-1}(t)]'\,Z'(\tau, t)Z(\tau, t)\,[X(\tau)X^{-1}(t)]\,d\tau\right]x$$

has property P. But this last form is the same as

$$x'\left[\int_t^\infty [\det X(\tau)X^{-1}(t)]^2\,d\tau\right]x = x'\,x\int_t^\infty [\det X(\tau)\,X^{-1}(t)]^2\,d\tau.$$

However we have the well known formula for the determinant of a fundamental matrix solution of (1.1), namely

$$\det X(t) = (\det X(t_0)) \exp \int_{t_0}^t \operatorname{Tr} A(s)\,ds,$$

which gives $\det X(\tau)X^{-1}(t) = \exp \int_t^\tau \operatorname{Tr} A(s)\,ds$. This shows that property P of (2.9) implies (2.6). The theorem is thus proved.

Proof of Theorem 2.4. Note that exponential stability of the solution $x = 0$ of (1.1) together with the condition (2.7) implies (2.6). This implies that the quadratic form (2.9) has property P. If V denotes the quadratic form (2.9), then $V^*_{(1.1)} = -x'x$, so that the conditions of Theorem 2.1 are satisfied. The solutions of (1.1), therefore, decay exponentially and the theorem is proved.

It is useful perhaps to give the following theorem, which is really a corollary of the above two results, but is equivalent to them.

THEOREM 2.5. *A necessary and sufficient condition for the existence of a quadratic form V such that V and $-V^*_{(1.1)}$ both have property P is that the solution $x = 0$ of* (1.1) *be exponentially stable and the condition* (2.7) *holds.*

3. The generalized exponential decay.

DEFINITION 3.1. *The solutions of the system* (1.1) *are said to exhibit generalized exponential decay* (g.e.d.), *if there exist a nondecreasing function* $\phi(t)$ *possessing a continuous derivative such that* $\phi(t) \to \infty$ *as* $t \to \infty$ *and four positive constants* α, a, β, b *such that every solution* $x(t)$, $x(t_0) = x_0$, *of* (1.1) *satisfies the inequalities*

$$\begin{aligned} &\| x_0 \| \beta \exp [-b(\phi(t) - \phi(t_0)] \\ &\qquad \leqq \| x(t) \| \leqq \| x_0 \| \alpha \exp [-a(\phi(t) - \phi(t_0))], \quad t \geqq t_0 . \end{aligned} \tag{3.1}$$

Remark 3.1. The g.e.d. does not imply exponential stability in general, as the following example shows. Consider the scalar equation

$$\dot{r} = -\frac{1}{1+t} r, \tag{3.2}$$

whose general solution is

$$r(t) = r_0 \exp [-(\log (t+1) - \log (t_0 + 1))].$$

We have thus g.e.d. with $\alpha = a = \beta = b = 1$ and $\phi(t) = \log (t+1)$. But we do not have exponential stability.

The following theorem gives a necessary and sufficient condition for g.e.d.

THEOREM 3.1. *The solutions of the system* (1.1) *exhibit generalized exponential decay if and only if there exist two quadratic forms* $V = x'B(t)x$ *and* $W = x'C(t)x$ *having property* P *and a nonnegative continuous function* $\theta(t)$ *such that* $\int_t^\infty \theta(\tau)\, d\tau = +\infty$ *and*

$$V^* = -\theta(t)W.$$

Proof. If the solutions of (1.1) exhibit g.e.d., so that (3.1) holds, then for any quadratic form $W = x'C(t)x$ having property P we set

$$V = \int_t^\infty \phi'(\tau)\, [X(\tau)X^{-1}(t)\, x]'C(\tau)\, [X(\tau)X^{-1}(t)x]\, d\tau.$$

We notice immediately that

$$V^* = -\phi'(t)W$$

and $\int_t^\infty \phi'(\tau)\, d\tau = +\infty$ by the assumptions on $\phi(t)$. Further V has property P. For if c_1, c_2 are positive constants such that

$$c_1 x'x \leqq W = x'C(t)x \leqq c_2 x'x, \qquad t \geqq 0,$$

we have

$$c_1 \int_t^\infty \phi'(\tau) \; \| X(\tau)X^{-1}(t) \, x\|^2 \, d\tau \leqq V \leqq c_2 \int_t^\infty \phi'(\tau) \; \| X(\tau)X^{-1}(t)x \, \|^2 \, d\tau.$$

This yields, on using inequalities (3.1),

$$\frac{c_1 \, \| \, x \, \|^2 \, \beta^2}{2b} \leqq V \leqq \frac{c_2 \, \| \, x \, \|^2 \, \alpha^2}{2a},$$

which is property P for V.

Suppose now that V, W, $\theta(t)$ exist as desired in the theorem. Set $\phi(t) = \int_0^t \theta(\tau) \, d\tau$. If b_1, b_2, c_1, c_2 are positive constants such that $b_1 x'x \leqq V \leqq b_2 x'x$ and $c_1 x'x \leqq W \leqq c_2 x'x$, $t \geqq 0$, then for any solution $x(t)$, $x(t_0) = x_0$, if $V(t) = x'(t)B(t)x(t)$, we have

$$- \, \theta(t) \, \frac{c_2}{b_1} \, V(t) \leqq \frac{dV(t)}{dt} \leqq - \, \theta(t) \, \frac{c_1}{b_2} \, V(t).$$

This yields on integration

$$V(t_0) \exp\left[-b(\phi(t) - \phi(t_0))\right] \leqq V(t) \leqq V(t_0) \exp\left[-a(\phi(t) - \phi(t_0))\right],$$

where $a = c_1/b_2$ and $b = c_2/b_1$. Further this inequality yields, because of the property P of V, the inequality (3.1) with $\alpha = b_2/b_1$ and $\beta = b_1/b_2$. This completely proves the theorem.

Example 3.2. The solutions of the second order system

$$\dot{x} = y, \qquad \dot{y} = -x - \frac{2}{t} \, y,$$

which is equivalent to the single differential equation

$$\ddot{x} + \frac{2}{t} \, \dot{x} + x = 0,$$

exhibit g.e.d. For we may consider the quadratic form

$$V = x^2 + y^2 + \frac{2}{t} \, xy$$

which has property P for $t \geqq 2$. Then

$$V^* = - \, \frac{2}{t} \left(x^2 + y^2 + \frac{3}{t} \, xy \right).$$

Thus we can set $W = x^2 + y^2 + \frac{3}{t} \, xy$ and $\theta(t) = \frac{2}{t}$. Notice that W has property P for $t \geqq 2$ and $\int^\infty \frac{2}{t} \, dt = +\infty$.

Remark 3.2. The g.e.d. indeed implies uniform stability. However, it does not in general imply uniform asymptotic stability which is equivalent to exponential stability for linear systems [1].

4. The stability theorem of Perron. In the case that the solution $x = 0$ of the system (1.1) is exponentially stable, the above discussion is inconclusive as to the existence of a quadratic form V with property P such that $-V^*_{(1.1)}$ is positive definite. Following Yoshizawa [9] we can prove the following useful theorem.

THEOREM 4.1. *The solution $x = 0$ of (1.1) is exponentially stable if and only if there exists a continuous function $v(t, x)$ having the following properties:*

(i) $v(t, 0) = 0$ *and there are positive constants a and b such that*

$$a \cdot \| x \| \leqq v(t, x) \leqq b \cdot \| x \|, \qquad t \geqq 0,$$

(ii) $v(t, x)$ *is locally lipschitzian in x and if we set*

$$v^{**}_{(1.1)} = \limsup_{h \to 0+} \frac{1}{h} [v(t + h, x + hA(t)\, x) - v(t, x)],$$

then

$$v^{**}_{(1.1)} \leqq -c\| x \|,$$

where c is a positive constant.

Proof. If $v(t, x)$ is a function having properties (i) and (ii), set

$$v(t) = v(t, x(t)),$$

where $x(t)$, $x(t_0) = x_0$, is any solution of (1.1). Then property (ii) implies

$$\limsup_{h \to 0+} \frac{v(t + h) - v(t)}{h} \leqq - \frac{c}{b} v(t).$$

This immediately gives

$$v(t) \leqq v(t_0) \exp \left[- \frac{c}{a} (t - t_0) \right], \qquad t \geqq t_0,$$

which in turn implies

$$\| x(t) \| \leqq \frac{b}{a} \| x_0 \| \exp \left[- \frac{c}{b} (t - t_0) \right], \qquad t \geqq t_0,$$

which is inequality (1.3).

Now suppose that (1.3) holds. Let ρ be any positive constant such that $0 < \rho < a$. Set

$$v(t, x) = \sup_{\tau \geqq 0} [\| X(t + \tau)X^{-1}(t)x \| \exp (\rho\tau)]. \tag{4.1}$$

We assert that this v has properties (i) and (ii). Note that

$$\| x \| \leqq v(t, x) \leqq \sup_{\tau \geqq 0} [\alpha \| x \| (\exp(-a\tau)) \exp(\rho\tau)] = \alpha \| x \|,$$

which is property (i). Further for $h > 0$, we have

$$\begin{aligned} &v(t + h, x + hA(t)x) \\ &\quad = \sup_{\tau \geqq 0} [\| X(t + h + \tau)X^{-1}(t + h)(x + hA(t)x) \| \exp(\rho\tau)] \\ &\quad = \exp(-\rho h) \cdot \sup_{\tau \geqq h} [\| X(t + \tau)X^{-1}(t + h)(x + hA(t)x) \| \exp(\rho\tau)] \\ &\quad \leqq \exp(-\rho h) \cdot \sup_{\tau \geqq 0} [\| X(t + \tau)X^{-1}(t + h)(x + hA(t)x) \| \exp(\rho\tau)] \\ &\quad \leqq \exp(-\rho h)[v(t, x) + h\psi(h)], \end{aligned}$$

where $\psi(h) \to 0$ as $h \to 0$. Thus

$$\begin{aligned} v^{**}_{(1.1)} &\leqq \limsup_{h \to 0+} \frac{1}{h} [\exp(-\rho h)\{v(t, x) + h\psi(h)\} - v(t, x)] \\ &= v(t, x) \lim_{h \to 0+} \frac{\exp(-\rho h) - 1}{h} = -\rho v(t, x) \leqq -\rho \| x \|. \end{aligned}$$

This proves the theorem.

Remark 4.1. We have defined exponential stability in the case of a linear system (1.1). However, if $\dot{x} = f(t, x)$, $f(t, 0) = 0$, is a nonlinear system, where x, f are n-vectors and the function $f(t, x)$ is such that every solution $x(t, t_0, x_0)$ with $x(t_0, t_0, x_0) = x_0$ exists for $t \geqq t_0$, then if the solutions of this system with $\| x_0 \| \leqq h$ (where $h > 0$) satisfy the inequality (1.3) we say that the origin $x = 0$ of this system is exponentially stable. It is clear that Theorem 4.1 is applicable to this case with the restriction that the conditions (i) and (ii) hold for $\| x \| \leqq R$, where R is a positive constant and $v^{**}_{(1.1)}$ is replaced by

$$v^{**} = \limsup_{h \to 0+} \frac{1}{h} [v(t + h, x + hf(t, x)) - v(t, x)].$$

We will now prove a stability theorem for the system (1.2), which is essentially due to Perron. But we do not make any assumption as to boundedness of the coefficients of $A(t)$ as in the classical result [5].

THEOREM 4.2. *Suppose that the origin $x = 0$ of the system* (1.1) *is exponentially stable and the function $f(t, x)$ is continuous and satisfies the condition*

$$f(t, x) = o(\| x \|). \tag{4.2}$$

Then the origin $x = 0$ of the system (1.2) *is exponentially stable.*

Proof. Set $v(t, x)$ as in (4.1), where $X(t)$ is any fundamental matrix solution of (1.1). Then $v(t, x)$ has property (i) and is lipschitzian. Now v^{**} calculated for the system (1.2) gives

$$\begin{aligned} v^{**} &= \limsup_{h\to 0+} \frac{1}{h} [\sup_{\tau\geqq 0} [\| X(t+h+\tau)X^{-1}(t+h)(x+hA(t) \\ &\qquad + hf(t,x))\| \exp(\rho\tau)] - \sup_{\tau\geqq 0} [\| X(t+\tau)X^{-1}(t)x \| \exp(\rho\tau)]] \\ &\leqq -\rho v(t,x) + \limsup_{h\to 0+} [\sup_{\tau\geqq 0} [\| X(t+\tau+h)X^{-1}(t+h)f(t,x) \| \exp(\rho\tau)]] \\ &\leqq -\rho v(t,x) + \sup_{\tau\geqq 0} [\alpha \| f(t,x)\| \exp[-(\alpha-\rho)\tau]] \\ &= -\rho v(t,x) + \alpha \| f(t,x) \| \\ &\leqq -\rho \| x \| + \alpha \| f(t,x)\| \\ &= -\rho \| x \| \left[1 - \frac{\alpha}{\rho} \frac{\| f(t,x)\|}{\| x \|} \right]. \end{aligned}$$

Since by hypothesis $\| f(t, x) \| / \| x \| \to 0$ as $\| x \| \to 0$ uniformly on $0 \leqq t < +\infty$, we conclude the existence of a positive constant h such that

$$v^{**} \leqq -\frac{\rho}{2} \| x \| \qquad \text{for } \| x \| \leqq h, t \geqq 0.$$

This implies by Remark 4.1 that the origin $x = 0$ of the system (1.2) is exponentially stable. We notice that this conclusion goes beyond the conclusion of Roseau [8].

REFERENCES

[1] H. A. Antosiewicz, and P. Davis, *Some implications of Liapunov's conditions of stability*, J. Rational Mech. Anal., 3 (1954), pp. 447–457.

[2] N. P. Bhatia, *Kriterien für die Definitheit quadratischer Formen mit veränderlichen Koeffizienten*, Math. Nachr., 22 (1960), pp. 365–370.

[3] Jack K. Hale, *Asymptotic behavior of the solutions of differential-difference equations*, Proc. International Symposium on Nonlinear Vibrations in Kiev (to appear).

[4] W. Hahn, *Theory and Application of Liapunov's Direct Method*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[5] S. Lefschetz, *Differential Equations Geometric Theory*, 2d ed., Interscience, New York, 1963.

[6] I. G. Malkin, *On the construction of Liapunov functions for systems of linear equations*, PMM, 16 (1952), pp. 239–242.

[7] M. Roseau, *Sur la seconde methode de Liapounoff*, C. R. Acad. Sci. Paris, 252 (1961), pp. 2056–2057.

[8] ——, *Sur la stabilité de la solution $x = 0$ du systѐme differential $\dot{x} = A(t)x + f(t, x)$*, J. de Math., Paris, XLI, 3 (1962), pp. 201–212.

[9] T. Yoshizawa, *Note on the equi-ultimate boundedness of solutions of $\dot{x} = F(t, x)$*, Mem. Coll. Sci., Univ. Kyoto, Ser. A, 31 (1958).

J.SIAM CONTROL
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

ON THE SOLUTIONS OF SYSTEMS OF SECOND ORDER DIFFERENTIAL EQUATIONS WITH VARIABLE COEFFICIENTS*

I. W. SANDBERG†

Introduction. In the study of parametrically excited dynamical systems, attention is frequently focused on the properties of differential equations of the form

$$\frac{d^2x}{dt^2} + A\frac{dx}{dt} + B(t)x = 0, \qquad t \geqq 0, \tag{1}$$

in which x is an n-vector valued function of t, A is a constant $n \times n$ matrix, and $B(t)$ is an $n \times n$ matrix-valued function of t. Usually, $B(t)$ varies periodically with t and one is primarily interested in determining whether or not the trivial solution $x = 0$ is stable.

The matrix A is often associated with the damping present in a physical system. For a given $B(t)$, it is reasonable to expect that the system will be stable if the damping is sufficiently large in some sense, and it is frequently desirable to actually determine the amount of damping necessary for stabilization.

The purpose of this note is to indicate in a simple manner the utility of the type of results of [1] in obtaining sufficient conditions under which all solutions of the nonhomogeneous equation

$$\frac{d^2x}{dt^2} + A\frac{dx}{dt} + B(t)x = y, \qquad t \geqq 0, \tag{2}$$

both approach zero (i.e., the zero vector) as $t \to \infty$ and belong to $\mathcal{L}_{2n}(0, \infty)$, the set of measurable complex n-vector-valued functions of t defined on $[0, \infty)$ such that the square of the modulus of each component is integrable on $[0, \infty)$. It is assumed throughout that A and $B(t)$ are complex matrices, that the elements of $B(t)$ are uniformly bounded and piecewise continuous, but not necessarily periodic, in t, and that y is an arbitrary element of $\mathcal{L}_{2n}(0, \infty)$.

In particular, for a subclass of equations of the type (2) of direct engineering interest (in which A and $B(t)$ are Hermitian for $t \geqq 0$), we show that if the smallest eigenvalue of A exceeds a number that depends in a simple manner on the eigenvalues of $B(t)$, then all solutions approach zero as $t \to \infty$ and belong to $\mathcal{L}_{2n}(0, \infty)$.

* Received by the editors July 9, 1964, and in revised form October 5, 1964.

† Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey.

Notation and definitions. Let M denote an arbitrary matrix. We shall denote by M', M^*, and M^{-1}, respectively, the transpose, the complex-conjugate transpose, and the inverse of M. The positive square-root of the largest eigenvalue of M^*M is denoted by $\Lambda\{M\}$; and, if M is Hermitian, $\bar{\lambda}\{M\}$ and $\underline{\lambda}\{M\}$, respectively, indicate the largest and smallest eigenvalues of M. The symbol 1_n denotes the identity matrix of order n; and s and ω, respectively, are complex and real scalar variables.

Let $\mathfrak{D}$ denote the set of points at which $B(t)$ or $y(t)$ is discontinuous. By a solution of (2) we mean any complex n-vector-valued function x which is twice differentiable everywhere on $[0, \infty)$ and satisfies (2) on the complement of $\mathfrak{D}$ with respect to $[0, \infty)$.

Results and discussion. It is intuitively palatable that all solutions of (2) should possess a given property if all solutions of the corresponding equation obtained by replacing $B(t)$ with some constant matrix C possess that property and $B(t)$ is sufficiently close (in some suitable sense) to C. The following theorem, which is proved in the next section, is a precise statement consistent with this notion.

THEOREM. *Suppose that there exists a constant $n \times n$ matrix C such that*

(i) $\det[s^2 1_n + As + C] \neq 0$ *for* $\operatorname{Re}[s] \geqq 0$, *and*

(ii) $\sup\limits_{t \geqq 0} \Lambda\{B(t) - C\} \sup\limits_{-\infty<\omega<\infty} \Lambda\{[-\omega^2 1_n + i\omega A + C]^{-1}\} < 1.$

Then all solutions of (2) *belong to* $\mathfrak{L}_{2n}(0, \infty)$ *and approach zero as* $t \to \infty$.

It is sometimes possible to considerably simplify the application of this result to specific cases by exploiting the (easily verified) inequality

$$\Lambda\{M\} \leqq n \max_{j,k} |m_{jk}|,$$

in which M is an arbitrary $n \times n$ matrix with elements m_{jk}.

The following corollary of the theorem provides a simple upper bound on the amount of damping necessary to stabilize a parametrically excited dynamical system of a very general type.

COROLLARY. *Let A be a positive-definite Hermitian matrix, and let $B(t)$ be a positive-definite Hermitian matrix for $t \geqq 0$. Suppose that*

$$\inf_{t \geqq 0} \underline{\lambda}\{B(t)\} > 0,$$

and that

$$\underline{\lambda}\{A\} > (\sup_{t \geqq 0} \bar{\lambda}\{B(t)\})^{1/2} - (\inf_{t \geqq 0} \underline{\lambda}\{B(t)\})^{1/2}.$$

Then all solutions of (2) *belong to* $\mathfrak{L}_{2n}(0, \infty)$ *and approach zero as* $t \to \infty$.

For similar results concerned with the special case in which $n = 1$, see [1] and [2, §3, p. 212].

TABLE 1

b_1/b_0	Q_1	Q_2
0.024	100	41.7
0.040	50	25
0.060	33	16.7
0.080	25	12.5
0.120	17	8.3
0.160	12.5	6.3
0.20	10	5

With regard to the necessity of the condition of the corollary, it is of interest to consider the recent results of Phillips [3] concerning the determination of the value of reactance variation in order that parametric oscillations can just be maintained in a time-varying damped resonant system governed by (1) with $n = 1$ and $B(t) = b_0 - b_1 \cos \omega_p t$, in which b_0, b_1, and ω_p are positive constants. Using a semigraphical technique and the results of McLachlan concerning the Mathieu equation, Phillips finds that if, with a given $b_1(b_0)^{-1} \leqq 0.2$, the quantity $\sqrt{b_0}/A$ exceeds the appropriate value of Q_1 in Table 1, then there exists ω_p for which all solutions of (1) do not approach zero as $t \rightarrow \infty$.

The values of Q_2 given in Table 1 were computed in accordance with the corollary and are such that if, for a given $b_1(b_0)^{-1}$, $\sqrt{b_0}/A$ does not exceed the corresponding value of Q_2, then for *any* real-valued $B(t)$ such that $(b_0 - b_1) \leqq B(t) \leqq (b_0 + b_1)$ for $t \geqq 0$, all solutions of (1) approach zero as $t \rightarrow \infty$. Observe that the values of Q_1 are only roughly twice the corresponding values of Q_2.

Proofs.

Proof of the theorem. We need the following lemma* which is a very simple version of the type of result proved in [1].

LEMMA. *Let $k(\cdot)$ and $Q(\cdot)$ denote measurable $n \times n$ matrix-valued functions of t defined on $[0, \infty)$. Let $k(t)$ possess elements $\{k_{ab}(t)\}$ such that for $p = 1, 2$,*

$$\int_0^\infty |k_{ab}(t)|^p \, dt < \infty, \qquad a, b = 1, 2, \cdots, n,$$

and let the elements of $Q(t)$ be uniformly bounded on $[0, \infty)$. Let g and f denote measurable n-vector-valued functions of t defined on $[0, \infty)$ such that $g \in \mathfrak{L}_{2n}(0, \infty)$, $g(t) \rightarrow 0$ as $t \rightarrow \infty$,

$$\int_0^\zeta f(t)^* f(t) \, dt < \infty \quad \mathit{for} \quad \zeta \in (0, \infty),$$

* For the sake of completeness, a proof of the Lemma is given in the Appendix.

and

$$g(t) = f(t) + \int_0^t k(t - \tau)Q(\tau)f(\tau)\, d\tau, \qquad t \geqq 0.$$

Suppose that, with

$$K(i\omega) = \int_0^\infty k(t)e^{-i\omega t}\, dt,$$

$$\sup_{t \geqq 0} \Lambda\{Q(t)\} \sup_{-\infty<\omega<\infty} \Lambda\{K(i\omega)\} < 1.$$

Then $f \in \mathcal{L}_{2n}(0, \infty)$ *and* $f(t) \rightarrow 0$ *as* $t \rightarrow \infty$.

Now consider (2). From

$$\frac{d^2x}{dt^2} + A\frac{dx}{dt} + Cx = y - [B(t) - C]x, \qquad t \geqq 0,$$

we obtain

$$x(t) + \int_0^t k(t - \tau)[B(\tau) - C]x(\tau)\, d\tau = u(t) + v(t)$$

for $t \geqq 0$, in which u is a solution of

$$\frac{d^2u}{dt^2} + A\frac{du}{dt} + Cu = 0,$$

$$v(t) = \int_0^t k(t - \tau)y(\tau)\, d\tau,$$

and $k(\cdot)$ is the inverse Laplace transform of $[s^2 1_n + sA + C]^{-1}$.

In accordance with our assumption that $\det[s^2 1_n + sA + C] \neq 0$ for $\operatorname{Re}[s] \geqq 0$, it follows that $k(\cdot)$ satisfies the conditions of the lemma, $u(t) + v(t) \rightarrow 0$ as $t \rightarrow \infty$ (see the proof of Theorem 6 of [1]), and $(u + v) \in \mathcal{L}_{2n}(0, \infty)$. Thus, the theorem follows from a direct application of the lemma.

Proof of the corollary. Let

$$\bar{b} = \sup_{t \geqq 0} \bar{\lambda}\{B(t)\}, \quad \text{and} \quad \underline{b} = \inf_{t \geqq 0} \underline{\lambda}\{B(t)\}.$$

Consider the theorem and let $C = \frac{1}{2}(\bar{b} + \underline{b})1_n$. Then, since A is assumed to be positive-definite, it is clear that condition (i) is satisfied. From the easily verified inequality

$$\sup_{t \geqq 0} \Lambda\{B(t) - \tfrac{1}{2}(\bar{b} + \underline{b})1_n\} \leqq \tfrac{1}{2}(\bar{b} - \underline{b}),$$

and the identity

$$\Lambda\{ [- \omega^2 1_n + i\omega A + \tfrac{1}{2}(\bar{b} + \underline{b})1_n]^{-1}\} = \Lambda\{[- \omega^2 1_n + i\omega D + \tfrac{1}{2}(\bar{b} + \underline{b})1_n]^{-1}\},$$

in which $D = \operatorname{diag}[d_1, d_2, \cdots, d_n]$ with $\{d_j\}$ the eigenvalues of A, it follows that condition (ii) is met if

$$\tfrac{1}{2}(\bar{b} - \underline{b}) \sup_{\omega} \mid [- \omega^2 + \tfrac{1}{2}(\bar{b} + \underline{b}) + i\omega\underline{\lambda}\{A\}]^{-1} \mid < 1. \tag{3}$$

Inequality (3) is satisfied if $\underline{\lambda}\{A\} > (\bar{b})^{1/2} - (\underline{b})^{1/2}$. This proves the corollary.

Appendix.

Proof of the lemma. For an arbitrary $h \in \mathfrak{L}_{2n}(0, \infty)$, let $\| h \|$ be defined by

$$\| h \| = \left(\int_0^\infty h(t)^* h(t) \, dt \right)^{1/2}.$$

Assume that the hypotheses of the lemma are satisfied. Let y be an arbitrary positive number, let

$$\chi(t) = \begin{cases} 1, & \text{if } 0 \leqq t \leqq y, \\ 0, & \text{if } t > y, \end{cases}$$

and let f_y and g_y be defined by

$$f_y(t) = \begin{cases} f(t), & \text{if } 0 \leqq t \leqq y, \\ 0, & \text{if } t > y, \end{cases}$$

$$g_y(t) = \begin{cases} g(t), & \text{if } 0 \leqq t \leqq y, \\ 0, & \text{if } t > y. \end{cases}$$

Finally, let e be defined by

$$e(t) = \int_0^t k(t - \tau) Q(\tau) f_y(\tau) \, d\tau, \qquad t \geqq 0.$$

Then, from

$$g(t) = f(t) + \int_0^t k(t - \tau) Q(\tau) f(\tau) \, d\tau, \qquad t \geqq 0,$$

we obtain

$$f_y(t) = g_y(t) - \chi(t) e(t), \qquad t \geqq 0,$$

which implies that $\| f_y \| \leqq \| g_y \| + \| \chi e \| \leqq \| g \| + \| e \|$.

Consider $\| e \|$. Let $p_y(t)$ and $P_y(i\omega)$ be defined by

$$p_y(t) = Q(t) f_y(t), \qquad t \geq 0,$$

$$P_y(i\omega) = \int_0^\infty e^{-i\omega t} p_y(t) \, dt, \qquad -\infty < \omega < \infty.$$

Using Parseval's identity and the well-known extremal property of the largest eigenvalue of a Hermitian matrix, we find that

$$\begin{aligned}\| e \|^2 &= (2\pi)^{-1} \int_{-\infty}^{\infty} P_y(i\omega)^* K(i\omega)^* K(i\omega) P_y(i\omega)\, d\omega \\ &\leqq \sup_{-\infty<\omega<\infty} \Lambda^2\{K(i\omega)\} (2\pi)^{-1} \int_{-\infty}^{\infty} P_y(i\omega)^* P_y(i\omega)\, d\omega \\ &\leqq \sup_{-\infty<\omega<\infty} \Lambda^2\{K(i\omega)\} \int_0^{\infty} f_y(t)^* Q(t)^* Q(t) f_y(t)\, dt \\ &\leqq \sup_{-\infty<\omega<\infty} \Lambda^2\{K(i\omega)\} \sup_{t\geqq 0} \Lambda^2\{Q(t)\} \| f_y \|^2.\end{aligned}$$

Thus, with

$$r = \sup_{-\infty<\omega<\infty} \Lambda\{K(i\omega)\} \sup_{t\geqq 0} \Lambda\{Q(t)\},$$

we have (recall that $r < 1$ by assumption) $\| f_y \| \leqq (1 - r)^{-1} \| g \|$ for *all* $y > 0$. It follows that $f \in \mathfrak{L}_{2n}(0, \infty)$.

It remains only to show that $f(t) \to 0$ as $t \to \infty$. Since by assumption $g(t) \to 0$ as $t \to \infty$, it clearly suffices to prove that if $f \in \mathfrak{L}_{2n}(0, \infty)$ and our assumptions concerning $k(\cdot)$ and $Q(\cdot)$ are satisfied, then

$$\int_0^t k(t - \tau) Q(\tau) f(\tau)\, d\tau \to 0 \quad \text{as} \quad t \to \infty.$$

Let $p(t) = Q(t)f(t)$ for $t \geqq 0$, and observe that $p \in \mathfrak{L}_{2n}(0, \infty)$. Thus,

$$\int_0^t k(t - \tau) p(\tau)\, d\tau = (2\pi)^{-1} \int_{-\infty}^{\infty} K(i\omega) P(i\omega) e^{i\omega t}\, d\omega, \qquad t \geqq 0,$$

in which

$$P(i\omega) = \operatorname*{l.i.m.}_{T\to\infty} \int_0^T e^{-i\omega t} p(t)\, dt.$$

Since $p \in \mathfrak{L}_{2n}(0, \infty)$, and

$$\int_0^{\infty} | k_{ab}(t) |^2\, dt < \infty, \qquad a, b = 1, 2, \cdots, n,$$

it follows that the modulus of each element of the n-vector $K(i\omega)P(i\omega)$ is integrable on the ω-set $(-\infty, \infty)$. Thus, by the Riemann-Lebesgue lemma,

$$\int_{-\infty}^{\infty} K(i\omega) P(i\omega) e^{i\omega t}\, d\omega \to 0 \quad \text{as} \quad t \to \infty.$$

This completes the proof of the lemma.

REFERENCES

[1] I. W. SANDBERG, *On the $\mathcal{L}^2$-boundedness of solutions of nonlinear functional equations*, Bell System Tech. J., 43 (1964), pp. 1581–1599.

[2] V. M. STARZINSKII, *A survey of works on the conditions of stability of the trivial solution of a system of linear differential equations with periodic coefficients*, Amer. Math. Soc. Transl., Ser. 2, Vol. 1, pp. 189–237.

[3] R. F. PHILLIPS, *Parametric oscillation in a damped resonant system*, IEEE-PGCT, CT 10 (1963), pp. 512–515.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

A GENERALIZATION OF LASALLE'S "BANG-BANG" PRINCIPLE*

HUBERT HALKIN†

Introduction. This paper is devoted to some new results for the minimum time problem for a time varying linear control system.

A fundamental result in this problem is the "Bang-Bang" Principle of J. P. LaSalle [1]: *If there exists an optimal steering function then there exists a "bang-bang" steering function which is optimal.* In LaSalle's version of this principle, a bang-bang steering function is a *measurable* function taking on values at the vertices of some hypercube. In this paper we prove the same principle under the additional restriction that a bang-bang steering function be *piecewise continuous,* i.e., continuous at all but a finite number of points.

This generalization of LaSalle's Principle transforms an interesting mathematical idea into a practical engineering tool.

The two main elements of this paper are the following:

(i) We assume that the time varying linear differential equation is *piecewise analytic* with respect to the time.

(ii) We use a generalization of Lyapounov's Theorem on the convexity and closure of the range of a vector integral [2].

We conjecture that other existence theorems in the theory of optimal control [3, 4] could be similarly strengthened.

The "bang-bang" principle. The present paper should be considered as a continuation of LaSalle's original paper [1] and we shall suppose that the reader has that paper at hand.

We shall assume that the elements of the matrices $A(t)$ and $B(t)$ are functions of t which are defined and piecewise analytic for $t \geqq 0$.

By a piecewise analytic function $f(t)$ for $t \geqq 0$ we mean the following: for each $\tau > 0$ there is a finite set $\{t_0, t_1, \cdots, t_k\}$ with $t_0 = 0 < t_1 < t_2 < \cdots < t_k = \tau$, a finite collection of functions $f_1(t), f_2(t), \cdots, f_k(t)$ and an $\epsilon > 0$ such that

(i) $f(t) = f_i(t)$ for all $t \in (t_{i-1}, t_i)$ and each $i = 1, 2, \cdots, k$,

(ii) $f_i(t)$ is defined and analytic on $(t_{i-1} - \epsilon, t_i + \epsilon)$ for each $i = 1, 2, \cdots, k$.

With LaSalle we shall denote by Ω the set of admissible steering functions and by Ω° the set of *measurable* bang-bang steering functions; we introduce

* Received by the editors June 6, 1964, and in revised form July 31, 1964.

† Analytical and Aerospace Mechanics Department, Bell Telephone Laboratories, Incorporated, Whippany, New Jersey.

the new notation Ω' to denote the set of *piecewise continuous* bang-bang steering functions.

In this paper we prove the following theorems.

THEOREM 1^*. *If of all piecewise continuous bang-bang steering functions there is an optimal one relative to Ω', then it is optimal (relative to Ω).*

THEOREM 2^*. *If there is an optimal steering function then there is always a piecewise continuous bang-bang steering function that is optimal.*

These two theorems are respective generalizations of Theorems 1 and 2 of LaSalle. In order to prove Theorems 1^* and 2^*, we need only replace Lemma 1 of LaSalle by the following result.

LEMMA 1^*. *Let M be the set of all real valued measurable functions $\alpha(t)$ on $[0, 1]$ with $|\alpha(t)| \leqq 1$. Let M^1 be the subset of piecewise continuous functions in M with $|\alpha(t)| = 1$. Let $y(t)$ be any n-dimensional function which is defined and piecewise analytic on $[0, 1]$. Define*

$$K = \left\{ \int_0^1 \alpha(t) y(t)\, dt: \quad \alpha \in M \right\}$$

and

$$K^1 = \left\{ \int_0^1 \alpha(t) y(t)\, dt: \quad \alpha \in M^1 \right\}.$$

Then K^1 is closed and $K = K^1$.

Let M^0 be the subset of functions in M with $|\alpha(t)| = 1$ and let

$$K^0 = \left\{ \int_0^1 \alpha(t) y(t)\, dt: \quad \alpha \in M^0 \right\}.$$

From LaSalle's Lemma 1 we know that K^0 is closed and $K = K^0$. From standard results in measure theory we know that $K^0 \subset \bar{K}^1$, where $\bar{K}^1$ denotes the closure of the set K^1. In order to prove Lemma 1^* it remains to prove that K^1 is closed or equivalently that the set $\left\{ \int_E y(t)\, dt: E \in \mathcal{A} \right\}$ is closed, where $\mathcal{A}$ is the set of subsets of $[0, 1]$ which are the union of a finite number of intervals. This last statement is a consequence of the following theorem which has been proved in [2]:

Suppose that $\mathcal{C}$ is a class of subsets of $[0, 1]$ and that $z(t)$ is an n-dimensional vector function on $[0, 1]$ possessing the following two properties:

(i) *$\mathcal{C}$ is an algebra of Borel sets such that, if C is any element of $\mathcal{C}$, then there exists a collection $\mathfrak{D}_C$ of sets D_α, defined for every α, $0 \leqq \alpha \leqq 1$, such that $\mathfrak{D}_C \subset \mathcal{C}$, $D_1 = C$, $\mu(D_\alpha) = \alpha\mu(C)$ where μ denotes the Lebesgue measure and $D_{\alpha_1} \subset D_{\alpha_2}$ if $\alpha_1 < \alpha_2$.*

(ii) *$z(t) \in L^1(0, 1)$ and for every n-vector p,*

$$\{t: p \cdot z(t) > 0, 0 \leqq t \leqq 1\} \in \mathcal{C},$$

where the dot denotes the scalar product. Then the set

$$\left\{\int_C z(t)\,dt:\quad C \in \mathcal{C}\right\}$$

is closed and convex.

It is trivial to show that the class $\mathcal{A}$ satisfies condition (i). Let us show that $\mathcal{A}$ and $y(t)$ also satisfy condition (ii). Since each component of $y(t)$ is piecewise analytic* on [0, 1], the real-valued function $p \cdot y(t)$ is also piecewise analytic for any p. It is thus sufficient to prove that

$$\{t: f_i(t) > 0, t \in [t_{i-1}, t_i]\} \in \mathcal{A}$$

for each $i = 1, 2, \cdots, k$, where $f_i(t)$ is analytic on $(t_{i-1} - \epsilon, t_i + \epsilon)$ for $\epsilon > 0$ and equal to $p \cdot y(t)$ on (t_{i-1}, t_i). This last statement follows immediately from the fact that if an analytic function is not identically zero then the set of its zeroes has no accumulation point in the interior of the domain of analyticity.

Final remarks.

1. If the system under consideration is *normal* (in the sense of LaSalle) then Theorems 1* and 2* are an immediate consequence of LaSalle's Theorem 3 which states that all optimal steering functions u^* are of the form

$$u^*(t) = \operatorname{sgn}[\eta Y(t)],$$

where η is some nonzero n-dimensional vector. (From the definition of normality we know that no component of $\eta Y(t)$, $\eta \neq 0$, is identically zero on an interval of positive length; we know also that no component of $\eta Y(t)$, $\eta \neq 0$, has a set of zeroes with an accumulation point in the interior of the domain of analyticity.) Accordingly Theorems 1^* and 2^* are true generalizations only in the case of nonnormal systems.

2. The results of the present paper could colloquially be summarized as follows: anything which can be done with an arbitrary admissible control can also be done with a relay control with a finite number of switching times.

Acknowledgment. I am deeply grateful to Dr. Lucien W. Neustadt for his many valuable comments and suggestions.

* In LaSalle's paper the vector $y(t)$ denotes a column of the matrix function $Y(t) = X^{-1}(t)\ B(t)$, where

$$\dot{X}(t) = A(t)\,X(t) \quad \text{and} \quad X(0) = I.$$

In this paper we have assumed that the elements of the matrices $A(t)$ and $B(t)$ are piecewise analytic. This assumption implies that the matrices $X(t)$ and $Y(t)$ are piecewise analytic.

REFERENCES

[1] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.

[2] H. Halkin, *On a generalization of a theorem of Lyapounov*, J. Math. Anal. Appl., to appear.

[3] L. W. Neustadt, *The existence of optimal control in the absence of convexity conditions*, Ibid., 7 (1963), pp. 110–117.

[4] A. F. Filippov, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ., Ser. Mat., Meh., Astr., Fiz., Him., 2 (1959), pp. 25–32; English trans., this Journal, 1 (1962), pp. 76–84.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

SINGULAR OPTIMAL CONTROLS FOR A CLASS OF MINIMUM EFFORT PROBLEMS*

DONALD R. SNOW†

1. Introduction. Suppose a system is described by the linear second-order scalar differential equation

$$\ddot{x} + a(t)\dot{x} + b(t)x + c(t) = u(t). \tag{1.1}$$

We are to find the function $u(t)$, $| u | \leqq 1$, which drives the system from a given initial state $[x(0), \dot{x}(0)]$ to the origin $(0, 0)$ within a given finite time T while minimizing the integral of $| u(t) |$. As is proved in [1, Appendix 1], any completely controllable second-order linear system can be described by (1.1) by using a suitable nonsingular linear transformation. Aspects of this minimum effort problem or solutions to specific examples have been discussed by various authors [2], [3], [4], [5]. A general discussion of minimum effort control problems is given in [6].

Since 1958 the standard method of solution of such problems has been the application of the Pontryagin maximum principle [7], [8] or some modification of it. It will be shown in this paper that for those systems where

$$b(t) \equiv \dot{a}(t), \tag{1.2}$$

this method breaks down for large regions of initial states because of insufficient characterization of the optimal control. Problems for which (1.2) holds therefore belong to the class of singular optimal control problems. The maximum principle, which is a necessary but not sufficient condition for optimality, is not useful here since an infinity of control functions are described by it, not all of which are optimal.

This paper will present a direct analytic method of solution for these singular problems. We first ask the question: for which initial states in the phase plane $(x, \dot{x})$ do there exist controls $u(t)$, $| u | \leqq 1$, which drive the system to the origin $(0, 0)$ within the given time T? This set of initial states is called the *T-controllable region* of the phase plane and equations

* Received by the editors March 23, 1964, and in revised form September 3, 1964.

† Department of Mathematics, University of Minnesota, Minneapolis, Minnesota. The research for this paper was done in connection with the author's Ph.D. dissertation submitted to the Department of Mathematics, Stanford University, Stanford, California. It has been supported by the Research Laboratories, Lockheed Missiles and Space Company, Palo Alto, California, through their Graduate Study Program. This paper is a revised version of a paper presented at the Joint AIAA-IMS-SIAM-ONR Symposium on Control and System Optimization, held at the United States Naval Postgraduate School, Monterey, California, January 27–29, 1964.

for the boundary of this region will be derived in terms of T, $a(t)$, and $c(t)$. The subregions of T-controllable initial states for which the maximum principle is not useful will be called the *singular subregions* and will be completely described.

We will then determine the optimal controls for any T-controllable initial state and it will be shown that for each initial state in the singular subregions there are infinitely many optimal controls, most of which are not bang-coast-bang. To have a well-posed problem (unique solution) in these cases an additional constraint must be specified. Such a constraint might be the requirement that the response time be smallest. This time would be less than or equal to the upper limit on the time, T. For initial states in the T-controllable region that are not in the singular subregions, the optimal controls will be shown to be unique. These controls could be obtained by the maximum principle method as well as by this new method; however, this new method is direct and eliminates the need of guessing the adjoint variable initial conditions.

Reference [9] is a recent paper on singular control problems in which the system equations as well as the integral to be optimized are linear in the controls. That paper does not cover the present case since the integral here depends on $| u(t) |$ which is nonlinear on $-1 \leqq u \leqq 1$.

To illustrate the method, the results will be applied to an important special case, namely to the system $\ddot{x} + a\dot{x} = u$, where a is a constant. The regions and optimal controls will be shown.

2. Statement of the problem. We will consider the following scalar differential equation

$$\ddot{x} + a(t)\dot{x} + b(t)x + c(t) = u(t), \tag{2.1}$$

where $a(t)$ and $b(t)$ are given continuous functions, $c(t)$ is a given piecewise continuous function, $u(t)$ is the control function, assumed to be in the class U, where U is the set of all piecewise continuous functions $u(t)$, $| u | \leqq 1$, on $0 \leqq t \leqq T$, where T is a given (fixed) number. The initial state and desired terminal state on $x(t)$ are

$$[x(0), \dot{x}(0)] = (\alpha_1, \alpha_2) \tag{2.2}$$

and

$$[x(T), \dot{x}(T)] = (0, 0). \tag{2.3}$$

The functional or payoff function to be minimized is taken to be

$$J[u] = \int_0^T | u(t) | \, dt. \tag{2.4}$$

The general problem may then be formulated as: find a function $u(t) \in U$ such that the solution of the differential equation (2.1) has initial state (2.2), terminal state (2.3), and makes $J[u]$ as small as possible.

We first find conditions on the coefficients of the differential equation such that the usual method of the Pontryagin maximum principle does not determine the optimal controls. Let $x_1 = x$ and $x_2 = \dot{x}$ be the state variables and p_1 and p_2 be the adjoint variables of the Pontryagin method. Using these variables the equivalent first order system, the Hamiltonian, and the adjoint system are:

$$\text{(2.5) Equivalent system:} \qquad \begin{aligned} \dot{x}_1 &= x_2 , \\ \dot{x}_2 &= -a(t)x_2 - b(t)x_1 - c(t) + u. \end{aligned}$$

$$\text{Hamiltonian:} \qquad \begin{aligned} H &= p_1\dot{x}_1 + p_2\dot{x}_2 - | u | \\ &= p_1x_2 - p_2a(t)x_2 - p_2b(t)x_1 - p_2c(t) + p_2u - | u | \cdot \end{aligned}$$

$$\text{(2.6) } \qquad \text{Adjoint system:} \qquad \begin{aligned} \dot{p}_1 &= -\frac{\partial H}{\partial x_1} = p_2b(t), \\ \dot{p}_2 &= -\frac{\partial H}{\partial x_2} = p_2a(t) - p_1 . \end{aligned}$$

According to the maximum principle, any optimal control $u(t)$ maximizes H as a function of u. Since the u-dependent terms $\bar{H}$ in H are given by $\bar{H} = p_2u - | u |$, we can see that an optimal u must satisfy the relation

$$\text{(2.7)} \qquad u(t) = \begin{cases} +1 & \text{if } p_2(t) > 1, \\ 0 & \text{if } | p_2(t) | < 1, \\ -1 & \text{if } p_2(t) < -1. \end{cases}$$

The general method now is to guess the values of $p_1(0)$, $p_2(0)$ (or use an iteration method as in [10]), use the given α_1 , α_2 , and integrate the system (2.5)-(2.6), choosing $u(t)$ according to requirement (2.7). If this trajectory does not have the desired terminal state (0, 0), we guess new values for $p_1(0)$, $p_2(0)$ and try again.

DEFINITION. A *regular initial state* is a T-controllable initial state (α_1 , α_2) for which the Pontryagin maximum principle characterizes at least one optimal control, i.e., there is at least one set of adjoint variable initial conditions such that (2.7) gives an optimal control. All other T-controllable initial states are called *singular initial states.*

PROPOSITION 2.1. *There are singular initial states for this problem if the coefficients of the differential equation* (2.1) *satisfy the relation*

$$\text{(2.8)} \qquad b(t) \equiv \dot{a}(t).$$

If this condition holds only on subintervals of $0 \leqq t \leqq T$, the problem can be treated separately on these "subintervals of singularity". Note that condition (2.8) does not depend on $c(t)$.

Proof of Proposition 2.1. We first show that (2.8) is equivalent to $p_2(t) \equiv \pm 1$ being permissible solutions of the adjoint system (2.6). When (2.8) holds, the adjoint system leads to the equation $\ddot{p}_2 = a(t)\dot{p}_2$, which has the solutions $p_2(t) \equiv \pm 1$. Conversely, substituting $p_2(t) \equiv \pm 1$ into (2.6), we get $p_1(t) \equiv \pm a(t)$ and $\dot{p}_1(t) \equiv \pm b(t)$. Hence (2.8) follows.

Now note that when $p_2(t)$ is given, (2.7) determines the optimal control uniquely except at the values of t for which $p_2(t) = \pm 1$. If these exceptional values of time are isolated, the value of the optimal control at these times is immaterial. But when $p_2(t) \equiv +1$, we have $\bar{H} = u - |u|$; and then, any $u \geqq 0$ will maximize $\bar{H}$ giving the maximum value $\bar{H} = 0$. When $p_2 \equiv -1$, any $u \leqq 0$ will maximize $\bar{H}$, again giving the maximum value $\bar{H} = 0$. Thus when $p_2(t) \equiv \pm 1$, the maximum principle does not characterize the optimal control except to indicate that it must not change sign. It will be shown in §5 that the two choices of adjoint variable initial conditions $[p_1(0), p_2(0)] = \pm[a(0), 1]$ which give $p_2(t) \equiv \pm 1$ correspond to two regions of T-controllable initial states. It will also be shown in §5 that there are no other adjoint variable initial conditions that lead to optimal controls for initial states in these regions. Hence these are regions of singular initial states.

3. A lemma. We now state a lemma without proof.

LEMMA 3.1. *Given any $K(t) \geqq 0$, continuous and strictly monotone increasing on $[0, T]$, and two real numbers A and M with $0 \leqq A \leqq MT$. Let $U_M{}^A$ be the class of all piecewise continuous functions with $0 \leqq u(t) \leqq M$ on $[0, T]$ or which*

$$\int_0^T u(t)\, dt = A.$$

For each $u \in U_M{}^A$, let

$$\int_0^T K(t)u(t)\, dt = B_u. \tag{3.1}$$

Let

$$u_L(t) = \begin{cases} M & \text{if} \quad 0 \leqq t \leqq A/M, \\ 0 & \text{if} \quad A/M < t \leqq T, \end{cases}$$

and

$$u_R(t) = \begin{cases} 0 & \text{if} \quad 0 \leqq t < T - A/M, \\ M & \text{if} \quad T - A/M \leqq t \leqq T. \end{cases}$$

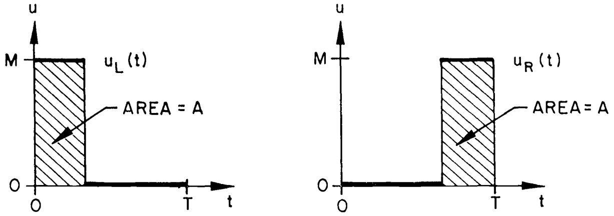


FIG. 1. *Functions u_L and u_R*

(See Fig. 1). Then, for all $u \in U_M{}^A$, we have

$$\int_0^T K(t)u_L(t)\,dt \leqq B_u \leqq \int_0^T K(t)u_R(t)\,dt, \tag{3.2}$$

and, for any number B in this range, we can find at least one $u \in U_M{}^A$ such that $B_u = B$.

When $K(t) \equiv t$ this lemma is just a formal statement of the fact that for a given area, the shape that has its centroid farthest to the left is the tallest allowable rectangle with this area at the left end. Similarly, for the right end. In all cases except when B is at the lower or upper limit in (3.2) there are infinitely many suitable u's; for example, the rectangle of maximum height with area A, shorter but wider rectangles located properly, triangles with area A, etc. This lemma will be the basis for the characterization of the optimal controls and T-controllable region.

4. Characterization of the T-controllable region. Throughout the remainder of the paper we assume that the differential equation satisfies (2.8). Then it can be written in the form

$$\frac{d}{dt}[\dot{x} + a(t)x] = u(t) - c(t). \tag{4.1}$$

DEFINITION. A control $u(t)$ will be called *admissible* if it is in U and if it drives the system from the initial state (2.2) to the terminal state (2.3).

Notation. The T-controllable region in state space (see definition in §1) will be denoted by R. The subregion for which there are nonnegative admissible controls will be denoted by P and that for which there are nonpositive admissible controls by N. The interiors of these regions will be denoted by R°, P°, and N°, respectively.

We now reduce the problem to an equivalent formulation.

THEOREM 4.1. *Given the initial state (α_1, α_2), let*

$$A = -\alpha_2 - a(0)\alpha_1 + \int_0^T c(t)\,dt, \qquad B = \alpha_1 + \int_0^T K(t)c(t)\,dt, \tag{4.2}$$

where

$$K(t) = \int_0^t \exp\left[\int_0^s a(r)\,dr\right] ds. \tag{4.3}$$

Then, $u \in U$ is admissible if and only if

$$A = \int_0^T u(t)\,dt, \tag{4.4}$$

$$B = \int_0^T K(t)u(t)\,dt. \tag{4.5}$$

Proof. The general solution of (4.1) can be written

$$x(t) = \frac{K(t)}{K'(t)}\left\{\int_0^t [u(s) - c(s)]\,ds + \alpha_2 + a(0)\alpha_1\right\} + \frac{1}{K'(t)}\left\{\alpha_1 - \int_0^t K(s)[u(s) - c(s)]\,ds\right\}.$$

From this it can easily be shown that $x(T) = \dot{x}(T) = 0$ if and only if

$$\int_0^T [u(s) - c(s)]\,ds + \alpha_2 + a(0)\alpha_1 = 0$$

and

$$\alpha_1 - \int_0^T K(s)[u(s) - c(s)]\,ds = 0.$$

Hence, for $u \in U$, u satisfying these conditions is equivalent to u being admissible, and these conditions are just (4.4) and (4.5).

For a given initial state, (4.4) and (4.5) give a characterization of the subclass of U that are admissible controls. To describe the region R we find the values that A and B can assume for controls in U and then determine the corresponding initial states (α_1, α_2). Region R and its subregions are independent of the functional to be minimized; but, it will be shown in §5 that the regions P° and N° are the regions of singular initial states when the particular functional (2.4) is used.

THEOREM 4.2. *An initial state (α_1, α_2) is in region R if and only if*

$$-T \leqq A \leqq T, \tag{4.6}$$

and

$$2\int_0^{(A+T)/2} K(t)\,dt \leqq B + \int_0^T K(t)\,dt \leqq 2\int_{(T-A)/2}^T K(t)\,dt, \tag{4.7}$$

where A, B, and $K(t)$ are defined by (4.2) and (4.3). The boundaries of the *region are given by considering equality to hold on the left and right, respectively, in inequality* (4.7).

Proof. For any admissible $u(t)$, let $v(t) = u(t) + 1$. Then $0 \leqq v \leqq 2$. In terms of $v(t)$, conditions (4.4) and (4.5) become

$$A + T = \int_0^T v(t)\,dt, \tag{4.8}$$

$$B + \int_0^T K(t)\,dt = \int_0^T K(t)v(t)\,dt. \tag{4.9}$$

Conversely, any piecewise continuous $v(t)$, $0 \leqq v \leqq 2$, that satisfies these two conditions corresponds to an admissible $u(t)$.

Now suppose $(\alpha_1, \alpha_2) \in R$. Then there is a $u \in U$ satisfying (4.4) and (4.5), and hence a $v(t)$ satisfying (4.8) and (4.9). Since $0 \leqq v \leqq 2$, (4.8) shows that $0 \leqq A + T \leqq 2T$, or $-T \leqq A \leqq T$. Use of Lemma 3.1 with $M = 2$ and the left hand sides of (4.8) and (4.9) as the constants gives

$$\int_0^T K(t)v_L(t)\,dt \leqq B + \int_0^T K(t)\,dt \leqq \int_0^T K(t)v_R(t)\,dt,$$

or

$$2\int_0^{(A+T)/2} K(t)\,dt \leqq B + \int_0^T K(t)\,dt \leqq 2\int_{T-(A+T)/2}^T K(t)\,dt.$$

Now suppose that the initial state (α_1, α_2) is such that A and B satisfy (4.6) and (4.7). By Lemma 3.1, there is at least one $v(t)$, $0 \leqq v \leqq 2$, that satisfies (4.8) and (4.9). This $v(t)$ corresponds to a $u \in U$ that is admissible. Hence $(\alpha_1, \alpha_2) \in R$.

THEOREM 4.3. *An initial state* (α_1, α_2) *is in region* P *if and only if*

$$0 \leqq A \leqq T, \tag{4.10}$$

and

$$\int_0^A K(t)\,dt \leqq B \leqq \int_{T-A}^T K(t)\,dt, \tag{4.11}$$

where A, B, *and* $K(t)$ *are defined by* (4.2) *and* (4.3). *The boundaries of the region are given by considering equality to hold on the left and right, respectively, in inequality* (4.11).

Proof. Suppose $(\alpha_1, \alpha_2) \in P$. Then there is a $u \in U$, $0 \leqq u \leqq 1$, satisfying (4.4) and (4.5). Condition (4.4) with $0 \leqq u \leqq 1$ gives $0 \leqq A \leqq T$. Use of Lemma 3.1 with $M = 1$ gives

$$\int_0^T K(t)u_L(t)\,dt \leqq B \leqq \int_0^T K(t)u_R(t)\,dt.$$

By the meaning of u_L and u_R in Lemma 3.1, this is just (4.11).

Now suppose initial state (α_1, α_2) is such that A and B satisfy (4.10) and (4.11). By Lemma 3.1 with $M = 1$ there is at least one $u \in U$ with $0 \leqq u \leqq 1$ that satisfies (4.4) and (4.5). By Theorem 4.1 this u is admissible and since $u \geqq 0$, we have $(\alpha_1, \alpha_2) \in P$.

By analogous reasoning with $-u(t)$, we can prove the following result.

THEOREM 4.4. *An initial state* (α_1, α_2) *is in region* N *if and only if*

$$-T \leqq A \leqq 0, \tag{4.12}$$

and

$$-\int_{T+A}^{T} K(t)\,dt \leqq B \leqq -\int_0^{-A} K(t)\,dt, \tag{4.13}$$

where A, B, *and* $K(t)$ *are defined by* (4.2) *and* (4.3). *The boundaries of the region are given by considering equality to hold on the left and right, respectively, in inequality* (4.13).

By (4.10), (4.12), (4.11), and (4.13), the regions P and N are disjoint except for the initial state corresponding to $A = B = 0$. This unique point in R is the initial state for which the control $u(t) \equiv 0$ is admissible, and is the origin if and only if $c(t) \equiv 0$. When $c(t) \equiv 0$, the T-controllable region is symmetric with respect to the origin since then $-u$ in place of u in (4.1) leads to the solution $-x$ instead of x. For this case region N is the image of region P under reflection in the origin. Region R transformed into AB-space is always symmetric with respect to the origin (regardless of $c(t)$) since conditions (4.4) and (4.5) are linear in $u(t)$ and the class U is symmetric.

We will now subdivide the set of initial states $R \sim (P^\circ \cup N^\circ)$ into four mutually disjoint sets R_i, $i = 1, 2, 3, 4$. It will be shown in §6 that the set $R \sim (P^\circ \cup N^\circ)$ is the set of regular initial states and that the optimal controls for initial states in each of the regions R_i have a specific form. To define these regions, we let

$$\begin{aligned} B_L &= 2\int_0^{(A+T)/2} K(t)\,dt - \int_0^T K(t)\,dt, \\ B_R &= 2\int_{(T-A)/2}^{T} K(t)\,dt - \int_0^T K(t)\,dt; \end{aligned} \tag{4.14}$$

and if $0 \leqq A \leqq T$, we let

$$B_{LP} = \int_0^A K(t)\,dt, \qquad B_{RP} = \int_{T-A}^{T} K(t)\,dt; \tag{4.15}$$

or, if $-T \leqq A \leqq 0$, we let

$$B_{LN} = -\int_{T+A}^{T} K(t)\,dt, \qquad B_{RN} = -\int_0^{-A} K(t)\,dt. \tag{4.16}$$

The description of region R, inequality (4.7), then becomes $B_L \leqq B \leqq B_R$. Region P, which requires $0 \leqq A \leqq T$, is described (see (4.11)) by $B_{LP} \leqq B \leqq B_{RP}$. Since $P \subset R$, we have $B_L \leqq B_{LP} \leqq B_{RP} \leqq B_R$. This inequality can also be verified by use of the monotonicity and positivity of $K(t)$. Then for $0 \leqq A \leqq T$ the regular initial states satisfy

$$R_1 : B_{RP} \leqq B \leqq B_R, \tag{4.17}$$

or else

$$R_2 : B_L \leqq B \leqq B_{LP}. \tag{4.18}$$

Analogously, we find that for $-T \leqq A \leqq 0$ the regular initial states satisfy

$$R_3 : B_L \leqq B \leqq B_{LN}, \tag{4.19}$$

or else

$$R_4 : B_{RN} \leqq B \leqq B_R. \tag{4.20}$$

For the example considered in §7, these regions are shown in Fig. 5.

5. Optimal controls for singular initial states. Given an initial state $(\alpha_1, \alpha_2) \in R$ we can now ask for those admissible controls which minimize $J[u]$ in (2.4) since we know at least one admissible control exists. Theorem 4.1 gives a more convenient characterization of the admissible controls. Note that the quantity $|A|$ in (4.2) and (4.4) gives a lower bound for J for the given initial state since

$$|A| = \left|\int_0^T u(t)\,dt\right| \leqq \int_0^T |u(t)|\,dt = J. \tag{5.1}$$

THEOREM 5.1. *Any admissible control that does not change sign is optimal.*

Proof. Equality holds in (5.1) if and only if $u(t)$ does not change sign. Since $|A|$ is the lower bound for J, any such u must be optimal.

Regions P and N contain all initial states for which there are admissible controls that do not change sign. Hence for initial states in P and N, Theorem 5.1 shows that the optimal controls are precisely those admissible controls that do not change sign. For these initial states there are also admissible controls that do change sign, but these are not optimal.

To determine the optimal controls for an initial state in P, we compute A and B from (4.2) and find all nonnegative functions $u(t)$ that satisfy (4.4) and (4.5). If the initial state is on the boundary of P, B will be at the lower or upper limit in (4.11). Hence the optimal control for such an initial state is the unique nonnegative admissible control for this state, namely the u_L or u_R shown in Fig. 1. But if the initial state is in P°, the interior of region P, strict inequality will hold on both sides of (4.11). Then, by the discussion following Lemma 3.1, there are infinitely many suitable u's. These are of arbitrary form as long as they satisfy $0 \leqq u \leqq 1$, have area A, and have $K(t)$-weighted integral B. The analogous result holds in region N.

An additional constraint may be imposed if it is desired to select only one of the infinitely many minimum effort optimal controls for a given initial state in P° or N°. This might be desirable in a computational scheme. An example of such a constraint would be to select that minimum effort control which makes the system arrive at the origin at the earliest time $t_2 \leqq T$. This constraint is simple to apply when $c(t) \equiv 0$, since then it can easily be shown that the system must be at the origin at the instant the control is turned off for the last time. Thus this control is the one with the area A concentrated as far to the left as permitted by B. It is shown in Fig. 2 for initial states in P°. This is a minimum fuel control regardless of $c(t)$, but if $c(t) \not\equiv 0$ on $t_2 \leqq t \leqq T$, the system does not reach the origin at time t_2, and hence this control is not a minimum time–minimum fuel control.

We have shown that there are infinitely many optimal controls for each initial state in $P^\circ \cup N^\circ$ and that most of these controls are composed of arcs that do not lie on $u = +1, 0$, or -1. Hence they are not characterized by the Pontryagin maximum principle (see (2.7)). We now show the following.

THEOREM 5.2. *Region $P^\circ \cup N^\circ$ is the region of singular initial states, i.e., there are no adjoint variable initial conditions that give a Pontryagin optimal control* (*see* (2.7)) *for any initial state in $P^\circ \cup N^\circ$.*

Proof. Condition (2.8) allows the adjoint system (2.6) to be solved explicitly as $p_2(t) = [a(0)p_2(0) - p_1(0)]K(t) + p_2(0)$, where $K(t)$ is given by (4.3). This is strictly monotone increasing or decreasing unless the initial conditions $[p_1(0), p_2(0)] = \pm[a(0), 1]$ are used, in which case $p_2(t) \equiv \pm 1$. For initial states in P° the only optimal controls with $u = 0$ and 1 have at least two jumps; for example $u = (1, 0, 1)$ or $u = (1, 0, 1, 0)$, where this notation means that $u \equiv 1$ during the first interval of time, $u \equiv 0$ during the next interval, etc. (Controls with only one jump, namely

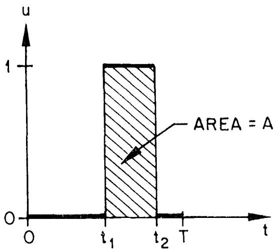


FIG. 2. *Minimum time–minimum fuel control in region P^c when $c(t) \equiv 0$*

$u = (1, 0)$ or $u = (0, 1)$, correspond to boundary points of P.) To obtain a control with at least two jumps from (2.7), $p_2(t)$ would have to change the sign of its slope at least once. But $p_2(t)$ is strictly monotone unless it is identically ± 1, in which case (2.7) does not characterize the optimal controls sufficiently anyway. Thus there are no adjoint variable initial conditions that give an optimal control by the Pontryagin maximum principle method for any initial state in P°. The analogous argument works for region N°. Hence $P^\circ \cup N^\circ$ is contained in the region of singular initial states. Theorem 6.1 (below) will show that all the T-controllable initial states outside $P^\circ \cup N^\circ$ are regular and therefore that $P^\circ \cup N^\circ$ is precisely the set of singular initial states.

The proof of Proposition 2.1 showed that $p_2(t) \equiv +1$ leads to nonnegative optimal controls. Thus we can think of the adjoint variable initial conditions $[p_1(0), p_2(0)] = [a(0), 1]$ as corresponding to all initial states in the singular region P°. Likewise, we can think of $[-a(0), -1]$ as corresponding to the singular region N°.

6. Optimal controls for regular initial states. Since P and N contain all the T-controllable initial states which have admissible controls of one sign, all the admissible controls for the remaining initial states must change sign. For any admissible $u(t)$, let

$$u(t) = u^+(t) - u^-(t), \tag{6.1}$$

where

$$u^+(t) = \begin{cases} u(t) & \text{if } u(t) \geqq 0, \\ 0 & \text{if } u(t) < 0, \end{cases}$$

and

$$u^-(t) = \begin{cases} -u(t) & \text{if } u(t) \leqq 0 \\ 0 & \text{if } u(t) > 0. \end{cases}$$

This decomposition is unique since at each t we require at least one of u^+ or u^- to be zero. Using this decomposition, $| u(t) | = u^+(t) + u^-(t)$.

THEOREM 6.1. *The initial states in regions R_i, $i = 1, 2, 3, 4$, are all of the regular initial states and have unique optimal controls of the forms*

$$\begin{aligned} u &= (-1, 0, +1) \quad \text{in} \quad R_1, \\ u &= (+1, 0, -1) \quad \text{in} \quad R_2, \\ u &= (+1, 0, -1) \quad \text{in} \quad R_3, \\ u &= (-1, 0, +1) \quad \text{in} \quad R_4, \end{aligned}$$

where this notation means that in R_1, the optimal control has $u \equiv -1$ during the first interval of time, $u \equiv 0$ during the next interval, and $u \equiv +1$ during the last interval, etc.

Proof. Using decomposition (6.1), the functional (2.4) and condition (4.4) become

$$J = \int_0^T u^+(t)\,dt + \int_0^T u^-(t)\,dt; \qquad A = \int_0^T u^+(t)\,dt - \int_0^T u^-(t)\,dt.$$

Adding, we have

$$J + A = 2\int_0^T u^+(t)\,dt.$$

Since A is a fixed number for a given initial state, minimizing J is equivalent to minimizing

$$J_1 = \int_0^T u^+(t)\,dt. \tag{6.2}$$

Using the transformation $v(t) = u(t) + 1$, the admissible controls correspond to the piecewise continuous functions $v(t)$, $0 \leqq v \leqq 2$, that satisfy (4.8) and (4.9), and minimizing (6.2) is equivalent to minimizing the area above the line $v = 1$ (the cross-hatched area in Fig. 3). Hence we need to find the function $v(t)$ that satisfies $0 \leqq v \leqq 2$, has a given area and a given $K(t)$-weighted integral, and has the smallest area above the line $v = 1$.

Consider an initial state in R_2 (see (4.18)); i.e., $0 \leqq A \leqq T$ and $B_L \leqq B \leqq B_{LP}$. We will show how to obtain the unique optimal control by modification of the optimal control for the same A but with $B = B_{LP}$. This is a boundary point of region P and the optimal control is given uniquely by $u_L(t)$ in Lemma 3.1. In terms of $v(t)$, it has total area $A + T$ and $K(t)$-weighted integral $B_{LP} + \int_0^T K(t)\,dt$ (see (4.8) and (4.9)) and is shown in Fig. 4. We will modify this function keeping the same total area but so that

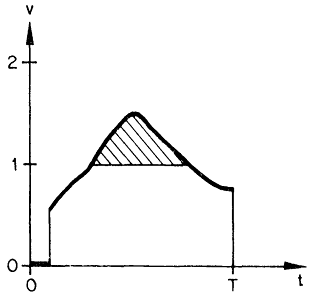


FIG. 3. *Geometrical interpretation of the problem*

the value of the $K(t)$-weighted integral (a "generalized centroid") decreases until it reaches the value $B + \int_0^T K(t)\,dt$. At the same time we need to do this by transferring as little area as possible from below $v = 1$ to above this line in order to minimize (6.2). Since $K(t)$ is monotone increasing we can accomplish this by transferring vertical strips from the far right below the line to as far left as possible above the line. Such a strip is shown in Fig. 4. By transferring enough such strips to move the generalized centroid left to the desired value, we obtain the optimal control in terms of $v(t)$. It is unique since any other admissible control corresponds to a $v(t)$ with more area above the line $v = 1$. Transformed back to $u(t)$, the unique optimal control for any initial state in R_2 is of the form $u = (+1, 0, -1)$ where the first jump is taken as early as possible consistent with the corresponding A and B. An analogous argument works for initial states in R_1, R_3, and R_4. It is readily seen that these unique optimal controls could have been obtained by application of the Pontryagin maximum principle (see (2.7)) with a suitable choice of adjoint variable initial conditions. Hence, these are regular initial states. Since Theorem 5.2 shows that the remaining initial states are all singular, region $R_1 \cup R_2 \cup R_3 \cup R_4$ is the set of all regular initial states.

The geometrical argument in the above proof could be replaced by a simple variational argument but the geometry gives more insight and shows the uniqueness immediately.

To compute the optimal control for a regular initial state we only need to find the jump points t_1 and t_2 since we know the general form of the control. To do this, we use (4.4) and (4.5) to get two relations between the jump points and let t_1 be the smallest value that satisfies these two equations; then, we find the corresponding t_2. This method is direct and eliminates the need of guessing the adjoint variable initial conditions of the Pontryagin maximum principle method. The optimal trajectory corresponding to an optimal control may be obtained by using the general solution of (4.1) which was given in the proof of Theorem 4.1.

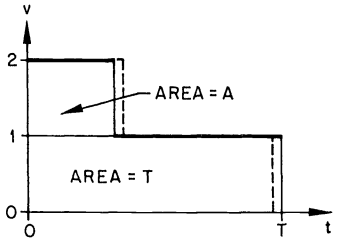


FIG. 4. *Optimal control for* $0 \leqq A \leqq T$ *and* B_{LP}

7. Example. Consider the differential equation $\ddot{x} + a\dot{x} = u$, where a is a constant. If $a = 0$, this describes the "$1/s^2$" plant. If $a \neq 0$, it describes a system with damping term proportional to the velocity. Both of these systems have many physical applications; e.g., satellite attitude controls, a body moving in a viscous fluid, etc.

From (4.2) and (4.3) we get

$$K(t) = \frac{1}{a}(e^{at} - 1),$$

$$A = -\alpha_2 - a\alpha_1,$$

$$B = \alpha_1 .$$

Using Theorem 4.2 and these quantities, the T-controllable region has boundary curves

$$\alpha_1 = -\frac{\alpha_2}{a} - \frac{2}{a^2}\log\left[\cosh\left(\frac{aT}{2}\right) - \frac{a}{2}e^{-aT/2}\alpha_2\right],$$

$$\alpha_1 = -\frac{\alpha_2}{a} + \frac{2}{a^2}\log\left[\cosh\left(\frac{aT}{2}\right) + \frac{a}{2}e^{-aT/2}\alpha_2\right],$$

where $-T \leqq \alpha_2 + a\alpha_1 \leqq T$. Note that one of these curves can be obtained from the other by replacing (α_1, α_2) by $(-\alpha_1, -\alpha_2)$. This occurs since $c(t) \equiv 0$ implies the T-controllable region is symmetric with respect to the origin (see the discussion following Theorem 4.4).

By Theorem 4.3, the singular subregion P° has boundary curves

$$\alpha_1 = -\frac{\alpha_2}{a} - \frac{1}{a^2}\log(1 - a\alpha_2),$$

$$\alpha_1 = -\frac{\alpha_2}{a} + \frac{1}{a^2}\log(1 + ae^{-aT}\alpha_2),$$

where $0 \leqq -\alpha_2 - a\alpha_1 \leqq T$. By Theorem 4.4, the singular subregion N° has boundary curves

$$\alpha_1 = -\frac{\alpha_2}{a} + \frac{1}{a^2}\log(1 + a\alpha_2),$$

$$\alpha_1 = -\frac{\alpha_2}{a} - \frac{1}{a^2}\log(1 - ae^{-aT}\alpha_2),$$

where $-T \leqq -\alpha_2 - a\alpha_1 \leqq 0$. Since $c(t) \equiv 0$, the boundaries of N° are those of P° with (α_1, α_2) replaced by $(-\alpha_1, -\alpha_2)$.

The boundaries of the regular regions R_i, given by equality in (4.17)–(4.20), consist of parts of the boundaries of R, P°, N°, and the line $A = 0$;

i.e., $\alpha_2 + a\alpha_1 = 0$ or $\alpha_1 = -\alpha_2/a$. (Note that the expression on the right here occurs in each boundary curve equation.)

The T-controllable region R and its subregions are shown in Fig. 5 for the case where $a = 1$, $T = 1$. For other values of a, the regions have the same general shape. Taking the limit as $a \to 0$, we can show that all the boundary curves for the case $a = 0$ are parabolas.

For any initial state in P the optimal controls are given by the u's, $0 \leqq u \leqq 1$, with area $\alpha_1/(-\alpha_2 - a\alpha_1)$. By the definition of region P there are such functions and in P° there are infinitely many (see the discussion following Theorem 5.1). Similarly, for region N.

For initial states in the regular regions R_i, the optimal controls are unique and have the forms given in Theorem 6.1. Using (4.4) and (4.5) to get relations between the jump points t_1 and t_2 for initial states in R_2, we find that

$$t_1 = \frac{1}{a}\log\left\{\frac{1 - a\alpha_2 + e^{aT}}{2} - \sqrt{\frac{(1 - a\alpha_2 + e^{aT})^2}{4} - e^{a(T+A)}}\right\},$$

$$t_2 = T + A - t_1 .$$

The optimal trajectory for an initial state in R_2 is shown in Fig. 5. The jump points and optimal trajectories for initial states in the other regular regions may be computed in a similar manner.

It may be shown that the optimal trajectory from any regular initial state enters one of the singular initial state subregions. It might appear then that these trajectories could not be unique. However, the size of the singular subregions depends on the available time left to go, which at the initial time is T. As the trajectory is traversed, the time left to go decreases and hence the size of the singular subregions *for the available time to go* decreases. Using this fact, it can be shown that the trajectory never enters a singular subregion corresponding to the available time left and that for $t_1 \leqq t \leqq T$, the trajectory is actually on the boundary of the decreasing singular subregion. Hence the regular initial state optimal trajectories can be unique even though they enter the initial state singular subregions.

8. Concluding remarks. We have shown a characterization of a class of singular minimum fuel problems and an analytic method for their solution. Equations were given describing the region of T-controllability and the subregions of singularity for which the Pontryagin maximum principle is not useful. We have also determined the optimal controls for any initial state in the T-controllable region and have shown that there are infinitely many optimal controls for each initial state in the singular regions. The T-controllable region does not depend on the functional to be minimized, of course, and hence holds for all differential equations (2.1) satisfying condition (2.8) with controls assumed to be piecewise continuous and $| u | \leqq 1$.

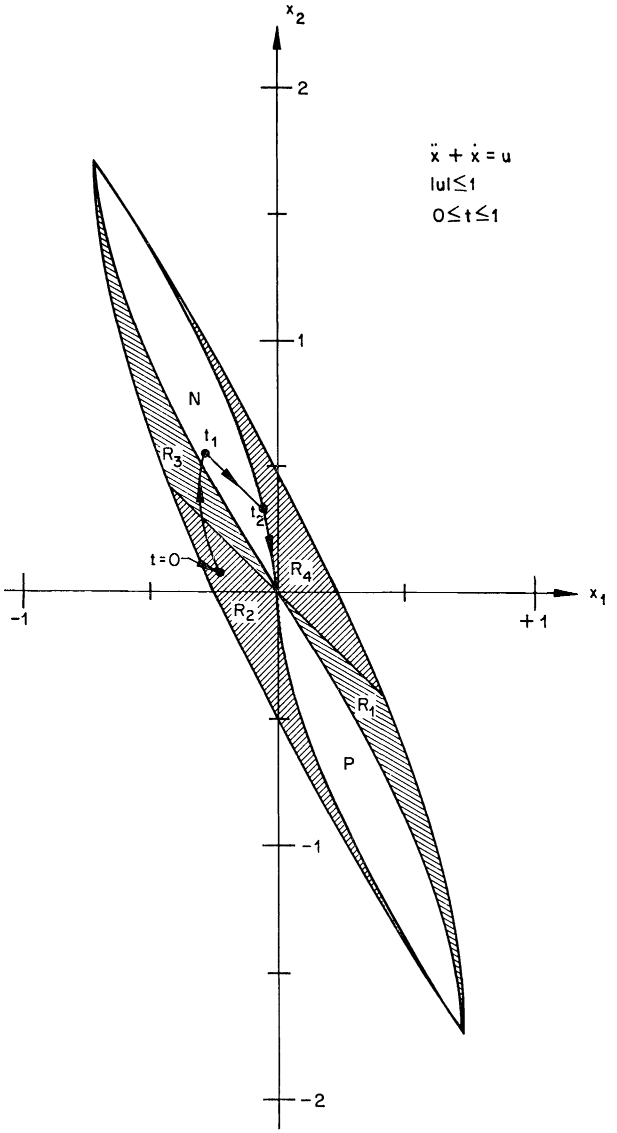


FIG. 5. *T-controllable region and subregions for example considered with* $a = 1$, $T = 1$.

With a few minor changes the results hold for measurable functions instead of piecewise continuous functions. This method can be extended to many additional classes of problems, some of which will be discussed in future papers by the author.

Acknowledgment. My thesis advisor at Stanford, Professor M. M. Schiffer, deserves much of the credit for this work, and my supervisor at Lockheed, Dr. John V. Breakwell, has also been very helpful. I would also like to thank the reviewer for his helpful suggestions on the manner of presentation of this material.

REFERENCES

[1] W. M. Wonham and C. D. Johnson, *Optimal bang-bang control with quadratic performance index*, Trans. ASME Ser. D. J. Appl. Mech., 86 (1964), pp. 107–115.

[2] I. Flügge-Lotz and H. Marbach, *The optimal control of some attitude control systems for different performance criteria*, Ibid., 85 (1963), pp. 165–176.

[3] H. O. Ladd, Jr. and B. Friedland, *Minimum fuel control of a second-order linear process with a constraint on time-to-run*, Ibid., 86 (1964), pp. 160–168.

[4] J. S. Meditch, *On minimal fuel satellite attitude controls*, Fourth Joint Automatic Control Conference, Minneapolis, Minnesota, preprints by A.I. Ch.E., 1963, pp. 558–564.

[5] M. Athans, *Minimum-fuel control of second-order systems with real poles*, Ibid., pp. 232–240.

[6] L. W. Neustadt, *Minimum effort control systems*, this Journal, 1 (1962), pp. 16–31.

[7] L. S. Pontryagin, et al., *The Mathematical Theory of Optimal Processes*, Interscience, Wiley, New York, 1962.

[8] R. E. Kopp, *Pontryagin maximum principle*, Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, Chap. 7.

[9] C. D. Johnson and J. E. Gibson, *Singular solutions in problems of optimal control*, IEEE Trans. on Automatic Control, AC-8 (1963), pp. 4–15.

[10] L. W. Neustadt and B. Paiewonsky, *On synthesizing optimal controls*, Presented at Second Congress of the International Federation of Automatic Control (IFAC), Basel, Switzerland, 1963.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

AN OPTIMAL REGULATOR PROBLEM*

ALBERT CHANG†

1. Problem statement. We consider an optimal regulator problem for systems whose behavior may be described by the linear differential equation

$$\dot{x}(t) = Ax(t) + Bu(t). \tag{1}$$

In (1), A and B are $n \times n$ and $n \times r$ matrices of real constants, respectively. The vector $x(t) = (x^1(t), x^2(t), \cdots, x^n(t))$ is called the state (at time t) and the function $t \rightarrow u(t) = (u^1(t), u^2(t), \cdots, u^r(t))$, the control. The control $t \rightarrow u(t)$, $0 \leqq t < \infty$, will be called an admissible control if it is measurable and if, in addition, for each t, $0 \leqq t < \infty$,

$$u(t) \in \Omega = \{(u^1, u^2, \cdots, u^r) \mid | u^i | \leqq 1, \quad i = 1, 2, \cdots, r\}.$$

The set of all admissible controls will be denoted by U. We shall regard two controls identical if they are equal almost everywhere (a.e.).

Let $u \in U$ and x_0 be any point in the state space R^n. The solution of (1) corresponding to u, which satisfies the initial condition $x(0; x_0, u) = x_0$, is given by

$$x(t; x_0, u) = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)\, ds. \tag{2}$$

The point x_0 is called the initial state, and u is said to transfer x_0 to the origin if $\lim_{t\rightarrow\infty} x(t; x_0, u) = 0$. For convenience, $x(t; x_0, u)$ will often be abbreviated to $x(t; u)$ or $x(t)$, when the initial state or control, or both, intended are clear from context.

We consider the function $J: U \times R^n \rightarrow R$ defined by

$$J(u, x_0) = \frac{1}{2}\int_0^\infty \langle x(t; x_0, u), Qx(t; x_0, u)\rangle + \langle u(t), Ru(t)\rangle\, dt. \tag{3}$$

Here Q is a nonnegative definite matrix and R is a positive definite matrix. J is well defined although it may assume the value $+\infty$ for certain u and x_0.

The following conditions are assumed to hold:

* Received by the editors July 24, 1964.

† Research Laboratory, International Business Machines Corporation, Monterey and Cottle Roads, San Jose, California. This research was done at the Department of Electrical Engineering, University of California, Berkeley, California, and was supported by the Air Force Office of Scientific Research under Contracts AF-AFOSR 292-63 and AF-AFOSR 139-63.

(a) The pair (A, B) is controllable [2], i.e.,

$$\operatorname{rank}\,[B, AB, \cdots, A^{n-1}B] = n.$$

(4) (b) The pair (A, Q) is observable [2], i.e.,

$$\operatorname{rank}\,[Q, AQ, \cdots, A^{n-1}Q] = n.$$

(c) The autonomous system $\dot{x} = Ax$ is Ljapunov stable.

We consider the following:

PROBLEM. Given the system (1), an initial state x_0, and the conditions (4), let $T \subset U$ be the set of all admissible controls which transfer x_0 to the origin. The problem is to find a $u \in T$ such that $J(u, x_0) = \inf_{v \in T} J(v, x_0)$.

A solution to the problem will be called an optimal control. It will be shown that there exists a unique optimal control for any choice of the initial state. Then, using a result of Rozonoer, we shall show Pontryagin's maximum principle, suitably strengthened, gives necessary and sufficient conditions for a control to be optimal. We then consider the problem of constructing a feedback control which yields optimal control.* It is shown that an optimal feedback control is a linear function of the state in a neighborhood of the origin. However, no closed form expression is derived for it that is valid for all states. Nevertheless, it may be computed by running the system backwards in time, in a way similar to that proposed by LaSalle for the time optimal control problem [1]. Our final result is that the optimal feedback control is a continuous function of the state which, in principle, makes this computation feasible.

Essentially the same problem under consideration was discussed in a series of papers by A. M. Letov [3], [4]. Krasovskii and Letov showed later that the solution proposed in [3] and [4] may be correct only for special choices of initial state [5]. More recently, Johnson and Wonham gave examples showing that the results in [3] and [4] are, in general, incorrect [6].

2. Existence of optimal controls.

In this section we shall show that our problem has a unique solution for any choice of the initial state.

If $0 < \lambda < 1$ and $u_1, u_2 \in U$, then the function u defined by

$$u(t) = \lambda u_1(t) + (1 - \lambda)u_2(t), \qquad 0 \leqq t < \infty, \tag{5}$$

is also a member of U because Ω is convex. Using (2) it is easy to verify that for arbitrary initial states x_0^1, x_0^2,

$$\begin{aligned} &x(t; \lambda x_0^1 + (1 - \lambda)x_0^2, u) \\ &\qquad = \lambda x(t; x_0^1, u_1) + (1 - \lambda)x(t; x_0^2, u_2), \quad 0 \leqq t < \infty. \end{aligned} \tag{6}$$

* See §5 for a definition of the term "feedback control."

In particular, setting $x_0^1 = x_0^2 = x_0$, note that if u_1 and u_2 transfer x_0 to the origin, then so does u.

Furthermore, now observe that the integrand of (3),

$$f(x, u) = \tfrac{1}{2}(\langle x, Qx\rangle + \langle u, Ru\rangle), \tag{7}$$

is a convex function; for all x_1, x_2, u_1, u_2 and $0 < \lambda < 1$,

$$\begin{aligned} &f(\lambda x_1 + (1 - \lambda)x_2, \lambda u_1 + (1 - \lambda)u_2) \\ &\qquad \leqq \lambda f(x_1, u_1) + (1 - \lambda)f(x_2, u_2). \end{aligned} \tag{8}$$

Moreover, since R is positive definite, equality holds in (8) only if $u_1 = u_2$. The following lemma will be needed in what follows.

LEMMA 1. *Let u_1, $u_2 \in U$ and let x_0^1 and x_0^2 be arbitrary initial states. Then if u is defined by* (5),

$$J(u, \lambda x_0^1 + (1 - \lambda)x_0^2) \leqq \lambda J(u_1, x_0^1) + (1 - \lambda)J(u_2, x_0^2).$$

Moreover, equality holds only if $u_1 = u_2$ whenever the quantities are finite.

Proof. Using (6) and (8), we have

$$\begin{aligned} J(u, \lambda x_0^1 + (1 - \lambda)x_0^2) &= \int_0^\infty f(x(t; \lambda x_0^1 + (1 - \lambda)x_0^2, u), u(t))\, dt \\ &= \int_0^\infty f(\lambda x(t; x_0^1, u_1) + (1 - \lambda)x(t; x_0^2, u_2), \\ &\qquad\qquad \lambda u_1(t) + (1 - \lambda)u_2(t))\, dt \\ &\leqq \lambda \int_0^\infty f(x(t; x_0^1, u_1), u_1(t))\, dt \\ &\qquad + (1 - \lambda) \int_0^\infty f(x(t; x_0^2, u_2), u_2(t))\, dt \\ &= \lambda J(u_1, x_0^1) + (1 - \lambda)J(u_2, x_0^2). \end{aligned}$$

In view of the property of f mentioned above, equality holds only if $u_1(t) = u_2(t)$, a.e., whenever the quantities are finite. This proves the lemma.

We deduce immediately from Lemma 1 the following corollary.

COROLLARY 1. *If an optimal control exists for initial state x_0, then it is unique.*

Proof. Suppose u_1 and u_2 are optimal controls for initial state x_0, but $u_1(t) \neq u_2(t)$, a.e. Setting $x_0^1 = x_0^2 = x_0$ in Lemma 1, $J(u, x_0) < \lambda J(u_1, x_0) + (1 - \lambda)J(u_2, x_0)$, which contradicts the optimality of u_1 and u_2.

Consider the function $V: R^n \to R$ defined by

$$V(x_0) = \inf_{u \in U} J(u, x_0).$$

Given an initial state x_0, the existence of an optimal control will be proven in two steps. First, we show there is a $u \in U$ such that $J(u, x_0) = V(x_0)$, and then we show that u necessarily transfers x_0 to the origin. It is then clear that u is the optimal control for the initial state x_0.

PROPOSITION 1. *Given an initial state* x_0, *there exists a* $u \in U$ *such that* $J(u, x_0) = V(x_0)$.

Proof. It follows from a lemma due to LaSalle [1] that there is a $t_1 < \infty$ and an admissible control $t \to u'(t)$ defined on $0 \leqq t \leqq t_1$ such that $x(t_1; x_0, u') = 0$. Setting $u(t) = u'(t)$ for $0 \leqq t \leqq t_1$ and $u(t) = 0$ for $t > t_1$, $J(u, x_0) < \infty$. Therefore, since R is positive definite, it suffices to consider only controls belonging to some bounded set, say

$$B = \left\{ u \in U \mid \| u \| = \int_0^\infty \sqrt{\langle u(t), u(t) \rangle}\, dt \leqq M \right\}.$$

Note that B is compact in the weak L_2 topology.

Let $d = \inf_{v \in B} J(v, x_0) = V(x_0)$. We have to show there is a $v \in B$ such that $J(u, x_0) = V(x_0)$. Let $\{u_n\}$ be a sequence in B such that

$$J(u_n, x_0) \leqq d + \frac{1}{2^n}.$$

Since B is weakly compact, there is a $u \in B$ and a subsequence of $\{u_n\}$ converging weakly to u. We show u is the required control.

The Banach-Saks theorem (renumbering indices if necessary) states there is a sequence of controls $\{v_n\}$ converging to u in norm with $v_k = (u_1 + u_2 + \cdots + u_k)/k$. Now $v_k \in B$ because B is convex. Then since $J(u, x_0)$ is a convex function of u (Lemma 1),

$$J(v_k, x_0) \leqq \frac{1}{k} \sum_{i=1}^{k} J(u_i, x_0) \leqq d + \frac{1}{k} \sum_{i=1}^{k} \frac{1}{2^i} < d + \frac{1}{k}.$$

Since $v_k \to u$ in norm, it is clear that $x(t; v_k) \to x(t; u)$ for each t. In addition, there is subsequence of $\{v_k\}$ converging to u, a.e.; so we may assume $v_k(t) \to u(t)$, a.e. Therefore, f being continuous,

$$\lim_{k \to \infty} f(x(t; v_k), v_k(t)) = f(x(t; u), u(t)), \qquad \text{a.e.}$$

The proposition then follows from the inequalities,

$$\begin{aligned} d &\leqq \int_0^\infty f(x(t; u), u(t))\, dt \\ &= \int_0^\infty \lim_k f(x(t; v_k), v_k(t))\, dt \\ &\leqq \liminf_k \int_0^\infty f(x(t; v_k), v_k(t))\, dt \\ &= \liminf_k J(v_k, x_0) = d. \end{aligned}$$

The third line follows from the Fatou-Lebesgue theorem.

In view of Proposition 1, we may define

$$V(x_0) = \min_{u \in U} J(u, x_0).$$

The function V has the following properties.

LEMMA 2. (a) *$V(x_0) = 0$ if and only if $x_0 = 0$.*
(b) *V is a convex function (and therefore continuous).*
(c) *If $a \leqq 1$, $V(ax_0) \leqq V(x_0)$.*

Proof. (a) If $V(x_0) = 0$, then $J(u_0, x_0) = 0$ for some $u_0 \in U$. Inasmuch as R is positive definite and Q is nonnegative definite, $u_0(t) = 0$, a.e.; therefore,

$$V(x_0) = \int_0^\infty \langle e^{At}x_0, Qe^{At}x_0\rangle \, dt.$$

It follows from the observability of (A, Q) that $x_0 = 0$ [2]. Conversely, it is clear that $V(0) = 0$.

(b) Given $x_0^1, x_0^2 \in R^n$, choose $u_1, u_2 \in U$ so that $V(x_0^1) = J(u_1, x_0)$ and $V(x_0^2) = J(u_2, x_0^2)$. Letting $0 < \lambda < 1$, the statement follows upon application of Lemma 1,

$$\begin{aligned} V(\lambda x_0^1 + (1 - \lambda)x_0^2) &\leqq J(\lambda x_0^1 + (1 - \lambda)x_0^2, \lambda u_1 + (1 - \lambda)u_2) \\ &\leqq \lambda J(u_1, x_0^1) + (1 - \lambda)J(u_2, x_0^2) \\ &= \lambda V(x_0^1) + (1 - \lambda)V(x_0^2). \end{aligned}$$

(c) This follows from (b) and the fact that V is nonnegative.

We now are in a position to show that a control which minimizes $J(u, x_0)$ necessarily transfers x_0 to the origin.

PROPOSITION 2. *Given x_0, suppose $u \in U$ and $J(u, x_0) = V(x_0)$. Then u transfers x_0 to the origin.*

Proof. Let $B_\delta = \{x \in R^n \mid \|x\| = \langle x, x\rangle^{1/2} = \delta\}$. Since V is continuous, $V(x) > 0$ for each $x \in B_\delta$, and B_δ is compact,

$$\inf_{x \in B_\delta} V(x) = b > 0.$$

It follows from Lemma 2(c) that $V(x) \geqq b$ whenever $\|x\| \geqq \delta$.

Now observe that

$$V(x_0) = V(x(t_1; x_0, u)) + \int_0^{t_1} f(x(t; x_0, u), u(t)) \, dt.$$

The last relation expresses the fact that the control $t \to u'(t) = u(t + t_1)$, $0 \leqq t \leqq \infty$, minimizes $J(\cdot, x(t_1; x_0, u))$. Since

$$V(x_0) = \lim_{t_1 \to \infty} \int_0^{t_1} f(x(t; x_0, u), u(t)) \, dt,$$

it follows that there exists $T < \infty$ such that for all $t_1 > T$, $V(x(t_1\,;\,x_0\,,u)) < b$, which implies $\| x(t_1\,;x_0\,,u)\| < \delta$. Since δ was arbitrary, this shows u transfers x_0 to the origin, completing the proof of the proposition.

Combining Corollary 1 and Propositions 1 and 2, we have the following theorem.

THEOREM 1. *For each initial state x_0, there exists a unique optimal control.*

3. Necessary conditions for optimality. In this section, we state Pontryagin's maximum principle, as it applies to our problem, and modify it to a form which yields both necessary and sufficient conditions for a control to be optimal.

The Hamiltonian for the problem is

$$\mathcal{H}(\bar{\psi}, \bar{x}, u) = \psi^0 f(x, u) + \langle \psi, Ax + Bu \rangle, \tag{9}$$

where

$$\psi = (\psi^1, \psi^2, \cdots, \psi^n),$$
$$\bar{\psi} = (\psi^0, \psi^1, \cdots, \psi^n),$$
$$x = (x^1, x^2, \cdots, x^n),$$
$$\bar{x} = (x^0, x^1, \cdots, x^n).$$

We consider the Hamiltonian system

$$\dot{x}^i(t) = \frac{\partial \mathcal{H}}{\partial \psi^i}, \qquad i = 0, 1, \cdots, n, \tag{10}$$

$$\dot{\psi}^i(t) = -\frac{\partial \mathcal{H}}{\partial x^i}, \qquad i = 0, 1, \cdots, n. \tag{11}$$

If $x_0 = (x^1, x^2, \cdots, x^n)$ is the initial state for (1), the corresponding initial condition for (10) is $\bar{x}_0 = (0, x^1, x^2, \cdots, x^n)$. The solution for the x^0 component of $\bar{x}$ is expressed in terms of $x(t; x_0\,, u)$ by the integral

$$x^0(t) = \int_0^t f(x(t; x_0\,, u), u(t))\, dt. \tag{12}$$

Hence, the value of $x^0(t)$ may be interpreted as the cost associated with the control u up to time t. The initial conditions for (11) are not specified.

Considering $\bar{\psi}$ and $\bar{x}$ fixed, the Hamiltonian is a function of $u \in \Omega$. Let $\mathfrak{M}(\bar{\psi}, \bar{x})$ denote the maximum of $\mathcal{H}$ over Ω:

$$\mathfrak{M}(\bar{\psi}, \bar{x}) = \max_{u \in \Omega} \mathcal{H}(\bar{\psi}, \bar{x}, u).$$

$\mathfrak{M}(\bar{\psi}, \bar{x})$ is well defined because Ω is compact and $\mathcal{H}$ is a continuous function of u.

Pontryagin's maximum principle as it applies to the problem is expressed as follows.

Theorem 2. [7] *If $u \in U$ is an optimal control, then there is a nontrivial solution $\bar{\psi}(\cdot)$ of* (11) *corresponding to u such that*
(a) $\mathcal{H}(\bar{\psi}(t), \bar{x}(t), u(t)) = \mathfrak{M}(\bar{\psi}(t), \bar{x}(t))$, a.e.,
(b) $\mathfrak{M}(\bar{\psi}(t), \bar{x}(t)) = 0$,
(c) *$\psi^0(t)$ is a nonpositive constant.*

We need to distinguish between the cases when $\psi^0 < 0$ and $\psi^0 = 0$. The problem is called normal when $\psi^0 < 0$, and abnormal when $\psi^0 = 0$. If the problem is normal, since $\mathcal{H}$ is linear and homogeneous in $\bar{\psi}$, we may and shall assume $\psi^0 = -1$.

We now proceed to strengthen Theorem 2 in two ways. First we show the problem is always normal; and second, the vector $\bar{\psi}(t)$ in the theorem must satisfy the boundary condition $\lim_{t\to\infty} \bar{\psi}(t) = (-1, 0, 0, \cdots, 0)$. With these additions, the maximum principle will then be shown to be a sufficient condition for optimality.

The proof of the next lemma depends on a construction used by Gamkrelidze in studying the time optimal control problem [8].

Lemma 3. *For every initial state, the problem is normal.*

Proof. Suppose u is an optimal control but $\psi^0(t) = 0$. In this case, we obtain from (9) and (11), the equations

$$\mathcal{H}(\bar{\psi}, \bar{x}, u) = \langle \psi, Ax + Bu \rangle, \qquad \dot{\psi} = -\psi A. \tag{13}$$

Then, since u maximizes the Hamiltonian, $u(t) = \operatorname{sgn}(\psi(t)\cdot B)$, $t \geqq 0$, where $\operatorname{sgn}: R^r \to R^r$ is the vector valued function whose ith component, $1 \leqq i \leqq r$, is

$$(\operatorname{sgn}(x^1, x^2, \cdots, x^r))_i = \begin{cases} 1 & \text{if } x^i > 0, \\ -1 & \text{if } x^i < 0, \\ \text{undefined} & \text{if } x^i = 0. \end{cases}$$

We shall obtain a contradiction by showing that

$$\int_0^\infty \langle u(t), Ru(t) \rangle \, dt = \infty. \tag{14}$$

First we observe that $\psi(t) \neq 0$, since $\psi^0 = 0$ and $\bar{\psi}(\cdot)$ is a nontrivial solution of (11). To establish (14), it suffices to show the set of points t for which $\psi(t)\cdot B = 0$ is a discrete set. Observe that the components of $\psi(t)\cdot B$ are analytic functions since $\psi(t) = \psi_0 \cdot e^{-At}$ for some nonzero $\psi_0 \in R^n$. Thus, if $\psi(t)\cdot B$ vanishes on a nondiscrete set, it vanishes identically, $\psi_0 \cdot e^{-At}B = 0$, $t \geqq 0$. Successively differentiating the last expression with respect to t and putting $t = 0$ yields

$$\psi_0 \cdot B = 0, \\ \psi_0 \cdot AB = 0, \\ \vdots \\ \psi_0 \cdot A^{n-1}B = 0. \tag{15}$$

But by hypothesis, (1) is controllable; so

$$\operatorname{rank}\,[B, AB, \cdots, A^{n-1}B] = n,$$

implying that (15) cannot hold, which proves the lemma.

In view of Lemma 3, we may put $\psi^0 = -1$. Then, if u maximizes the Hamiltonian, from (9) we have

$$u(t) = \operatorname{sat}\,(R^{-1}B^*\psi(t)), \qquad t \geqq 0, \tag{16}$$

where B^* is the transpose of B and sat: $R^r \to R^r$ is the function whose ith component, $1 \leqq i \leqq r$, is

$$(\operatorname{sat}\,(x^1, x^2, \cdots, x^r))_i = \begin{cases} 1 & \text{if } x^i \geqq 1, \\ x^i & \text{if } |x^i| < 1, \\ -1 & \text{if } x^i \leqq -1. \end{cases}$$

In addition, with $\psi^0 = -1$, (11) becomes

$$\dot{\psi}(t) = -A^*\psi(t) + Qx(t), \tag{17}$$

where A^* is the transpose of A.

LEMMA 4. *If $u \in U$ is an optimal control and $\psi(\cdot)$ is a solution of* (12) *such that $u(t) = \operatorname{sat}\,(R^{-1}B^*\psi(t))$, then $\lim_{t\to\infty} \psi(t) = 0$.*

Proof. We shall only indicate the idea of the proof because it is similar to that of Lemma 2. Since u is an optimal control $\lim_{t\to\infty} x(t; u) = 0$, and therefore for large t, the solution of (17) differs little from the solution of (13). With this observation, Lemma 4 can be proved by the same method used in proving Lemma 3.

The results of this section are summarized in the following.

THEOREM 2′. *If $u \in U$ is an optimal control, then there is a solution $\psi(\cdot)$ of* (17) *corresponding to u such that*
(a) $u(t) = \operatorname{sat}\,(R^{-1}B^*\psi(t))$,
(b) $\lim_{t\to\infty} \psi(t) = 0$.

4. Sufficient conditions for optimality. In this section, the conditions in Theorem 2′ will be shown to be sufficient conditions for a control to be optimal. Throughout the discussion, the initial state x_0 will be assumed given and fixed.

If $u_1, u_2 \in U$, let

$$d(u_1, u_2, t_1) = \int_0^{t_1} \sqrt{\langle u_1(t) - u_2(t), u_1(t) - u_2(t)\rangle}\, dt.$$

Similarly, if $u \in U$, let

$$J(u, t_1) = \int_0^{t_1} f(x(t; u), u(t))\, dt.$$

The following lemma is a particularization to our problem of a more general result due to Rozonoer [9].

LEMMA 5. *Let $u, u_1 \in U$. Suppose there is a solution $\bar{\psi}(\cdot)$ of (11) corresponding to u such that*

$$\mathcal{H}(\bar{\psi}(t), \bar{x}(t; u), u(t)) = \mathfrak{M}(\bar{\psi}(t), \bar{x}(t; u));$$

then

$$\langle \bar{\psi}(t_1), \bar{x}(t_1; u_1) - \bar{x}(t_1; u)\rangle + o(\epsilon) \leqq 0,$$

whenever $d(u, u_1, t_1) \leqq \epsilon$.

The strengthening of the hypothesis of Lemma 4 to conditions (a) and (b) of Theorem 2′ allows us to prove the next lemma.

LEMMA 5′. *Let $u, u_1 \in U$. Suppose there is a solution $\bar{\psi}(\cdot)$ of (17) corresponding to u such that*
(a) $u(t) = \text{sat}\,(R^{-1}B^*\psi(t))$,
(b) $\lim_{t\to\infty} \psi(t) = 0$.
Then for any $\delta > 0$ and $N > 0$, there exists a $t_1 > N$ such that

$$J(u, t_1) - J(u_1, t_1) \leqq \delta \,\| x(t_1; u_1) - x(t_1; u)\| + o(\epsilon),$$

whenever $d(u, u_1, t_1) \leqq \epsilon$.

Proof. Since $\psi^0 = -1$, we have, upon application of Lemma 5,

$$(18) \quad J(u, t_1) - J(u_1, t_1) + \langle \psi(t_1), x(t; u_1) - x(t; u)\rangle + o(\epsilon) \leqq 0,$$

whenever $d(u, u_1, t_1) \leqq \epsilon$. Given N and $\delta > 0$ and using condition (b), we can find $t_1 > N$ such that

$$|\langle \psi(t_1), x(t_1; u_1) - x(t; u)\rangle| \leqq \delta \,\| x(t_1; u_1) - x(t; u)\|.$$

The last expression together with (18) proves the lemma.

We are now in the position to prove the principal result of this section.

THEOREM 3. *Let $u \in U$ transfer x_0 to the origin. Then a necessary and sufficient condition for u to be optimal is that there exist a solution $\psi(\cdot)$ of (17) corresponding to u such that*
(a) $u(t) = \text{sat}\,(R^{-1}B^*\psi(t))$,
(b) $\lim_{t\to\infty} \psi(t) = 0$.

Proof. The necessity part is just a restatement of Theorem 2′.

To prove the sufficiency, suppose u_0 is the optimal control for the initial state x_0, but u is nonoptimal. Then for some $\alpha > 0$ and $N > 0$, and for all $t_1 > N$,

$$J(u, t_1) - J(u_0, t_1) \geqq \alpha. \tag{19}$$

For $0 < \lambda < 1$, consider the controls u_λ defined by

$$u_\lambda(t) = (1 - \lambda)u(t) + \lambda u_0(t), \qquad t \geqq 0. \tag{20}$$

We remark that the formula in Lemma 1 is valid if $J(u, \lambda x_0^1 + (1 - \lambda)x_0^2)$, $J(u_1, x_0^1)$, and $J(u_2, x_0^2)$ are replaced by $J(u_\lambda, t_1)$, $J(u_0, t_1)$, and $J(u, t_1)$ respectively; thus $J(u_\lambda, t_1) \leqq \lambda J(u_0, t_1) + (1 - \lambda)J(u, t_1)$. Subtracting $J(u, t_1)$ from both sides of this inequality and using (19) yields the inequality

$$J(u, t_1) - J(u_\lambda, t_1) \geqq \lambda\alpha. \tag{21}$$

Now we use Lemma 5′ to reach a contradiction. Both $x(t; u)$ and $x(t; u_0)$ are bounded because u and u_0 transfer x_0 to the origin. Application of (6) yields

$$x(t; u_\lambda) - x(t; u) = \lambda(x(t; u_0) - x(t; u)),$$

and therefore there is a constant β such that for all t_1,

$$\| x(t_1; u) - x(t; u_\lambda) \| \leqq \beta\lambda.$$

We note that for $t_1 < \infty$, $d(u, u_\lambda, t_1) = \lambda\gamma$, and $\gamma < \infty$. Finally, using Lemma 5′,

$$J(u, t_1) - J(u, t_1) \leqq \delta\beta\lambda + o(\gamma\lambda).$$

Since δ can be made arbitrarily small, the last expression contradicts (21). Therefore u must have been optimal.

5. Optimal feedback control. A function $v: R^n \rightarrow \Omega$ will be called an admissible feedback control if the autonomous differential equation,

$$\dot{x} = Ax + v(x), \tag{22}$$

has a unique solution for any initial state. Let $x(t; x_0)$ be the solution of (22) which satisfies the boundary condition $x(0; x_0) = x_0$. Then v is said to be an optimal feedback control if for every x_0, the time function

$$t \rightarrow v(x(t; x_0)), \qquad t \geqq 0,$$

is the optimal control for the initial state x_0. In this section we shall consider the problem of determining an optimal feedback control.

It is expedient to combine (1), (16), and (17) into the single equation

$$(23)\qquad \begin{pmatrix} \dot{x} \\ \dot{\psi} \end{pmatrix} = \begin{pmatrix} A & 0 \\ Q & -A^* \end{pmatrix} \begin{pmatrix} x \\ \psi \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} \text{sat } (R^{-1}B^*\psi),$$

where $(x, \psi) = (x^1, x^2, \cdots, x^n, \psi^1, \psi^2, \cdots, \psi^n)$.

In the sequel, $(x(\cdot\,; x_0), \psi(\cdot\,; \psi_0))$ will denote the solution of (23) which satisfies the boundary conditions $x(0; x_0) = x_0$ and $\psi(0; \psi_0) = \psi_0$.

The significance of (23) is this: given an initial state x_0 according to Theorem 3, the problem of determining the optimal control for x_0 is equivalent to finding a ψ_0 such that $x(t; x_0) \to 0$ and $\psi(t; \psi_0) \to 0$ as $t \to \infty$. Since ψ_0 and (16) define the control when the state is x_0, the problem of determining an optimal feedback control is one of choosing an appropriate ψ for each state x.

Unfortunately, because (23) is nonlinear, there is no simple relation between x and the proper ψ, in general. However, for all states sufficiently close to the origin, we shall show there is a linear relation between ψ and x of the form $\psi = Px$, where P is an $n \times n$ real matrix.

We observe that if the control region Ω is replaced by R^r, i.e., the magnitude constraints on the control are removed but otherwise the problem is the same, then everything done so far is valid if everywhere $R^{-1}B^*\psi$ is substituted for sat $(R^{-1}B^*\psi)$. In fact, the only places where we explicitly used the assumption that the control region was Ω are in (16) and Lemma 3. When the control region is R^r, (16) is valid with the above substitution, and the proof of Lemma 3 carries over essentially without change.

Assuming the control region is R^r, (23) becomes the linear equation

$$(24)\qquad \begin{pmatrix} \dot{x} \\ \dot{\psi} \end{pmatrix} = \begin{pmatrix} A & BR^{-1}B^* \\ Q & -A^* \end{pmatrix} \begin{pmatrix} x \\ \psi \end{pmatrix}.$$

For brevity, let

$$D = \begin{pmatrix} A & BR^{-1}B^* \\ Q & -A^* \end{pmatrix}.$$

Then taking the Laplace transform of (24),

$$(25)\qquad S\begin{pmatrix} \hat{x}(s) \\ \hat{\psi}(s) \end{pmatrix} - \begin{pmatrix} x(0) \\ \psi(0) \end{pmatrix} = D\begin{pmatrix} \hat{x}(s) \\ \hat{\psi}(s) \end{pmatrix},$$

where $\hat{x}(s)$ and $\hat{\psi}(s)$ are the Laplace transforms at $x(\cdot)$ and $\psi(\cdot)$, respectively. From (25),

$$(26)\qquad \begin{pmatrix} \hat{x}(s) \\ \hat{\psi}(s) \end{pmatrix} = (sI - D)^{-1} \begin{pmatrix} x(0) \\ \psi(0) \end{pmatrix}.$$

The right hand side of (26) is a column vector of rational functions in s. Given $x(0)$, in order to satisfy the desired boundary conditions at ∞, it is necessary to choose $\psi(0)$ so that the poles of each of these rational functions have negative real parts. Such a choice is possible because of Theorems 1 and 3. If we let $x(0)$ take on the values $e_j = (0, 0, \cdots, 0, 1, 0, \cdots, 0)$ with the 1 in the jth coordinate, $1 \leqq j \leqq n$, we get corresponding to each e_j, for an appropriate $\psi(0)$, a vector $P_j = (P_{1j}, P_{2j}, \cdots, P_{nj})$. Now from the linearity of (26), it is clear that if $x(0) = \sum_{j=1}^{n} \alpha_j e_j$, then an appropriate choice for $\psi(0)$ is $\sum_{j=1}^{n} \alpha_j P_j$. In matrix notation, this can be expressed

$$(27) \qquad \psi(0) = Px(0), \qquad P = (P_{ij}).$$

It follows that in the case when the control region is R^r, the optimal feedback control is, in view of (16) and (27), given by*

$$(28) \qquad v(x) = R^{-1}B^*Px.$$

Now consider the asymptotically stable system,

$$(29) \qquad \dot{x} = Ax + BR^{-1}B^*Px.$$

Let $G = \{x \in R^n \mid | (R^{-1}B^*Px)_i | \leqq 1, \quad i = 1, 2, \cdots, r\}$. G is a neighborhood of the origin, and because (29) is stable, we can find a neighborhood S of the origin such that if $x_0 \in S$, then the whole trajectory of (29) starting at x_0 remains within G. It is then clear that the optimal feedback control for states in S is the same regardless of whether the control region is Ω or R^r. We state this conclusion as follows.

PROPOSITION 3. *There is a neighborhood S of $x = 0$ such that if $x \in S$, the optimal feedback control is $v(x) = R^{-1}B^*Px$ for some*† *$n \times n$ matrix P.*

We have seen that if $x(0) \in S$, an appropriate choice for $\psi(0)$ in (23) to satisfy the desired boundary conditions at ∞ is given by (27). The next lemma shows this is the only appropriate choice.

LEMMA 6. *For each x_0, there is a unique ψ_0 such that the solution of* (23) *satisfies the boundary condition*

$$(30) \qquad \lim_{t \to \infty} (x(t; x_0), \psi(t; \psi_0)) = 0.$$

Proof. The existence of at least one such ψ_0 follows from Theorems 1 and 3. Suppose $(x_1(\cdot; x_0), \psi_1(\cdot; \psi_0^1))$ and $(x_1(\cdot; x_0), \psi_1(\cdot; \psi_0^2))$ are two solutions of (23) which satisfy the boundary conditions (30). We have to show $\psi_0^1 = \psi_0^2$. From Theorem 3, it follows that the controls defined by

* This result was previously obtained by Kalman [10], who also gave numerically efficient means of computing P.

† Later by Lemma 6, we show P is unique.

$$u_1(t) = \text{sat}\,(R^{-1}B^*\psi_1(t;\psi_0^1)),$$
$$u_2(t) = \text{sat}\,(R^{-1}B^*\psi_2(t;\psi_0^2)),$$

are both optimal for the initial state x_0. By Corollary 1, $u_1(t) = u_2(t)$, $0 \leqq t < \infty$, and therefore $x_1(t; x_0) = x_2(t; x_0)$ on the same interval. Then from (23),

$$\frac{d}{dt}(\psi_1 - \psi_2) = -A^*(\psi_1 - \psi_2).$$

Solving the last equation,

$$\psi_1(t;\psi_0^1) - \psi_2(t;\psi_0^2) = e^{-A^*t}(\psi_0^1 - \psi_0^2).$$

Since $u_1(t) = u_2(t)$ and both $\psi_1(t;\psi_0^1)$ and $\psi_2(t;\psi_0^2)$ converge to 0, for some $T < \infty$,

$$R^{-1}B^*\psi_1(t;\psi_0^1) = R^{-1}B^*\psi_2(t;\psi_0^2), \qquad t \geqq T.$$

Hence, after transposing, $(\psi_0^1 - \psi_0^2)e^{-At}B = 0$, $t \geqq T$. We showed in proving Lemma 3 that the last relation holds only if $\psi_0^1 = \psi_0^2$, which proves the lemma.

Lemma 6 implies that the matrix P in Proposition 3 is unique. Combining Lemma 6 and (27), we have the following.

Lemma 7. *There is a neighborhood S of $x = 0$ and a unique matrix P such that if $x_0 \in S$, the solution of* (23), $(x(\cdot\,; x_0), \psi(\cdot\,; \psi_0))$, *satisfies* (30) *if and only if $\psi_0 = Px_0$.*

We now are in a position to describe how the optimal feedback control may be computed. For states in S, it is given by Proposition 3, but for states not in S, we do not have a closed form solution. However, using Lemma 7 to determine the appropriate initial conditions, integrating (23) backwards in time with $x_0 \in S$, yields an optimal trajectory. We are assured that every optimal trajectory can, in principle, be determined by such an integration because of Theorem 3 and Lemma 7. By making the number of such trajectories sufficiently large, the feedback control can be determined on an arbitrarily dense set of states.

In practice, with the above method of determining $v(x)$, the value of $v(x)$ can only be determined on a proper subset of state space, and since no numerical computation can be perfect, the following result is of some importance.

Proposition 4. *The optimal feedback control is a continuous function of the state.*

Proof. By Lemma 6, there is assigned to each x_0 a unique ψ_0 such that if $(x(\cdot\,; x_0), \psi(\cdot\,; \psi_0))$ is the solution of (23), then $x(t; x_0) \to 0$ and $\psi(t; \psi_0) \to 0$ as $t \to \infty$. Let F be the function defined by this assignment: $F(x_0) = \psi_0$. Then, in view of (16), to prove the assertion, it suffices to show that F is a continuous function.

Let $W \subset S$ be a compact neighborhood of the origin. Given x_0 and assuming $\psi_0 = F(x_0)$, there is a $t_1 > 0$ such that $x(t_1 ; x_0) \in W^\circ$, the interior of W. We consider the map $G: W \rightarrow R^n$ defined by $G(y) = x(-t_1 ; y)$, where $x(-t_1 ; y)$ is the first component of $(x(-t_1 ; y), \psi(-t_1 ; Py))$. G is one-to-one because of Lemma 6. Moreover G is continuous because the solution of (23) depends continuously on its initial conditions. Since V is compact, G^{-1}, the inverse of G, is continuous. Observe that, by construction, $G^{-1}(x_0) = x(t_1 ; x_0)$. Hence since $x(t_1 ; x_0) \in W^\circ$, there is a neighborhood N of x_0 such that $G^{-1}(N) \subset W^\circ$. Then if $x \in N$,

$$F(x) = \psi(-t_1 ; F(G^{-1}(x))) = \psi(-t_1 ; P \cdot G^{-1}(x)).$$

Since $\psi(-t_1 ; y)$ depends continuously on y, the last expression proves F is continuous, completing the proof of the proposition.

6. Remarks. Most of the results in §§2, 3, 4 can be extended to more general cost functions than (3). The essential property of (3) for the arguments in these sections is the convexity of the integrand. However, the method for determining the optimal feedback control in §5 relies substantially on the quadratic form of the integrand.

Efficient methods for computing the matrix P of Proposition 3 were devised by Kalman [10]. Further research is needed to discover practical and efficient ways of computing or approximating the optimal feedback control.

REFERENCES

[1] J. P. LaSalle, *The time-optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. 5, Princeton, 1960, pp. 1–24.

[2] L. A. Zadeh and C. A. Desoer, *Linear System Theory: The State Space Approach*, McGraw-Hill, New York, 1963, Chap. 11.

[3] A. M. Letov, *Analytic controller design*, Automat. Remote Control, 21 (1960), pp. 389–393.

[4] ———, *The analytic design*, Ibid., 22 (1961), pp. 363–372.

[5] N. N. Krasovskii and A. M. Letov, *The theory of analytical design of controllers*, Ibid., 23 (1962), pp. 644–656.

[6] C. D. Johnson and W. M. Wonham, *On a problem of Letov in optimal control*, Proceedings, Joint Automatic Control Conference, Stanford, California, 1964, pp. 317–328.

[7] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience, John Wiley, New York, 1962, pp. 189–191.

[8] R. V. Gamkrelidze, *The theory of time-optimal processes in linear systems*, Izv. Akad. Nauk SSSR, Ser. Mat., 22 (1958), pp. 449–474.

[9] L. I. Rozonoer, *The L. S. Pontryagin maximum principle in the theory of optimal systems*, Automat. Remote Control, 20 (1959), pp. 1288–1302.

[10] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

A TRANSFORMATION APPROACH TO SINGULAR SUBARCS IN OPTIMAL TRAJECTORY AND CONTROL PROBLEMS*

HENRY J. KELLEY†

Abstract. Mayer variational problems in which the control variable appears linearly are considered and a canonical form sought for the system equations which is somewhat analogous to that adopted by Wonham and Johnson for linear constant coefficient systems with cost functional quadratic in the state variables. A means of synthesizing a transformation to the canonical form in terms of the mutually independent solutions of a first order linear homogeneous partial differential equation is described. It is then shown how the Legendre-Clebsch necessary condition applied in the transformed system of variables may be employed to obtain information on the singular extremals of the problem and the possible appearance of singular subarcs in the solution.

Two examples are employed for illustration, one a simple servomechanism problem and the other Goddard's problem of optimal thrust programming for a sounding rocket.

Introduction. Optimal control problems in which the control variable appears linearly yield to conventional treatment if the optimal control has a bang-bang character. Difficulties arise with the possibility of intervals during which the optimal control may be intermediate between the specified limits, such segments of the solution being *singular* subarcs, in classical variational terminology. The Green's Theorem technique of Miele [1] is a powerful tool for solution of such problems, applicable, however, only if the state space is of very limited dimension. There is presently available no general theory for determining singular arcs, deciding as to whether or not they are minimizing even locally, i.e., over a short time interval, or for determining their role as subarcs of a composite solution. In the special case of systems linear in the state variables, investigated by LaSalle [2], the appearance of singular subarcs corresponds to degeneracy in the sense of nonuniqueness of solution.

In the present paper we investigate the possibility of a canonical form

* Received by the editors February 10, 1963, and in revised form September 19, 1964.

† Analytical Mechanics Associates, Incorporated, 941 Front Street, Uniondale, New York. The investigation reported in this paper was performed under Contract NAS 8-1549 with Marshall Space Flight Center of the National Aeronautics and Space Administration and Contract AF 29(600)-2671 with the United States Air Force Office of Scientific Research while the writer was employed in the Research Department of the Grumman Aircraft Engineering Corporation.

for such problems which resembles that chosen by Wonham and Johnson [3] for study of problems featuring linear constant coefficient systems and cost functional an integral quadratic form in the state variables. Initially our concern will be with synthesis of the desired form by means of an appropriate transformation. Following this, attention will be turned to the application of optimality criteria in the transformed system of variables.

It has been called to the writer's attention that the transformation scheme presented herein is similar to a scheme developed by Faulkner for treatment of the case in which the differential equations of state are reducible to a single total differential equation [4, 5, 6]. The two schemes appear to be equivalent for problems of that type, which are of fairly frequent occurrence in application.

Transformation to canonical form. Our analysis begins with the usual Mayer problem statement. A minimum is sought of a function P of the terminal values of variables $x_1, \cdots, x_n$ and the terminal value of the independent variable, time t. The variables $x_1, \cdots, x_n$ are *state* variables satisfying a system of first order differential equations of the form

$$\dot{x}_i = p_i(x_1, \cdots, x_n, t) + q_i(x_1, \cdots, x_n, t)y, \quad i = 1, \cdots, n. \tag{1}$$

In the class of problems of present interest, the differential equations are linear in a single *control* variable y, as indicated. Initial conditions numbering at most $n + 1$ and terminal conditions numbering at most n may be imposed upon the variables $x_1, \cdots, x_n$ and t. The variable y is subject to an inequality constraint of the form

$$y_1 \leqq y \leqq y_2 . \tag{2}$$

We wish to consider the possibility of introducing new variables

$$z_j = f_j(x_1, \cdots, x_n, t), \qquad j = 1, \cdots, m, \tag{3}$$

satisfying equations of state whose right members are not dependent upon the variable y explicitly:

$$\dot{z}_j = \sum_{i=1}^{n} \frac{\partial f_j}{\partial x_i} p_i + \frac{\partial f_j}{\partial t}, \qquad j = 1, \cdots, m. \tag{4}$$

Evidently $m < n$ unless all the q_i are identically zero. The vanishing of the collected coefficient of the variable y,

$$\sum_{i=1}^{n} \frac{\partial f_j}{\partial x_i} q_i = 0, \qquad j = 1, \cdots, m, \tag{5}$$

has been assumed in (4).

For the purpose of determining functions f_j having the desired property, we seek the solutions of the linear homogeneous first order partial differential equation (5). From the theory of characteristics [7] we are led to consideration of the ordinary differential equations

$$\frac{dx_i}{ds} = q_i(x_1, \cdots, x_n, t), \qquad i = 1, \cdots, n, \tag{6}$$

in which s is a parameter and t is fixed. If one of the quantities x_i is adopted instead of s as independent variable, the general solution may be represented in terms of $n - 1$ parameters:

$$C_k = \varphi_k(x_1, \cdots, x_n, t), \quad k = 1, \cdots, n - 1. \tag{7}$$

The C_k are constants of integration and the φ_k are *mutually independent* integrals of the system. Each integral φ_k is a solution of the partial differential equation (5).

The first $n - 1$ of the new variables z_j are then to be defined according to $f_j = \varphi_j$. The nth variable z_n we define as

$$z_n = x_l, \tag{8}$$

choosing l such that $q_l \neq 0$ over the domain of interest, a choice which we assume for the time being open to us. The mutual independence of the functions φ_j together with $q_l \neq 0$ insures that the transformation between the variables z and x is nonsingular by the nonvanishing of the Jacobian determinant

$$\Delta = \frac{\partial(z_1, \cdots, z_n)}{\partial(x_1, \cdots, x_n)} \neq 0. \tag{9}$$

The Legendre-Clebsch condition in the transformed variables. To provide intuitive motivation for our next step, we digress momentarily, considering the possibilities offered by our transformation in (rarely occurring) problems devoid of inequality constraints on the control variable. In such cases we are led to an equivalent problem in a state space of smaller dimension, the z_j, $j = 1, \cdots, n - 1$, becoming the state variables and $z_n = x_l$ the control variable. This comes about through the identical vanishing of the Lagrange multiplier associated with the nth equation of state

$$\dot{z}_n = p_l + q_l y. \tag{10}$$

In this equation as well as in the first $n - 1$ equations of state (4), the variables x_i are presumed eliminated in favor of the z_j by use of the inverse transformation. It should be noted that jump discontinuities in the new control variable $z_n(t) = x_l(t)$ occurring at corner points of the solution imply impulsive behavior of $y(t)$. Such behavior would be admissible in the absence of an inequality constraint on y, which we have momentarily assumed, and the Weierstrass necessary condition would then be directly applicable.

Unless the transformed equations were linear in the new control variable $z_n = x_l$, the Weierstrass necessary condition could then be employed in conjunction with the Euler equations for the transformed problem to yield information not obtainable via the corresponding condition in the original problem. The extremals of the transformed problem are the singular extremals of the original, and those satisfying the strengthened version of the Weierstrass condition are minimizing, at least over short intervals. In the special case in which the transformed equations of state (4) are linear in the new control variable x_l, an additional transformation to a state space of still smaller dimension is indicated.

Redirecting attention to the problem of main interest, in which the inequality constraint (2) is operative, we perceive that the course of action just described is not open to us. We may, however, examine sub-arcs over which the control variable y takes on values intermediate between the specified bounds

$$y_1 < y < y_2, \tag{11}$$

with similar considerations in mind. If $y = \bar{y}(t)$ is the optimal control, we must, evidently, restrict attention to small variations $\delta y(t) = \epsilon\eta(t)$, where $\eta(t)$ is an arbitrary piecewise continuous function and the magnitude of the variation, ϵ, is vanishingly small so that $y = \bar{y} + \delta y$ satisfies (11). In the literature of classical variational theory, such variations are often referred to as *weak* variations, and the Legendre-Clebsch condition, necessary for a *weak* relative minimum, plays a role loosely analogous to that of the Weierstrass condition whenever a restriction to vanishingly small variations is either assumed or imposed.

We rewrite (4) with the notation a_j for the functions appearing on the right as

$$\dot{z}_j = a_j(x_1, \cdots, x_n, t), \qquad j = 1, \cdots, n-1, \tag{12}$$

and with the variables x_i eliminated in favor of z_j, as

$$\dot{z}_j = b_j(z_1, \cdots, z_n, t), \qquad j = 1, \cdots, n-1. \tag{13}$$

Introducing the usual Lagrange multipliers λ_j, $j = 1, \cdots, n-1$, we form the Hamiltonian

$$H \equiv \sum_{j=1}^{n-1} \lambda_j b_j, \tag{14}$$

and write the Euler-Lagrange equations corresponding to the z_j,

$$\dot{\lambda}_j = -\frac{\partial H}{\partial z_j}, \qquad j = 1, \cdots, n-1, \tag{15}$$

and that corresponding to z_n,

$$\frac{\partial H}{\partial z_n} = 0. \tag{16}$$

The Legendre-Clebsch necessary condition is

$$\frac{\partial^2 H}{\partial {z_n}^2} \delta {z_n}^2 \geqq 0, \tag{17}$$

for $\delta z_n \neq 0$, or

$$\frac{\partial^2 H}{\partial {z_n}^2} \geqq 0. \tag{18}$$

Solutions of the system (13), (15) and (16) are the extremals of the transformed problem and the condition (18) provides an additional criterion for screening these candidates. If the left member of (18) is positive, the singular subarc is locally minimizing, i.e., over short time intervals; if negative, maximizing. The vanishing of the left member of (18) corresponds to the special case, mentioned earlier, in which z_n enters the function H linearly. Thus along singular arcs of the original problem, (18) partially fills the gap created by the Weierstrass necessary condition's being trivially satisfied.

If it is not possible to choose the variable z_n according to the scheme $z_n = x_l$, $q_l \neq 0$, or if it is inconvenient to invert the transformation $Z(X)$ analytically, one may deal with the equations of state in the form (12) adjoining the $n - 1$ equations (3) as constraints by means of additional Lagrange multipliers. The more complex form of the Legendre-Clebsch necessary condition as given in [8], for example, must then be applied.

Examples. 1. *A servomechanism problem.* In [1] and [9] the following problem has been studied in some detail. Given the system

$$\dot{x}_1 = x_2 + y, \tag{19}$$

$$\dot{x}_2 = -y, \tag{20}$$

$$\dot{x}_3 = \frac{{x_1}^2}{2}, \tag{21}$$

$$| y | \leqq 1, \tag{22}$$

the control taking the system from a specified initial state to $x_1 = x_2 = 0$ and extremizing the final value of x_3 is sought. The structure of the solution of this problem is rather complex, belying its innocuous appearance.

An application of the transformation scheme just described leads to $z_1 = x_1 + x_2$, $z_2 = x_1^2/2$, and $z_3 = x_1$. An examination of the Legendre-Clebsch condition leads to the conclusions:

(a) The singular subarcs of [1] and [9] are locally minimizing for the case of a minimum of the final value of x_3.

(b) The singular subarcs are *not* minimizing if the function whose minimum is sought is the *negative* of the final value of x_3, and the optimal control is bang-bang.

In [1] a result stronger than (a) is obtained, and for problems which fit the linear/quadratic format of [1] this will generally be the case. Owing to an assumed restriction in the problem statement of [1], the results do not apply to case (b). It should be noted that the family of solutions contain not only the singular arcs found in [9] for unlimited final time, but also other singular arcs for finite final time, as pointed out to the writer by A. E. Bryson of Harvard University in personal communication.

2. *Goddard's problem.* The problem of determining the optimal thrust program for the vertical flight of a sounding rocket is one which has been extensively studied in the astronautical literature. The state variables are altitude h, velocity V, and mass m, satisfying

$$\dot{h} = V,$$

$$\dot{V} = \frac{T - D(h, V)}{m} - g(h),$$

$$\dot{m} = -\frac{T}{c},$$

in which rocket thrust T is bounded above and below according to

$$0 \leqq T \leqq \bar{T}.$$

The function D is aerodynamic drag, g is the acceleration of gravity, and c is rocket exhaust velocity. The problem usually of interest is the minimization of propellant expenditure $m_0 - m_f$ with fixed initial mass for attainment of fixed final altitude, final velocity and time unspecified.

The transformation scheme leads to $z_1 = h$, $z_2 = me^{V/c}$ as new state variables and $z_3 = V$ as new control variable. The problem is nonsingular in the state space of reduced dimension. The version of the problem featuring drag proportional to the square of the velocity has been investigated fairly thoroughly and in this case a single intermediate thrust subarc enters the solution, which the Legendre-Clebsch condition confirms as locally minimizing. The advantage of employing such variables in the sounding rocket problem was first recognized by Faulkner [5] and later, independently, by Ross [10]. In the case of a general drag law, e.g., one which exhibits sharp variation in the vicinity of sonic velocity, the Legendre-Clebsch condition may rule out intermediate thrust operation over a certain velocity range.

Concluding remarks. The transformation scheme and application of the Legendre-Clebsch condition appear to be useful for examination of singular subarcs in Mayer problems linear in a single control variable, although, of course, the information obtained is only a fragment of that needed for complete analysis of such problems. Perhaps the most interesting and suggestive feature of the approach is the idea of treating problems of this kind in a state space of reduced dimension. A different but somewhat related approach to the testing of singular arcs is presented in [11].

REFERENCES

[1] A. Miele, *Extremization of linear integrals by Green's theorem,* Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, chap. 3.

[2] J. P. LaSalle, *The time optimal control problem,* Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960.

[3] W. M. Wonham and C. D. Johnson, *Optimal bang-bang control with quadratic performance index,* Joint Automatic Control Conference, Minneapolis, Minnesota, June 19–21, 1963.

[4] F. D. Faulkner, *A degenerate problem of Bolza,* Proc. Amer. Math. Soc., 6 (1955), pp. 847–854.

[5] ———, *The problem of Goddard and optimum thrust programming,* Advances in the Astronautical Sciences, vol. I, Plenum Press, New York, 1957.

[6] ———, *Direct methods,* Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, chap. 2.

[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics, vol. II,* Wiley, New York, 1962.

[8] G. A. Bliss, *Lectures on the Calculus of Variations,* University of Chicago Press, Chicago, 1946.

[9] C. D. Johnson and J. E. Gibson, *Singular solutions in problems of optimal control,* IEEE Trans. Automatic Control, 8 (1963), pp. 4–15.

[10] S. Ross, *Minimality for problems in vertical and horizontal rocket flight,* Jet Propulsion, 28 (1958), pp. 55–56.

[11] H. J. Kelley, *A second variation test for singular extremals,* AIAA J., 2 (1964), pp. 1380–1382.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

CONTROLLABILITY AND THE SINGULAR PROBLEM*

H. HERMES†

Introduction. The concept of complete controllability of linear systems was introduced by R. E. Kalman [1]. It is part of the purpose of this paper to extend the concept to nonlinear systems, with control appearing linearly. All systems considered are of this form.

Geometrically, a linear system is completely controllable at time t_0 if any state can be attained in finite time by a trajectory of the system having arbitrary initial data x_0 at time t_0 . The motivation for the extension of this concept to nonlinear systems came largely from results obtained in [2] and from the geometric interpretation of nonintegrability of Pfaffians given in [3] and [4]. In particular, Carathéodory gives an argument to show that if, for a single Pfaffian equation, there are points in every neighborhood of a given point which are not "reachable" from the given point by curves satisfying the equation, the equation is integrable. This result was generalized to systems of Pfaffians in [4]. There is a difficulty in applying these ideas to Pfaffian systems which are quite naturally associated with control systems having control appearing linearly. (See §2.) The reason for this is that usually the independent variable t appears explicitly in the Pfaffian system, hence its integral curves (which can be related back to solutions of the control system, and are used to connect neighboring points to a given point) must have t parametrized as $t(\sigma)$, a monotone function of σ. This is *not* the case in the proofs in [3] and [4], and with this restriction, in general the results of these papers are no longer valid.

The relation between singular problems and controllability arises quite naturally from the Pfaffian approach and can be anticipated from results obtained by LaSalle in [5]. In §2 we define the concept of a totally singular arc, i.e., an arc satisfying the differential constraining equations, for which there exists an adjoint vector such that the maximum principle yields no information as to the optimality of *any* of the components of the control along this arc. In particular, if the system were linear and admitted no totally singular arc, the system would be proper in the sense of LaSalle [5] and completely controllable in the sense of Kalman [6]. Even if the controls are merely restricted to be $\mathfrak{L}_2$ (Lebesgue square integrable) functions, it is

* Received by the editors November 12, 1963, and in revised form September 21, 1964.

† The Martin Company, Denver, Colorado. Now at Division of Applied Mathematics, Brown University, Providence, Rhode Island. This work was supported in part by the National Aeronautics and Space Administration under Contract NAS 8-11020.

shown that totally singular arcs can exist and comprise some or all of the boundary of the attainable set, thereby being optimal trajectories for certain time optimal control problems. These are also precisely the arcs along which the system need not be locally controllable, i.e., if we assume initial data x_0 given at time t_0, there *may* exist points in every state space neighborhood of a point $\varphi^v(t_1)$ of a totally singular arc φ^v, which are not attainable in time $t_1 > t_0$ by trajectories of the system with $\mathfrak{L}_2$ controls. Here φ^v denotes the solution of the system with control v. Precisely, if for every $t_1 > t_0$ there exist points in every state space neighborhood of $\varphi^v(t_1)$, which are not attainable with $\mathfrak{L}_2$ control in time t_1, the arc φ^v is totally singular. However, it is shown by example that there do exist totally singular arcs about which the system is locally controllable.

1. Complete controllability for linear and mildly nonlinear systems. Throughout this section H will denote an $n \times r$ matrix valued function of t, which is in $\mathfrak{L}_2[t_0, t_1]$ for any given finite $t_1 > t_0$. Controls will be $\mathfrak{L}_2$ vector-valued functions. We begin with the following basic lemma.

LEMMA 1.1. *A necessary and sufficient condition that there exist an $r \times n$ matrix valued function $V(t)$ in $\mathfrak{L}_2[t_0, t_1]$ such that for some $t_1 > t_0$, $\int_{t_0}^{t_1} H(\tau)V(\tau)d\tau$ is nonsingular is that for some $t_1 > t_0$, $\int_{t_0}^{t_1} H(\tau)H^T(\tau)d\tau$ is nonsingular.*

Proof. Sufficiency is immediate by choosing $V(\tau) = H^T(\tau)$. To show necessity assume there exist V, $t_1 > t_0$, such that $\int_{t_0}^{t_1} H(\tau)V(\tau)d\tau$ is nonsingular, but $\int_{t_0}^{\bar{t}} H(\tau)H^T(\tau)d\tau$ is singular for all $\bar{t} > t_0$, in particular $\bar{t} = t_1$. This implies there exists a constant vector $c \neq 0$ such that $c\left(\int_{t_0}^{t_1} H(\tau)H^T(\tau)\, d\tau\right)c^T = 0$, and since $H(\tau)H^T(\tau)$ is positive semidefinite, we obtain $cH(t) = 0$ almost everywhere in $[t_0, t_1]$. Thus $\int_{t_0}^{t_1} cH(\tau)V(\tau)d\tau = 0$, which contradicts the nonsingularity of $\int_{t_0}^{t_1} H(\tau) \cdot V(\tau)d\tau$.

We next consider the system

$$\dot{x}(t) = H(t)u(t), x(t_0) = x_0, u \in \mathfrak{L}_2[t_0, t_1]. \tag{1.1}$$

Define

$$M(t_0, t_1) \equiv \int_{t_0}^{t_1} H(\tau)H^T(\tau)\, d\tau.$$

THEOREM 1.1. *A necessary and sufficient condition for the system* (1.1) *to be completely controllable at t_0 is that there exists $t_1 > t_0$ such that $M(t_0, t_1)$ is nonsingular.*

Proof. (Sufficiency.) Let $\bar{x}$ be any given point in E^n, Euclidean n-space. We will show $\bar{x}$ is attainable from x_0 at time t_1. Indeed pick $u(t) = H^T(t)\,\xi$, $\xi \in E^n$. We desire $\bar{x} = x(t_1) = x(t_0) + \left(\int_{t_0}^{t_1} H(\tau)H^T(\tau)\,d\tau\right)\xi$ or $\xi = M^{-1}(t_0, t_1)\,(\bar{x} - x(t_0))$.

(Necessity.) Assume $M(t_0, t_1)$ is singular for all $t_1 > t_0$. This implies (see proof of Lemma 1.1) that there exists a constant vector $c \neq 0$ such that $cH(t) \equiv 0$ a.e. Since x_0 is arbitrary, let it be such that $c \cdot x_0 = 0$. We will show the point c is not attainable from x_0. Indeed suppose for some u and t_1, $c = x_0 + \int_{t_0}^{t_1} H(\tau)u(\tau)\,d\tau$. Then

$$c \cdot c = \|c\|^2 = c \cdot x_0 + c \int_{t_0}^{t_1} H(\tau)u(\tau)\,d\tau = 0,$$

a contradiction to the fact that $c \neq 0$.

COROLLARY 1.1 (*Kalman*). *The linear system*

$$\dot{x}(t) = A(t)x(t) + H(t)u(t),\ x(t_0) = x_0, \tag{1.2}$$

is completely controllable at t_0 and and only if

$$\int_{t_0}^{t_1} \Phi(t_0, \tau)H(\tau)H^T(\tau)\Phi^T(t_0, \tau)\,d\tau$$

is nonsingular for some $t_1 > t_0$. Here $\Phi(t, \tau)$ denotes a fundamental solution of the homogeneous system $\dot{x}(t) = A(t)x(t)$.

Proof. Make the transformation $y(t) = \Phi^{-1}(t, t_0)x(t)$. Then x satisfies (1.2) if and only if y satisfies

$$\dot{y}(t) = \Phi(t_0, t)H(t)u(t),\ y(t_0) = x_0. \tag{1.3}$$

(Note $\Phi(t_0, t) = \Phi^{-1}(t, t_0)$.) From the transformation, it follows that the system (1.2) is completely controllable if and only if the system (1.3) is completely controllable, i.e., from Theorem 1.1, that there exists a $t_1 > t_0$ such that

$$\int_{t_0}^{t_1} \Phi(t_0, \tau)H(\tau)H^T(\tau)\Phi^T(t_0, \tau)\,d\tau$$

is nonsingular.

Some special results for nonlinear systems. We next consider the nonlinear system

$$\dot{x}(t) = g(t, x(t)) + H(t)u(t),\ x(t_0) = x_0, \tag{1.4}$$

with the assumptions:
i) $| g^j(t, x) | \leqq N, j = 1, 2, \cdots, n$.
ii) $| g^j(t, x) - g^j(t, \bar{x}) | \leqq m \, \|x - \bar{x}\|, j = 1, 2, \cdots, n$.
iii) g is continuous as a function of t for each x.

Again let $$M(t_0, t_1) = \int_{t_0}^{t_1} H(\tau)H^T(\tau) \, d\tau.$$

THEOREM 1.2. *A sufficient condition that the set of points attainable by trajectories of the system* (1.4) *with* $\mathfrak{L}_2$ *control be all of* E^n *is that* $M(t_0, t_1)$ *be nonsingular for some* $t_1 > t_0$.

Remark. Rather than state the theorem in this manner, one might consider merely saying that the system (1.4) is completely controllable at t_0. However, this notion has not been defined for nonlinear systems, and it does not seem reasonable to this author to define it in such a global fashion for these systems.

Proof. For arbitrary u, (1.4) has a solution designated φ^u which satisfies

$$(1.5) \qquad \varphi^u(t) \equiv x_0 + \int_{t_0}^{t} g(\tau, \varphi^u(\tau)) \, d\tau + \int_{t_0}^{t} H(\tau)u(\tau) \, d\tau.$$

Let $\bar{x}$ be any given point in E^n. We desire a control such that for some finite point $t_1 > t_0$, $\varphi^u(t_1) = \bar{x}$. It suffices to consider controls which come from a finite dimensional subspace of $\mathfrak{L}_2$, in particular the controls considered will be of the form $u(t) = H^T(t)\xi$ where $\xi \in E^n$. Hence the notation φ^ξ rather than φ^u will be used.

Define a mapping $\mathfrak{F}: E^n \to E^n$ as follows. Let $\alpha(\xi) \equiv \int_{t_0}^{t_1} g(\tau, \varphi^\xi(\tau)) \, d\tau$, and define $\mathfrak{F}(\xi) \equiv M^{-1}(t_0, t_1)[\bar{x} - \alpha(\xi) - x_0)]$. From (1.5) it follows that a fixed point of $\mathfrak{F}$ will yield a value ξ such that $\varphi^\xi(t_1) = \bar{x}$.

It is well known that with the conditions imposed on g [7, Theorem 7.4, Chap. I], φ^ξ is a continuous function of ξ in the topology $C[t_0, t_1]$, i.e., the topology induced by the supremum norm. Thus $\alpha(\xi)$ is a continuous function of ξ, and $\mathfrak{F}$ is a continuous function of ξ.

We next show that there exists a K such that $\| \xi \| \leqq K$ implies $\| \mathfrak{F}(\xi) \| \leqq K$. Letting $\| \xi \| = \sum_{i=1}^{n} | \xi_i |$ and $\| M^{-1} \|$ be any matrix norm, since $| g^j | \leqq N$, for any ξ, $\| \alpha(\xi) \| \leqq n(t_1 - t_0)N$. Letting

$$K \equiv \| M^{-1}(t_0, t_1) \| \, [\| \bar{x} \| + nN(t_1 - t_0) + \| x_0 \|],$$

it follows that for any ξ, $\| \mathfrak{F}(\xi) \| \leqq K$; hence, in particular, $\mathfrak{F}$ maps the ball $\{\xi \in E^n: \| \xi \| \leqq K\}$ continuously into itself. Thus $\mathfrak{F}$ has a fixed point.

Remark. The result obtained in this theorem is not surprising in view of Theorem 1.1 and the boundedness condition on the vector g, which excludes linear systems. Also the condition $M(t_0, t_1)$ nonsingular for some $t_1 > t_0$ is much stronger than it need be. For example, if we consider a linear system of the form (1.2) and $H(t)$ is a column vector with one component zero, then $M(t_0, t_1)$ is singular for all $t_1 \geqq t_0$, yet the system can certainly be completely controllable.

2. Nonlinear systems with linear control; the singular problem. In this section, we consider extending the notion of complete controllability to systems of the form

$$\dot{x}(t) = g(t, x(t)) + H(t, x(t))u(t), \tag{2.1}$$

where g is an n-vector, H an $n \times r$ matrix, while u is an $\mathfrak{L}_2$ control vector. It is assumed that g and H are C^1 in all arguments. Throughout, the stipulation $1 \leqq r < n$ is required to hold.

Let $B(t, x)$ be a C^1, $(n - r) \times n$ matrix with rank (n − rank H) at each point (t, x) in some domain $\mathfrak{D}$ of interest, such that

$$B(t, x)H(t, x) \equiv 0, \qquad (t, x) \in \mathfrak{D}. \tag{2.2}$$

Since $r < n$, we know that rank $B \geqq 1$ for all (t, x).

With the system (2.1), associate the Pfaffian system

$$B(t, x)\, dx - B(t, x)g(t, x)\, dt = 0. \tag{2.3}$$

Let b be an arbitrary linear combination of the rows b^ν of B, taken with C^1 scalar valued coefficients $\alpha_\nu(t, x)$, i.e.,

$$b(t, x) = \sum_\nu \alpha_\nu(t, x) b^\nu(t, x).$$

Throughout, b will be used to denote such a linear combination which is *not* identically zero.

DEFINITION 2.1. *The Pfaffian system* (2.3) *is integrable at the point* $(\bar{t}, \bar{x})$ *if there exists a* C^1 *scalar valued function* $\psi(t, x)$ *and an* $\epsilon > 0$ *such that for some* b,

$$\psi_x(t, x) = b(t, x), \qquad \psi_t(t, x) = -b(t, x)\cdot g(t, x)$$

for $\bar{t} \leqq t < \bar{t} + \epsilon$, $|x - \bar{x}| < \epsilon$. Here $\psi_x(t, x)$ denotes the vector with components $\partial\psi(t, x)/\partial x_i$. This notation will be used, when convenient.

Essentially this states that for some b,

$$b(t, x)\, dx - b(t, x)\cdot g(t, x)\, dt \tag{2.4}$$

is an exact differential in a "neighborhood" of $(\bar{t}, \bar{x})$. It should be noted that any integrating factor can be included in the coefficients of the linear combination of the rows b^ν.

The notion of integrability of a Pfaffian system is, of course, related to the property of completeness of an associated system of partial differential equations. To show the relation, let $C(x)$, $x \in E^n$, be a smooth $(n - r) \times n$ matrix, and $K(x)$ a smooth $n \times r$ matrix, both of maximum rank, such that $C(x)K(x) \equiv 0$. With the Pfaffian system

$$C(x)\,dx = 0 \tag{2.5}$$

we associate the system of partial differential equations $K^T(x)\ \partial f(x)/\partial x = 0$. Each row k^i of K^T can be considered as defining a vector field X^i which locally generates a one parameter semigroup of diffeomorphisms $\{T_i(t)\}$ (see, for example, [8, p. 10]). In turn, such a semigroup determines a vector field. If for each $i, j = 1, 2, \cdots, r$ and for all arbitrarily small fixed τ, the vector field determined by $\{T_j(\tau)T_i(t)T_j(-\tau)\}$ is linearly dependent on the fields X^i, the system of partial differential equations is said to be complete. If it is not complete, the number m of linearly independent fields formed in this manner is called the index of both the Pfaffian system and the associated partial differential equation system [4].

From the results in [4], it easily follows that *the Pfaffian system* (2.5) *is integrable* (Definition 2.1) *if and only if the index m is such that $m + r < n$.* If the index m is such that $m + r = n$, Chow [4] shows that there is a neighborhood of a point $x_0 \in E^n$ such that all points in this neighborhood are attainable by curves satisfying (2.5). From the viewpoint of local controllability for a control system, we can interpret this as follows. *If the Pfaffian system associated with the control system*

$$\dot{x}(t) = K(x(t))u(t), \qquad x(t_0) = x_0$$

has index m, where K is a continuous $n \times r$ matrix function of $x \in E^n$ with constant rank r, and $m + r = n$, then every point in some neighborhood of x_0 is attainable by an admissible trajectory. Indeed, since all points in some neighborhood of x_0 are attainable by absolutely continuous curves satisfying $C(x(t))\dot{x}(t) = 0$ almost everywhere, we must only show that such a curve also satisfies the differential equation. But $C(x(t))\dot{x}(t) = 0$ implies $\dot{x}(t)$ is a linear combination of the columns of $K(x(t))$, since $CK \equiv 0$. Thus there exists $u(t)$ such that $\dot{x}(t) = K(x(t))u(t)$ for almost all t. Since K has rank r, it has a continuous left inverse on its range, from which it follows that u is measurable.

Before stating an explicit criterion for complete controllability of a system of the form (2.1), one may ask: "What should one expect the definition to yield?" This can presently be answered as follows. Since the definition should extend that given for a linear system of the form (1.2) which is a special case of (2.1), one expects:

(a) If $g(t, x) \equiv A(t)x$, $H(t, x) \equiv H(t)$, then the criterion which defines complete controllability at t_0 for (2.1) should be equivalent with the condition $\int_{t_0}^{t_1} \Phi(t_0, t)H(t)H^T(t)\Phi^T(t_0, t)\,dt$ nonsingular for some $t_1 > t_0$, as given in Corollary 1.1.

(b) There should be a geometric interpretation of the condition; e.g., what points are attainable from the initial point in finite time? In the linear system there were global attainability results, i.e., any point could be attained from the initial point via a trajectory of the system. In the nonlinear problem, one would expect at most local results of this nature.

The approach will be to state a criterion for complete controllability of (2.1) which we will show satisfies (a). We then use this criterion to try to establish a geometric interpretation as mentioned in (b). Of course, how the definition of complete controllability should be extended is somewhat a matter of personal opinion.

Definition 2.2. *The system* (2.1) *is completely controllable at* $(\bar{t}, \bar{x}) \in \mathfrak{D}$ *if the associated Pfaffian system* (2.3) *is not integrable at* $(\bar{t}, \bar{x})$.

It will next be shown that this criterion is equivalent to the condition given in Corollary 1.1 for the special case of the linear system (1.2). In this case it suffices to take $B = B(t)$ in forming the Pfaffian system equivalent (2.3). Also, in taking the linear combination of the rows of B to form the single Pfaffian as in (2.4), we can consider the scalar functions α_ν as functions of only t. Indeed we must only show that if the Pfaffian form

$$b(t)\,dx - b(t)A(t)x\,dt \tag{2.6}$$

has an integrating factor, then this integrating factor, denoted by μ, can be taken as a function of only t. To obtain this, suppose $\bar{\mu}(t, x)$ is such that $\bar{\mu}(t, x)b(t)\,dx - \bar{\mu}(t, x)b(t)A(t)x\,dt$ is an exact differential. Then $\bar{\mu}_{x_j}b^i - \bar{u}_{x_i}b^j = 0$ for all $i, j = 1, 2, \cdots, n$, and $\bar{\mu}_t b + \bar{\mu}\dot{b} = -\bar{\mu}_x bAx - \bar{\mu}bA$. Define $\mu(t) = \bar{\mu}(t, 0)$, noting that for the linear system, $\mathfrak{D} = (t_0, \infty) \times E^n$ which implies $(t, 0) \in \mathfrak{D}$ for $t > t_0$. It follows that $\mu(t)$ is also an integrating factor.

Since it is sufficient to consider both μ and the α_ν as functions of only t, there is no loss of generality in considering that if the Pfaffian system

$$B(t)\,dx - B(t)A(t)x\,dt = 0 \tag{2.7}$$

associated with (1.2) is integrable, then (2.6) is an exact differential.

Since x appears linearly, Definition 2.1 simplifies for such systems, and is: *The Pfaffian system* (2.7) *is integrable at the point* $\bar{t}$ *if there exists a* C^1 *scalar valued function* $\psi(t, x)$ *and an* $\epsilon > 0$ *such that for some* b,

$$\psi_x(t, x) = b(t), \qquad \psi_t(t, x) = -b(t)A(t)x$$

for $\bar{t} \leqq t < \bar{t} + \epsilon$. (Note: Under the assumptions on B and H, ψ_{xt} and ψ_{tx} exist and are equal.)

Define:

$$W(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_0, t)H(t)H^T(t)\Phi^T(t_0, t)\,dt.$$

Then Corollary 1.1 states that the system (1.2) is completely controllable at t_0 if and only if there exists $t_1 > t_0$ such that $W(t_0, t_1)$ is nonsingular.

Remark 1. If A and H are constant matrices, Kalman [1] shows that this condition is equivalent to the condition: rank $[H, AH, \cdots, A^{n-1}H] = n$.

Remark 2. While the above condition given for the constant coefficient case can be directly checked, $W(t_0, t_1)$ depends on knowledge of a fundamental solution $\Phi(t, t_0)$ which is *not* always easily obtainable.

Remark 3. It is easily seen that $W(t_0, t_1)$ is a positive semidefinite matrix. Thus if $W(t_0, t_1)$ is nonsingular, $W(t_0, t)$ is nonsingular for all $t \geqq t_1$.

The main purpose of this section will be to show that the condition 2.2 for complete controllability of (1.2) is equivalent to $W(t_0, t_1)$ being nonsingular for some $t_1 > t_0$. This condition has the advantage of not depending on knowledge of a fundamental solution.

Before stating the main theorem, a simple computation yields, for $t_0 < t_1 < t_2$,

$$W(t_0, t_2) = W(t_0, t_1) + \Phi(t_0, t_1)W(t_1, t_2)\Phi^T(t_0, t_1).$$

Thus if $W(t_1, t_2)$ is nonsingular (positive definite) it follows that $W(t_0, t_2)$ is also nonsingular (positive definite). The reverse implication need not be true.

Theorem 2.1. *A necessary and sufficient condition that $W(t_1, t_2)$ be nonsingular for all $t_2 > t_1$ is that the Pfaffian* (2.7) *be not integrable at t_1.*

For ease in both using and proving this theorem, we list the implications and their contrapositives.

(A) *Necessary condition.* $W(t_1, t_2)$ nonsingular for all $t_2 > t_1$ implies Pfaffian (2.7) is not integrable at t_1.

(B) *Necessary; contrapositive.* Pfaffian (2.7) integrable at t_1 implies $W(t_1, t_2)$ is singular for some $t_2 > t_1$.

(C) *Sufficient condition.* Pfaffian (2.7) not integrable at t_1 implies $W(t_1, t_2)$ is nonsingular for all $t_2 > t_1$.

(D) *Sufficient; contrapositive.* $W(t_1, t_2)$ singular for some $t_2 > t_1$ implies Pfaffian (2.7) is integrable at t_1.

Proof. We shall prove (B) and (D).

Assume the Pfaffian (2.7) is integrable at t_1. Then there is a vector b, which is a linear combination of the rows of B, and an $\epsilon > 0$ such that $\dot{b}(t) = -b(t)A(t)$ for $t_1 \leqq t \leqq t_1 + \epsilon$. Let $\Phi(t, t_1)$, $\Phi(t_1, t_1) = I$, be the fundamental solution of $\dot{x} = A(t)x$. Then the vector b admits the representation $b(t) = c\Phi^{-1}(t, t_1) = c\Phi(t_1, t)$ for some constant vector c. Let $h(t)$ be any column of $H(t)$. Then $0 = b(t)h(t) = c\Phi(t_1, t)h(t)$. Since h is an arbitrary column of H, and W is positive semidefinite, we have $cW(t_1, t)c^T = 0$ for $t_1 \leqq t \leqq t_1 + \epsilon$, showing that there exists a $t_2 > t_1$ such that $W(t_1, t_2)$ is singular.

Assume, next, that $W(t_1, t_2)$ is singular for some $t_2 > t_1$. From Remark

3, it follows that $W(t_1, t)$ is singular for all $t_1 \leqq t \leqq t_2$. This implies there exists a vector $c(t_2)$ such that $c(t_2)W(t_1, t_2)c^T(t_2) = 0$. Since the integrand of the integral defining $W(t_1, t_2)$ is continuous,

$$c(t_2)\Phi(t_1, t)H(t)H^T(t)\Phi^T(t_1, t)c^T(t_2) \equiv 0 \quad \text{for} \quad t_1 \leqq t \leqq t_2 .$$

It follows that $0 \equiv c(t_2)\Phi(t_1, t)H(t) \equiv c(t_2)\Phi^{-1}(t, t_1)H(t)$. Thus b defined by $b(t) \equiv c(t_2)\Phi^{-1}(t, t_1)$ is an admissible vector in the sense that $b(t)H(t) \equiv 0$, i.e., b lies in the subspace spanned by the rows of B.

Define the scalar valued function $\psi(t, x) = c(t_2)\Phi^{-1}(t, t_1)x$. Then $\psi_x(t, x) = b(t)$, $\psi_t(t, x) = -b(t)A(t)x$ for $t_1 \leqq t \leqq t_2$, showing that the Pfaffian (2.7) is integrable at t_1.

The following illustrates the advantage of a definition of complete controllability for linear systems which does not depend on knowledge of a fundamental solution.

It is known that an n-dimensional system which is formed from a single nth *order equation having constant coefficients and the control as forcing term is completely controllable. We next show that this is also true for time varying systems of the form*

$$x^{(n)}(t) + a_1(t)x^{(n-1)}(t) + \cdots + a_n(t)x(t) = u(t).$$

Specifically we shall show that for any t_0 the associated Pfaffian is not integrable, implying $W(t_0, t_1)$ is nonsingular for *all* $t_1 > t_0$.

We take the equivalent linear system of the form $\dot{y}(t) = A(t)y(t) + h(t)u(t)$, where

$$A(t) = \begin{bmatrix} 0 & 1 & 0 \cdots & 0 \\ 0 & 0 & 1 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & & \cdot 0 & 1 \\ -a_n & -a_{n-1} & \cdots\cdot & -a_1 \end{bmatrix}, \qquad h(t) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

One can choose $B(t)$ as the $(n - 1) \times n$ matrix

$$B(t) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

The Pfaffian system equivalent to (2.7) is then

$$\begin{aligned} dx_1 - x_2\,dt &= 0 \\ dx_2 - x_3\,dt &= 0 \\ &\vdots \\ dx_{n-1} - x_n\,dt &= 0. \end{aligned} \tag{2.8}$$

If (2.8) is to be integrable there must exist scalar valued functions $\alpha_j(t)$, not all zero, so that the single Pfaffian

$$\sum_{j=1}^{n-1} \alpha_j(t)\, dx_j + 0\, dx_n - \sum_{j=1}^{n-1} \alpha_j(t) x_{j+1}\, dt$$

is an exact differential. But this would imply $\alpha_j(t) = 0$, $j = 1, 2, \cdots, (n-1)$, which shows (2.8) is not integrable for any t_0.

Geometric interpretation, local controllability, and the singular problem. By associating a Pfaffian system of the form (2.3) with the system (2.1), it is conspicuous that the stress is taken away from the functional form of the elements of the matrix H, and placed only on what the range of $H(t, x)$, considered as an operator on E^r, is. This obviously should be the case when controls are required to be only $\mathfrak{L}_2$ functions.

In [9] Markus and Lee consider a system of the form $\dot{x} = f(x, u)$, $f \in C^1$ in $E^n \times \Omega$, where Ω is a compact set contained in E^r with 0 in its interior and is the range set of the control. Assuming $f(0, 0) = 0$ and letting $A = f_x(0, 0)$, $H = f_u(0, 0)$, it is shown that if the linear system $\dot{x} = Ax + Hu$ is completely controllable, then the set of points from which the origin can be reached in finite time by trajectories of $\dot{x} = f(x, u)$ is an open connected set containing the origin. Kalman [10] pointed out that a similar result can be obtained for a system of the form $\dot{x} = f(t, x, u)$ by assuming the linear approximation is completely controllable in terms of the criterion given in Corollary 1.1.

The system

$$\dot{x}(t) = f(t, x(t), u(t)), \qquad x(t_0) = x_0, \tag{2.9}$$

where x is an n-vector, f is a C^2 vector-valued function and u is an r vector-valued measurable control, is said to be *locally controllable* along a solution φ^v corresponding to control v if for some $t_1 > t_0$ all points in some state space (n dimensional) neighborhood of $\varphi^v(t_1)$ are attainable in time t_1 by trajectories of (2.9) with admissible control.

It would be somewhat fallacious to say that a time dependent system is locally controllable, say at the origin, if all points in a neighborhood of the origin in state space are attainable by trajectories of the system in finite time. To see this, we consider the following example of G. Haynes.

Example.

$$\dot{x}_1 = -x_2 + (\cos t)u, \qquad x(0) = 0, \qquad |\, u(t)\, | \leqq 1,$$
$$\dot{x}_2 = x_1 + (\sin t)u.$$

An integral of the motion is seen to be $x_1 \sin t - x_2 \cos t = 0$, which one can picture as a rotating (with time) line in x_1, x_2 space. As t varies from

0 to 2π, all points of E^2 are swept out by this line. Now multiply the first equation by cos t, the second by sin t and one obtains by adding:

$$\frac{d}{dt}(x_1 \cos t + x_2 \sin t) = u,$$

or

$$x_1 \cos t + x_2 \sin t = \int_0^t u(\tau)\, d\tau.$$

Combining this with the integral of the motion gives

$$x_1^2(t) + x_2^2(t) = \left[\int_0^t u(\tau)\, d\tau\right]^2,$$

implying that as time increases, the two-dimensional neighborhoods of the origin of E^2 which are attainable also increase.

Since all solutions lie on a surface in (t, x) space, one would hardly feel that the system should be termed locally controllable; indeed it is *not* locally controllable by the definition given above.

We next proceed with an analysis, similar to that used in [9] and [10], to examine local controllability about a given trajectory of the system (2.1). Let $x(t_0) = 0$ be initial data for this system, v an arbitrary $\mathfrak{L}_2$ control and φ^v the corresponding solution. Let $u(t; \xi)$, $\xi \in E^n$, be a family of controls such that $u(t; 0) = v(t)$, u_ξ exists, and denote $x(\cdot\,; \xi)$ as the response to $u(\cdot\,; \xi)$. Then $x(\cdot\,; \xi)$ satisfies

$$x(t; \xi) \equiv \int_{t_0}^t [g(\tau, x(\tau; \xi)) + H(\tau, x(\tau; \xi))u(\tau; \xi)]\, d\tau.$$

$$x_\xi(t; 0) \equiv \int_{t_0}^t [g_x(\tau, \varphi^v(\tau)) + H_x(\tau, \varphi^v(\tau))v(\tau)]x_\xi(\tau; 0) + H(\tau, \varphi^v(\tau))u_\xi(\tau; 0)\, d\tau,$$

where $H_x v$ is an $n \times n$ matrix with i, jth element

$$\sum_{\nu=1}^{r} H_{x_j}^{i\nu} v^\nu.$$

For each $t \geqq t_0$, we view $x(t; \xi)$ as a mapping $\xi \to x$ with $0 \to \varphi^v(t)$. Let $Z(t; \varphi^v, u_\xi)$ denote the Jacobian matrix $x_\xi(t; 0)$. We have: *If for some $\bar{t}$, u_ξ, $Z(t; \varphi^v, u_\xi)$ is nonsingular, the attainable set at $\bar{t}$ contains a neighborhood of the point $\varphi^v(\bar{t})$.* Let $\Phi(t, t_0)$ be a fundamental solution matrix of the system

$$\dot{x}(t) = [g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t)]x(t).$$

Then

$$Z(t; \varphi^v, u_\xi) \equiv \int_{t_0}^{t} \Phi(t, \tau)H(\tau, \varphi^v(\tau))u_\xi(\tau; 0)\, d\tau.$$

From Lemma 1.1 and Corollary 1.1 we have:

THEOREM 2.2 (*Kalman*). *A necessary and sufficient condition that there exist an $r \times n$ matrix u_ξ such that $Z(t_1 ; \varphi^v, u_\xi)$ is nonsingular for some $t_1 > t_0$ is that the linear system*

$$\dot{y}(t) = [g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t)]y(t) + H(t, \varphi^v(t))u(t)$$

is completely controllable.

In terms of the Pfaffian approach the equivalent theorem is the following.

THEOREM 2.3. *A necessary and sufficient condition that there exist an $r \times n$ matrix u_ξ such that $Z(t_1 ; \varphi^v, u_\xi)$ is nonsingular for some $t_1 > t_0$, is that the Pfaffian system*

$$B(t, \varphi^v(t))\, dx - B(t, \varphi^v(t))[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t)]x\, dt = 0$$

be nonintegrable for some $t_1 \geqq t_0$, i.e., that

$$(2.10) \quad b(t, \varphi^v(t))dx - b(t, \varphi^v(t))[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t)]\, x\, dt$$

is not an exact differential for any b which is a linear combination of the rows of B.

The same method, when applied to a system of the form (2.9) yields:

THEOREM 2.3′. *A sufficient condition that there exists a $t_1 \geqq t_0$, such that all points in some state space neighborhood of $\varphi^v(t_2)$ for all $t_2 > t_1$ are attainable in time t_2 by trajectories of* (2.9) *with admissible controls, is that there exists a $t_1 \geqq t_0$ such that the Pfaffian system*

$$B(t; v)\, dy - B(t; v)f_x(t, \varphi^v(t), v(t))y\, dt = 0$$

is not integrable at t_1.

[The notation $B(t;\ v)$ is used to denote the dependence of B on the reference trajectory, specifically $B(t; v)f_u(t, \varphi^v(t), v(t)) \equiv 0$.]

It is interesting at this point to see the implications of the assumption that (2.10) *is* an exact differential. This implies and is implied by

$$(2.11) \quad \frac{d}{dt} b(t, \varphi^v(t)) \equiv -b(t, \varphi^v(t))[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t)],$$

which can be recognized as the so-called adjoint system of the maximum principle [11] approach to the time optimal problem for system (2.1). It should be noted that if $b(t,\ \varphi^v(t))$ satisfies (2.11), then it is an adjoint vector which is orthogonal to all of the columns of H. Since the maximum principle (for control components bounded by one in absolute value) implies one chooses

$$u^j(t) = \operatorname{sgn} \sum_{i=1}^{n} b^i(t, \varphi^v(t))H^{ij}(t, \varphi^v(t)),$$

in this case it yields no information.

I shall designate such a problem as one which admits a *totally singular* arc φ^v, i.e., where the maximum principle yields no information in the time optimal problem for any components of the optimal control. The arc would be singular, but not totally singular, if there is an adjoint vector orthogonal to some, but not all, columns of H.

THEOREM 2.4. *The Pfaffian form* (2.10) *is an exact differential if and only if φ^v is a totally singular arc.*

Proof. It has been shown above that if (2.10) is an exact differential, then the vector b satisfies (2.11), which implies φ^v is a totally singular arc. If φ^v is a totally singular arc, there exists a vector $p(t)$ such that

$$\text{(i)} \qquad p(t)H(t, \varphi^v(t)) \equiv 0$$

and

$$\text{(ii)} \qquad \dot{p}(t) = -p(t)[g_x(t, \varphi^v(t)) + H_x(t, \varphi^v(t))v(t)].$$

From (i) we conclude that $p(t)$ is a linear combination of the rows of $B(t, \varphi^v(t))$, while (ii) implies that this linear combination, (2.10), is an exact differential.

To summarize: φ^v *not* a totally singular arc implies the Pfaffian form (2.10) is *not* an exact differential, which implies there exist $\bar{t} \geqq t_0$ and u_ξ such that $Z(\bar{t}; \varphi^v, u_\xi)$ is nonsingular and the attainable set at time $\bar{t}$ contains a neighborhood of the point $\varphi^v(\bar{t})$. The contrapositive of this statement provides an interesting characterization of totally singular arcs, i.e., if for every $t_1 > t_0$ there exist points in every state space neighborhood of $\varphi^v(t_1)$ which are not attainable in time t_1 with $\mathfrak{L}_2$ controls, the arc φ^v is totally singular. On the other hand, as will be shown by example, a totally singular arc can remain on the boundary of the attainable set, and thus provide a time optimal trajectory.

THEOREM 2.5. *If the system* (2.1) *is not completely controllable at t_0, $Z(t; \varphi^v, u_\xi)$ is singular for all $t \geqq t_0$, u_ξ and all reference trajectories φ^v, i.e., every trajectory φ^v is totally singular.*

Proof. Any vector b, which is a linear combination of the rows of B, satisfies $b(t, x)H(t, x) \equiv 0$. Thus for any vector $v(t)$,

$$\frac{\partial}{\partial x}[b(t, x)H(t, x)v(t)] \equiv 0,$$

or

$$v(t)H^T(t, x)b_x(t, x) \equiv -b(t, x)H_x(t, x)v(t).$$

Evaluation of this identity at the point $(t, \varphi^v(t))$, substitution into (2.11) and expansion of the left side yield

$$(2.12)\quad \begin{aligned} &b_t(t, \varphi^v(t)) + b(t, \varphi^v(t))g_x(t, \varphi^v(t)) + g(t, \varphi^v(t))b_x^T(t, \varphi^v(t)) \\ &\qquad \equiv v(t)H^T(t, \varphi^v(t))[b_x(t, \varphi^v(t) - b_x^T(t, \varphi^v(t))]. \end{aligned}$$

This identity provides a necessary and sufficient condition that (2.10) be an exact differential, i.e., that φ^v be totally singular.

Now assume the system (2.1) is not completely controllable. This means that for some b, a linear combination of the rows of B, the Pfaffian form $b(t, x)\,dx - b(t, x)g(t, x)\,dt$ is an exact differential, or

$$\begin{aligned} b_t(t, x) &\equiv -b(t, x)g_x(t, x) - g(t, x)b_x^T(t, x), \\ b_x(t, x) &\equiv -b_x^T(t, x) \equiv 0. \end{aligned}$$

Evaluating these two identities at $(t, \varphi^v(t))$ for an arbitrary control v shows that (2.12) is satisfied, hence every trajectory φ^v is totally singular.

A conjecture which one might be tempted to make is that if the system (2.1) is completely controllable, it admits no totally singular arcs. This is *not* true, as the following example from [2] shows.

Example 2.1.

$$\begin{aligned} \dot{x}_1 &= x_1^2 - x_1^2x_2u, \qquad & x_1(0) &= 1, \\ \dot{x}_2 &= -x_2 + u, & x_2(0) &= 0. \end{aligned}$$

For the time optimal problem of reaching the point (2, 0), it is shown in [2] that $u \equiv 0$ is the optimal control, if the restriction $|\, u(t) \,| \leqq 1$ is imposed, and it easily follows that this is also optimal in the class of $\mathfrak{L}_2$ controls.

For this problem, one can use for the matrix B the single vector $b = (1, x_1^2x_2)$. The associated Pfaffian equation is

$$dx_1 + x_1^2x_2\,dx_2 + x_1^2(x_2^2 - 1)\,dt = 0.$$

Let $x = (x_1, x_2)$, $a(x) = (1, x_1^2x_2, x_1^2(x_2^2 - 1))$. Then

$$(\mathrm{curl}\; a(x))\cdot a(x) = 2x_2x_1^2 \neq 0;$$

thus the Pfaffian is not integrable.

The optimal path from the point (1, 0) to $(\alpha, 0)$, $\alpha > 1$, is obtained with control $u \equiv 0$, and is

$$\varphi^0(t) = \begin{cases} \dfrac{1}{1 - t} \\ 0 \end{cases}.$$

This is a totally singular arc. To show this, we note $b(t, \varphi^0(t)) \equiv (1, 0)$.

$$b(t, \varphi^0(t))\,dx - b(t, \varphi^0(t))[g_x(t, \varphi^0(t)) + H_x(t, \varphi^0(t))\cdot 0]x\,dt$$
$$= dx_1 + 0\,dx_2 - \frac{2x_1}{1-t}\,dt.$$

Let

$$\bar{a}(x, t) \equiv \left(1, 0, -\frac{2x_1}{1-t}\right).$$

Then (curl $\bar{a}$)$\cdot\bar{a} \equiv 0$, which implies the Pfaffian

$$dx_1 + 0\,dx_2 - \frac{2x_1}{1-t}\,dt = 0$$

is integrable, and φ^0 is a totally singular arc. Here the arc φ^0 is on the boundary of the attainable set.

It should be stressed at this point that while the nonsingularity of the matrix $Z(t; \varphi^v, u_\xi)$ is sufficient for local controllability, it has *not* been shown and is *not* true that this is a necessary condition. To show this we will construct a time optimal problem (Example 2.2) possessing a totally singular arc which yields neither a maximum or minimum. This arc, together with the control which produces it and the nonzero solution of the corresponding adjoint system, satisfies the maximum principle. For one thing this points out that the maximum principle is, of course, only a necessary condition; but, of more importance, the example is constructed such that any theory based on a linearized or variational principle will be inconclusive. The abovementioned trajectory can be thought of as being an "inflection point" in function space for the functional (time). It is easily seen that a problem can be constructed such that at the inflection point (trajectory) the values of the functional are "so flat that all order derivatives vanish." It becomes difficult to show that such a trajectory is in the interior of the attainable set, and therefore cannot be optimal. We shall first prove a rather special theorem (Theorem 2.6) which will allow us to show local controllability along the totally singular arc of Example 2.2.

Consider a control system of the form studied in [2], i.e.,

$$\begin{aligned} \dot{x}_1(t) &= A_1(x(t)) + B_1(x(t))u(t), \qquad x(0) = x_0, \\ \dot{x}_2(t) &= A_2(x(t)) + B_2(x(t))u(t), \qquad |u(t)| \leqq 1. \end{aligned} \tag{2.13}$$

Assume that in some region of interest $\mathfrak{D}$ of state space

$$\Delta(x) \equiv -B_2(x)A_1(x) + B_1(x)A_2(x) \neq 0, \tag{2.14}$$

and that A_i, B_i, $i = 1, 2$, are $C^1(\mathfrak{D})$.

The Pfaffian system associated with (2.13) is the single Pfaffian equation

$$B_2(x)\,dx_1 - B_1(x)\,dx_2 + \Delta(x)\,dt = 0. \tag{2.15}$$

Since $\Delta(x) \neq 0$ and multiplication by a factor does not change integrability, this can be rewritten as

$$\frac{B_2(x)}{\Delta(x)}\,dx_1 - \frac{B_2(x)}{\Delta(x)}\,dx_2 + dt = 0. \tag{2.16}$$

Let

$$Z(x) = \left(\frac{B_2(x)}{\Delta(x)}, -\frac{B_1(x)}{\Delta(x)}, 1\right);$$

then a necessary and sufficient condition that the Pfaffian (2.16) be integrable at a point (t, x) is that $Z(x)\cdot\text{curl}\; Z(x) \equiv 0$ in a neighborhood of x. Computing yields

$$Z(x)\cdot\text{curl}\; Z(x) \equiv -\left[\frac{\partial}{\partial x_1}\left(\frac{B_1(x)}{\Delta(x)}\right) + \frac{\partial}{\partial x_2}\left(\frac{B_2(x)}{\Delta(x)}\right)\right] \equiv -\omega(x),$$

where $\omega(x)$ (using the notation of [2]) can be directly computed from the right sides of the differential equations (2.13).

Let v be a continuous control (this is sufficient continuity when the control appears linearly) satisfying $|\, v(t)\, | < 1$, and let φ^v be the corresponding trajectory of (2.13).

THEOREM 2.6. *If for some* $t_1 \geqq t_0$, $\varphi^v(t_1)$ *is not a zero of* ω, *then for any* $t_2 > t_1$ *all points in some state space neighborhood of* $\varphi^v(t_2)$ *are attainable by trajectories of* (2.13), *in time* t_2, *with admissible controls.*

Proof. The variational equation for the system (2.13) about the trajectory φ^v is given by

$$\dot{y}(t) = [A_x(\varphi^v(t)) + v(t)B_x(\varphi^v(t))]y(t) + B(\varphi^v(t))u(t),$$

where

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

The Pfaffian equivalent to (2.10) for this variational equation is

$$\begin{aligned} &B_2(\varphi^v(t))\,dy_1 - B_1(\varphi^v(t))\,dy_2 \\ &\qquad + (-B_2(\varphi^v(t)), B_1(\varphi^v(t)))[A_x(\varphi^v(t)) \\ &\qquad\qquad + v(t)B_x(\varphi^v(t))]y\,dt = 0. \end{aligned} \tag{2.17}$$

A sufficient condition that (2.17) be *not* integrable at t_1 is that

$$\begin{aligned} \frac{d}{dt}\,(B_2(\varphi^v(t)), B_1(\varphi^v(t)))\,|_{t=t_1} &\neq (-B_2(\varphi^v(t_1)), B_1(\varphi^v(t_1))) \\ &\quad \cdot[A_x(\varphi^v(t_1)) + v(t_1)B_x(\varphi^v(t_1))], \end{aligned} \tag{2.18}$$

which is implied by $\omega(\varphi^v(t)) \neq 0$, as can be shown by a straightforward calculation. [In terms of Theorem 2.4, (2.18) states that $\varphi^v(t_1)$ is *not* a point of a singular arc. In [2, p. 97] it is shown that for systems of this type, singular arcs are characterized by the fact that ω is zero along them. It follows that if $\varphi^v(t_1)$ is *not* a zero of ω, then it is *not* a point of a singular arc; hence, (2.17) is not integrable and the conclusion of the theorem follows.]

It should be stressed that the integrability of (2.16) requires $\omega(x) = Z(x)\cdot\text{curl}\, Z(x)$ to be zero in a neighborhood of a point, while Theorem 2.6 deals only with the value of ω at a point. It is possible (see Example 2.1) to have the Pfaffian (2.16) not integrable at a point $(\bar{t}, \bar{x})$ at which $\omega(\bar{x}) = 0$, and yet have a trajectory φ^v such that $\varphi^v(\bar{t}) = \bar{x}$ and the system is not locally controllable about φ^v.

We next give the example of a problem which is locally controllable along a totally singular arc.

Example 2.2 (A singular arc $\varphi^0(t)$ such that all points in a neighborhood of $\varphi^0(t_1)$ are attainable in time t_1). Consider the system

$$\begin{aligned} \dot{x}_1 &= u, & |\, u(t)\, | &\leqq 1, \\ \dot{x}_2 &= 1 + x_2 x_1^2 u, & x(0) &= 0. \end{aligned}$$

Then $\Delta(x) = 1$, $\omega(x) = x_1^2$. Hence, if we consider the time optimal problem of reaching the final point $x_f(0, \frac{1}{2})$, the Green's Theorem approach [2] yields Fig. 1. The optimal arc is shown by the arrows. There is an arc along which $\omega = 0$, i.e., $x_1 \equiv 0$, and while this can be attained with the control $u \equiv 0$ it yields neither a maximum nor minimum to the time optimal problem. This arc we designate as φ^0:

$$\varphi^0(t) = \begin{cases} \varphi_1^0(t) \equiv 0, \\ \varphi_2^0(t) \equiv t. \end{cases}$$

It is easily checked that the variational equation along φ^0 is *not* completely controllable.

Now consider a relation $x_1 = k_1 \sin k_2 x_2$, where $k_1 > 0$ and $k_2 > 4\pi$. It will be shown that for k_1 sufficiently small, there exists a unique admissible

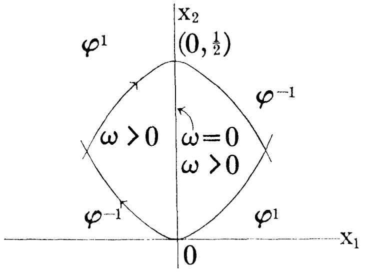


FIG. 1

continuous control $\bar{u}(t)$ with trajectory $\varphi^{\bar{u}}$ which has $\{(x_1, x_2): x_1 = k_1 \cdot \sin k_2 x_2, x_2 \geqq 0\}$ as its track.

From the Green's Theorem approach [2] and the symmetry of $\omega(x)$ about the line $x_1 = 0$, the parametrization of $\varphi^{\bar{u}}$ must be such that at the even numbered crossings of the x_2-axis, counting only crossings which occur for $x_2 > 0$, one must have

$$\varphi_1^{\bar{u}}\left(\frac{2n\pi}{k_2}\right) = 0 = \varphi_1^{0}\left(\frac{2n\pi}{k_2}\right),$$
$$\varphi_2^{\bar{u}}\left(\frac{2n\pi}{k_2}\right) = \frac{2n\pi}{k_2} = \varphi_2^{0}\left(\frac{2n\pi}{k_2}\right).$$

We will be interested in the case $n = 1$, so that $2\pi/k_2 < \frac{1}{2}$. It will be shown that there is local controllability along $\varphi^{\bar{u}}$, and since $\varphi^{\bar{u}}(2\pi/k_2) = \varphi^{0}(2\pi/k_2)$, it will follow that a neighborhood of $\varphi^{0}(2\pi/k_2)$ is attainable in time $2\pi/k_2$.

First we will show that for k_1 sufficiently small, there is a unique continuous u which leads to a trajectory φ^{u} having $\{(x_1, x_2): x_1 = k_1 \sin k_2 x_2, x_2 \geqq 0\}$ as its track. Differentiation of the track relation with respect to t yields

$$\dot{x}_1(t) = k_1 k_2 [\cos k_2 x_2(t)] \dot{x}_2(t).$$

Substitution from the system equations leaves

$$u(t) = k_1 k_2 [\cos k_2 x_2(t)][1 + x_2(t) x_1^2(t) u(t)]. \tag{2.19}$$

For any control u,

$$x_1(t) = \int_0^t u(\tau)\, d\tau,$$
$$x_2(t) = \exp\left[\int_0^t u(\tau)\left(\int_0^\tau u(\sigma)\, d\sigma\right)^2 d\tau\right] \cdot \int_0^t \exp\left[-\int_0^\tau u(\sigma)\left(\int_0^\sigma u(\gamma)\, d\gamma\right)^2 d\sigma\right] d\tau.$$

Substituting these in (2.19) yields an expression of the form

$$u(t) = k_1(\mathfrak{F}u)(t),$$

where the definition of the nonlinear operator $\mathfrak{F}$ is obvious. Let $C[0, \frac{1}{2}]$ denote the space of continuous vector-valued functions u on the interval $[0, \frac{1}{2}]$ with the supremum norm, and $B^{1/2}$ the closed ball of radius $\frac{1}{2}$ in this space. It is easily shown that for k_1 sufficiently small but positive, $u \in B^{1/2}$ implies $k_1 \mathfrak{F} u \in B^{1/2}$, and $k_1 \mathfrak{F}$ is a contracting map. Thus $k_1 \mathfrak{F}$ has a unique fixed point in $B^{1/2}$. Call this point $\bar{u}$. Then $\varphi^{\bar{u}}$ is *not* a singular trajectory, since k_1 positive implies $\bar{u}(t) \not\equiv 0$, and $\varphi^{\bar{u}}$ has the desired track.

Now for $0 < t_1 < \pi/k_2$, $\varphi^{\bar{u}}(t_1)$ is not a point of the singular arc, hence not a zero of ω. From Theorem 2.6 it follows that all points in some neighborhood of $\varphi^{\bar{u}}(t_2)$, for any $t_2 > t_1$, are attainable in time t_2 by trajectories with admissible controls; hence this is true for $t_2 = 2\pi/k_2$.

To determine local controllability along $\varphi^{\bar{u}}$ by use of the fundamental solution of the variational equation about this trajectory would be a virtually impossible task.

In concluding, it should be noted that totally singular arcs were defined with no mention made of transversality conditions. It is possible to use these conditions, in very special cases, to rule out the existence of singular arcs in the optimal strategy. Also, for a time optimal problem for a system of the form

$$\dot{x}(t) = g(x(t)) + H(x(t))u(t), \tag{2.20}$$

the maximum principle yields the fact that the Hamiltonian is constant along the optimal path. We shall show that this cannot be used to rule out totally singular arcs, since such arcs automatically satisfy the condition even though the Hamiltonian is seemingly a function of time along them.

For the system (2.20) with any given control $u(t)$ we define the Hamiltonian for the time optimal problem as

$$\mathcal{H}(t, x, p) \equiv p \cdot g(x) + p \cdot H(x)u(t) + 1.$$

A necessary condition is that $\mathcal{H}$ is a constant along the optimal trajectory; it need not be so on a nonoptimal trajectory. Define the adjoint system as

$$\dot{p}(t) = -p(t)g_x(x, (t)) - p(t)H_x(x(t))u(t). \tag{2.21}$$

THEOREM 2.7. *The Hamiltonian for the system* (2.20) *is constant along any totally singular arc.*

Proof. We defined a totally singular arc as an arc φ^u which satisfies (2.20) and for which there exists an adjoint vector $p(t)$ satisfying (2.21) such that $p(t)H(\varphi^u(t)) \equiv 0$ for a set of t values having positive measure. Then

$$\frac{d}{dt}\mathcal{H}(t, \varphi^u(t), p(t)) \equiv \frac{d}{dt}[p(t) \cdot g(\varphi^u(t)) + 1] \equiv \dot{p}_i g^i + p_i g^i_{x_\nu} \dot{\varphi}_\nu^u. \tag{2.22}$$

From (2.20),

$$g^i \equiv \dot{\varphi}_i^u - H^{ik}u_k .$$

From (2.21),

$$p^i g^i_{x_\nu} = -\dot{p}_\nu - p_i H^{ik}_{x_\nu} u_k .$$

Substituting in (2.22),

$$\frac{d}{dt}\mathcal{H}(t, \varphi^u(t), p(t)) \equiv \dot{p}_i[\dot{\varphi}_i^u - H^{ik}u_k] + [-\dot{p}_\nu - p_i H^{ik}_x u_k]\dot{\varphi}_\nu^u$$

$$= [-\dot{p}_i H^{ik} - p_i H^{ik}_{x_\nu} \dot{\varphi}_\nu^u]u_k \equiv -\left\{\frac{d}{dt}[p(t)H(\varphi^u(t))]\right\} u = 0,$$

from the condition $p(t)H(\varphi^u(t)) = 0$.

REFERENCES

[1] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, (1960), pp. 102–119.

[2] H. Hermes and G. Haynes, *On the nonlinear control problem with control appearing linearly*, this Journal, 1 (1963), pp. 85–107.

[3] C. Carathéodory, *Untersuchungen über die Grundlagen der Thermodynamik*, Math. Ann., (1909), pp. 355–386.

[4] W. L. Chow, *Über Systeme von linear partiellen Differentialgleichungen erster Ordnung*, Math. Ann., (1940), pp. 95–105.

[5] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol 5, Princeton University Press, Princeton, 1960, pp. 1–24.

[6] R. E. Kalman, Y. C. Ho, and K. S. Narendra, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1, pp. 189–213.

[7] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[8] J. Milnor, *Morse Theory*, Annals of Math. Studies, **51**, Princeton University Press, Princeton, 1963.

[9] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[10] R. E. Kalman, *Discussion of* [9], Trans. ASME Ser. D. J. Basic Engrg., 84 (1962).

[11] L. S. Pontryagin, V. G. Boltyanski, R. V. Gankrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

J.SIAM Control
Ser. A, Vol. 2, No. 2
Printed in U.S.A., 1965

THE SYNTHESIS OF LINEAR OPTIMAL SYSTEMS*

T. G. BABUNASHVILI†

A series of works has been devoted to the synthesis of linear optimal systems. Among these we note the work of Neustadt [1], and N. N. Krasovskii [2]. In [1] the synthesis problem for homogeneous linear systems was solved completely. The general case of inhomogeneous linear systems was considered in [2], but the method proposed there is too complicated.

Here we shall describe a new synthesis method that is suitable for arbitrary inhomogeneous (nondegenerate, see below) linear systems. In connection with the method presented here, also see the work of Antosiewicz [3].

1. Statement of the problem. Let there be given the equation

$$\dot{x} = A(t)x + B(t)u + f(t). \tag{1}$$

Here, x is an n-dimensional phase column vector, $A(t)$ is a summable $n \times n$ matrix (i.e., a matrix whose elements are summable on any bounded interval of the time axis), u is an r-dimensional control column vector, $B(t)$ is a summable $n \times r$ matrix, and $f(t)$ is a summable n-dimensional column vector. The control u is sought in the class of measurable functions with values in a given convex, compact polyhedron U in r-space that contains the origin.

The problem consists in finding, for a given initial position x_0 in phase space, the optimal control that transfers the phase point along the corresponding (optimal) trajectory of (1) from x_0 to the origin in minimum time.

Let us write the equation

$$\dot{\psi} = -\psi A(t), \tag{2}$$

where ψ is an n-dimensional row vector, and let us define the "norm"

* Originally published in Dokl. Akad. Nauk SSSR, 155 (1964), pp. 295–298. Translated from the Russian by L. W. Neustadt, University of Michigan, Ann Arbor, Michigan. The original article was submitted on October 24, 1963, and was presented by academician L. S. Pontryagin on November 25, 1963. The work was carried out in the seminar of L. S. Pontryagin on the theory of oscillations and automatic control.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† The V. A. Steklov Mathematical Institute of the Academy of Sciences of the USSR.

$\| v \|$ in the space of r-dimensional row vectors v according to the formula

$$\| v \| = \max_{u \in U} vu.$$

A necessary condition for optimality (the maximum principle, see [4]) can now be formulated as follows.

For every optimal control $u(t)$, $0 \leqq t \leqq T$, *there exists a nonzero solution* $\psi(t)$, $0 \leqq t \leqq T$, *of* (2) *such that, almost everywhere in* $[0, T]$,

$$\psi(t)B(t)u(t) = \| \psi(t)B(t) \|. \tag{3}$$

Equation (1) is assumed to be nondegenerate (see [4]); this is equivalent to the assertion that, for any given nonzero solution $\psi(t)$ of (2), the control $u(t)$ is uniquely defined by the maximum condition (3) for almost all t.

Thus, if the origin can be attained from a given x_0, the optimal problem (for the given x_0) will be solved if we can compute the initial value $\psi_{x_0} = \psi(0)$ of the corresponding solution $\psi(t)$ of (2). We shall call the computation of the vector ψ_{x_0} corresponding to the vector x_0 the synthesis of the optimal system described by (1).

2. Derivation of the fundamental equation (6) (also see [2] and [3]). The solution $x(t)$ of (1) with initial condition $x(0) = x_0$ has the form

$$x(t) = \Phi(t) \left[x_0 + \int_0^t \Phi^{-1}(\tau)(B(\tau)u(\tau) + f(\tau))\, d\tau \right],$$

where $\Phi(t)$ is the fundamental matrix for the homogeneous equation $\dot{x} = A(t)x$, normalized at $t = 0$. Let T_{x_0} be the optimal transfer time from x_0 to the origin. Finding the optimal control $u_{x_0}(t)$, $0 \leqq t \leqq T_{x_0}$, is equivalent to solving the equation

$$\begin{aligned} z(T) = -\left(x_0 + \int_0^T \Phi^{-1}(t)f(t)\, dt \right) &= \int_0^T \Phi^{-1}(t)B(t)u(t)\, dt \\ &= \int_0^T K(t)u(t)\, dt, \end{aligned} \tag{4}$$

for the unknowns T and $u(t)$, $0 \leqq t \leqq T$; moreover, T is to be taken as the smallest positive root of this equation. We shall call the solution T_{x_0}, $u_{x_0}(t)$, $0 \leqq t \leqq T_{x_0}$, the optimal solution of (4).

In order that, for any given $T > 0$, *equation* (4) *have a solution* $u(t)$ *with* $u(t) \in U$, $0 \leqq t \leqq T$, *it is necessary and sufficient that*

$$\chi z(T) \leqq \int_0^T \| \chi K(t) \|\, dt, \tag{5}$$

for every n-dimensional row vector χ. For a proof, see [3].

THEOREM. *In order that* (4) *have an optimal solution* T_{x_0}, $u_{x_0}(t)$, $0 \leqq t \leqq T_{x_0}$, *it is necessary and sufficient that there exist a nonzero row vector* ψ_0 *that satisfies the equation*

$$(6)\quad \psi_0 z(T_{x_0}) = \int_0^{T_{x_0}} \| \psi_0 K(t) \| \, dt = \min_{\chi z(T_{x_0}) = \psi_0 z(T_{x_0})} \int_0^{T_{x_0}} \| \chi K(t) \| \, dt,$$

(*i.e., the minimum is taken over all* χ *that satisfy the condition* $\chi z(T_{x_0}) = \psi_0 z(T_{x_0})$). *Any solution* ψ_0 *of* (6) *may be taken for the vector* ψ_{x_0} *in defining the optimal control* $u_{x_0}(t)$, $0 \leqq t \leqq T_{x_0}$, *through the maximum condition* (3) *wherein* $\psi(t)$, $0 \leqq t \leqq T_{x_0}$, *is the solution of* (2) *with initial condition* $\psi_{x_0} = \psi(0)$.

Proof. Let χ be an arbitrary n-dimensional row vector that satisfies the condition $\chi z(T_{x_0}) > 0$, and let $\alpha \chi z(T_{x_0}) = \psi_0 z(T_{x_0})$. It follows from the nondegeneracy of (1) that $\alpha > 0$. It therefore follows from (6) that

$$\alpha \chi z(T_{x_0}) = \int_0^{T_{x_0}} \| \psi_0 K(t) \| \, dt \leqq \int_0^{T_{x_0}} \| \alpha \chi K(t) \| \, dt,$$

i.e., (5), or equivalently (4), is satisfied. Conversely, if T_{x_0}, $u_{x_0}(t)$, $0 \leqq t \leqq T_{x_0}$ is an optimal solution of (4), then, according to the maximum principle, there exists a solution $\psi(t) = \psi_{x_0} \Phi^{-1}(t)$ of (2) such that $\psi(t) B(t) u_{x_0}(t) = \| \psi_{x_0} K(t) \|$. Consequently, multiplying (4) by ψ_{x_0}, we obtain

$$\psi_{x_0} z(T_{x_0}) = \int_0^{T_{x_0}} \psi_{x_0} K(t) u_{x_0}(t) \, dt = \int_0^{T_{x_0}} \| \psi_{x_0} K(t) \| \, dt,$$

i.e., (6) holds.

It is easily seen that the control $u_0(t)$, $0 \leqq t \leqq T_{x_0}$, given by the equation $\psi_0 K(t) u_0(t) = \| \psi_0 K(t) \|$, where ψ_0 is any nonzero solution of (6), is optimal: $u_0(t) = u_{x_0}(t)$, $0 \leqq t \leqq T_{x_0}$. Indeed, if $u_0(t) \neq u_{x_0}(t)$ on a set of positive measure, then

$$\psi_0 z(T_{x_0}) = \int_0^{T_{x_0}} \psi_0 K(t) u_{x_0}(t) \, dt < \int_0^{T_{x_0}} \psi_0 K(t) u_0(t) \, dt$$

$$= \int_0^{T_{x_0}} \| \psi_0 K(t) \| \, dt,$$

contradicting (6).

Thus, the synthesis problem is equivalent to that of solving (6) for the unknowns ψ_0 and T_{x_0}; moreover, T_{x_0} must be taken as the smallest positive root of this equation.

3. The solution of equation (6). Equation (6) may be solved by the method of gradient descent as based on the following proposition.

For any $T > 0$, *the gradient of the function* $g(\chi) = \int_0^T \| \chi K(t) \| \, dt$ *with respect to* χ *is continuous. Every relative minimum of the function* $g(\chi)$ *under the condition* $\chi z(T) = \text{const.} > 0$ *is an absolute minimum of the function (under the given condition* $\chi z(T) = \text{const.}$*).*

Proof. By virtue of the nondegeneracy of (1), $\| \chi K(t) \| = \chi K(t) v_\chi(t)$, where $v_\chi(t)$ is a function that depends on $\chi \neq 0$, is piecewise constant on a set (of t) of full measure, and has values on the vertices of the polyhedron U. For a small change in χ, the function $v_\chi(t)$ changes on a set of small measure. Consequently, $\operatorname{grad} g(\chi) = \int_0^T K(t) v_\chi(t) \, dt$ changes continuously with χ. Let χ_1 and χ_2 be two stationary points of the function $g(\chi)$ under the condition $\chi z(T) = \text{const.} > 0$; we shall show that $g(\chi_1) = g(\chi_2)$. Let us assume the contrary, and let $g(\chi_1) > g(\chi_2)$. We have that

$$\operatorname{grad} g(\chi_i) = \int_0^T K(t) v_{\chi_i}(t) \, dt = \lambda_i z(T), \qquad i = 1, 2,$$

$$\chi_i \cdot \operatorname{grad} g(\chi_i) = \int_0^T \chi_i K(t) v_{\chi_i}(t) \, dt = \int_0^T \| \chi_i K(t) \| \, dt = g(\chi_i)$$

$$= \lambda_i \cdot \text{const.}, \qquad i = 1, 2.$$

Consequently, $\lambda_1 > \lambda_2$.

Further, we have that

$$\int_0^T K(t) (v_{\chi_1}(t) - v_{\chi_2}(t)) \, dt = (\lambda_1 - \lambda_2) z(T).$$

Multiplying both sides by χ_2, we obtain the relation

$$\int_0^T \chi_2 K(t) v_{\chi_1}(t) \, dt - \int_0^T \| \chi_2 K(t) \| \, dt = (\lambda_1 - \lambda_2) \cdot \text{const} > 0,$$

which is a contradiction since

$$\int_0^T \chi_2 K(t) v_{\chi_1}(t) \leqq \int_0^T \| \chi_2 K(t) \| \, dt.$$

The proposition that has just been proved yields the following method for the solution of (6).

We make a "first approximation" χ_1 for the solution ψ_0, subject to the single condition $\chi_1 z(0) > 0$, and begin to increase the time t from 0 up to the first instant t_1 (the "first approximation" to T_{x_0}) when $\chi_1 z(t_1) = \int_0^{t_1} \| \chi_1 K(t) \| \, dt$. (If $\chi_1 z(t) > \int_0^t \| \chi_1 K(\tau) \| \, d\tau$ for every $t > 0$, then the optimal problem with the given initial value x_0 evidently has no solution.) After this we find, by the method of gradient descent, the minimum of the function $g_1(\chi) = \int_0^{t_1} \| \chi K(t) \| \, dt$ under the condition $\chi z(t_1) = \chi_1 z(t_1)$. If the minimum point $\chi_2 \neq \chi_1$, then $\chi_2 z(t_1) > \int_0^{t_1} \| \chi_2 K(t) \| \, dt$, and we begin to increase the time from t_1 up to the instant t_2 when once again $\chi_2 z(t_2) = \int_0^{t_2} \| \chi_2 K(t) \| \, dt$, obtain "second approximations" t_2, χ_2, etc. It is easy to see that the increasing sequence $t_1 \leqq t_2 \leqq \cdots$ has a finite upper bound if, and only if, the optimal problem with the given initial value x_0 has a solution, and this upper bound is equal to the optimal time T_{x_0}. In case T_{x_0} is finite, the sequence of unit vectors $\chi_1/\| \chi_1 \|$, $\chi_2/\| \chi_2 \|, \cdots$ ($\| \chi \|$ is an arbitrary vector norm) converges to some compact set of vectors that make up all the solutions of unit length of (6).

REFERENCES

[1] L. W. NEUSTADT, *Synthesizing time optimal control systems*, J. Math. Anal. Appl., 1(1960), pp. 484–493.

[2] N. N. KRASOVSKII, *On a method of constructing optimal trajectories*, Mat. Sb., 53(95) (1961), pp. 195–206. (Russian)

[3] H. A. ANTOSIEWICZ, *Linear control systems*, Arch. Rational Mech. Anal., 12(1963), pp. 313–324.

[4] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience Division of John Wiley and Sons, New York, 1962.

*[5] J. H. EATON, *An iterative solution to time-optimal control*, J. Math. Anal. Appl., 5(1962), pp. 329–344.

*[6] L. W. NEUSTADT, *Minimum effort control systems*, this Journal, 1(1962), pp. 16–31.

*[7] B. PAIEWONSKY, P. WOODROW, W. BRUNNER, AND P. HALBERT, *Synthesis of optimal controllers using hybrid analog-digital computers*, Computing Methods in Optimization Problems, A. V. Balakrishnan and L. W. Neustadt, ed., Academic Press, New York, 1964, pp. 285-303.

* References added by translator.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

AN ANALYTIC THEORY OF MODELING FOR A CLASS OF MINIMAL-ENERGY CONTROL SYSTEMS (DISTURBANCE-FREE CASE)*

WALTER J. CULVER†

Summary. A quantitative theory is developed for modeling a class of optimal control systems. A mathematical representation—a model system—is fit to an actual system solely on the basis of the respective optimal performances of the two systems, where performance is defined by a generalized quadratic criterion of the minimum energy, minimal endpoint-error type. The plant to be controlled is assumed to be linear time-varying (at least in the small), and the model is taken to be linear, but constant-coefficient.

Necessary and sufficient conditions are derived for achieving certain pertinent tasks of performance prediction and optimal control, wherein particular attention is paid to the accomplishment of these tasks by computer methods. It is found that the very structure of the plant representation may prohibit some model activities, e.g., if a certain inequality relation is not maintained between the respective dimensions of the state and control vectors.

Finally, the given performance index is used to partition the universe of linear systems into equivalence classes, and the conditions are presented for two systems to be *performance-equivalent*. These are shown to be the necessary and sufficient conditions for the optimal control laws of nonidentical systems to be, in fact, interchangeable in the large.

1. Introduction. Modeling is perhaps the most fundamental aspect of purposeful behavior, being the foundation for virtually all processes of learning, identification, prediction, and control. In this paper we are concerned primarily with the latter two activities, and more specifically, with mathematical models by which real physical processes can be represented for purposes of estimating their future behavior, controlling their future behavior, or both.

We have developed what is believed to be a new approach to the problem of modeling a system which is to be optimally controlled, wherein the model is chosen so as to match the *performances*—not the *responses*—of the modeled and actual systems. If the outputs or states of the system are not contained in the pertinent performance index, then the model determination is *explicitly independent* of such states, an approach which is in sharp contrast to more classical modeling (or identification) tech-

* Received by the editors April 23, 1964, and in revised form July 23, 1964.

† Surface Division, Westinghouse Electric Corporation, P. O. Box 1897, Baltimore, Maryland. Formerly at the Systems Research Center, Case Institute of Technology, Cleveland, Ohio. The research presented here was supported in part by the Office of Naval Research, contract NOnr 1141(12), and the personal support of the author came primarily from the Department of Health, Education and Welfare under the terms of a National Defense Graduate Fellowship.

niques [1], [2], [3]. Note that if the criterion of performance, say L, is *really* meaningful for the problem at hand, then we should not particularly care what the respective states do, so long as the respective performances, in terms of L, are close enough.

At the time of this writing, our accomplished research along these lines falls naturally into two parts.

The first of these parts outlines the fundamental properties of modeling in our context, dealing especially with what a model can and cannot do in an ideal, disturbance-free environment. This, considered as an indication of the upper limit on model performance, is what the present paper is concerned with.

The second part comprises the more physically realistic study of a disturbance-contaminated environment. It depends upon much of the work of the first part, indicates the possible superiority of model-control over control based upon the exact plant equations, and will be submitted for publication as a forthcoming paper [18].

2. The problem formulation.

The plant. In specific terms, our study here is concerned with known systems of time-varying linear differential equations, of which a typical decoupled subsystem might be

$$\frac{d^K x}{dt^K} + \sum_{k=0}^{K-1} p_k(t) \frac{d^k x}{dt^k} = \sum_{k=0}^{K-1} q_k(t) \frac{d^k m}{dt^k} . \tag{1}$$

We will call the collection of such subsystems, together with their natural interactions, the "plant."[1]

Employing the so-called state variable transformations [4], [5], we can put the plant equations in the vector-matrix form

$$\dot{\mathbf{x}} = P(t)\mathbf{x} + Q(t)\mathbf{m}, \tag{2}$$

where $\mathbf{x}$ is an n-dimensional state vector, $\mathbf{m}$ is a p-dimensional manipulated or control vector, and $P(t)$ and $Q(t)$ are matrices of obvious dimensions with typical elements $p_{ij}(t)$ and $q_{ij}(t)$ in ith rows and jth columns.

For purposes of generality in application, we do not assume these matrices to be continuous, for if (2) is obtained from a perturbational analysis of a nonlinear system, a denumerable number of discontinuities may arise.[2] Rather, we require only that these matrices be Riemann integrable and bounded in norm [7, p. 97, prob.1] which, when applied to $\mathbf{m}$ as well, insures that every solution vector $\mathbf{x}(t)$ be *absolutely continuous* [8]. We need these

[1] The plant plus the proper equations of control we will call the control system, or just "the system."

[2] For an extended discussion of this and other general points, see [6].

properties only on the closed interval $[t_0, T]$, which we take to be the time interval over which the system is to operate, and, of course, (2) has meaning here only when the derivative $\dot{\mathbf{x}}(t)$ exists. Note finally that even though the states in this representation are *absolutely continuous*, the actual physical states need not be, as a perusal of the state-variable transformations quickly indicates.

The model. In keeping with the introductory remarks, we now postulate a constant-coefficient model

$$\dot{\bar{\mathbf{x}}} = A\bar{\mathbf{x}} + B\,\bar{\mathbf{m}}, \tag{3}$$

where the barred variables have the same dimensions as the original ones, A and B being constant matrices yet to be determined.

The justification for such a model rests upon the ever-present need to simplify the computational aspects of on-line[3] prediction and control. For even though the plant is taken to be linear, still it is time-varying, a fact which almost surely eliminates the possibility of completely closed-form calculations. In fact, as we will see, the so-called fundamental matrix of the plant has a prominent place in the requisite mathematics. This matrix is an array of the general homogeneous solutions to (2), and has the following important properties:

$$\dot{\Phi}(t, s) = P(t)\Phi(t, s), \qquad \textit{almost everywhere}, \tag{4a}$$

$$\Phi(t, t) = I, \tag{4b}$$

$$\Phi(t, u)\Phi(u, s) = \Phi(t, s), \tag{4c}$$

$$\det \Phi(t, s) \neq 0; \tag{4d}$$

where (4a) has no meaning on the set of measure zero for which $P(t)$ is not defined, I is the identity matrix, and det ( ) denotes determinant.

Unfortunately, except in the very special cases [9] where $P(t)$ commutes with its integral $\int_s^t P(u)\,du$, $\Phi(t, s)$ cannot be written in closed form. Rather, as in [11], it must be expressed as the iterate solution to a matrix Volterra equation of the second kind, which must be tabulated for discrete increments of time and read out as required.

In replacing the plant with a time-invariant model, we would be gaining simplicity in many respects, but especially in that the model fundamental matrix, say $\bar{\Phi}(t, s)$, can be written immediately as

$$\bar{\Phi}(t, s) = e^{A(t-s)}. \tag{5}$$

[3] That is, while the process is operating, as opposed to pre-programmed operation. Here, elapsed computation time can be a vital factor, and in aerospace vehicles, could conceivably determine weight and power to be allotted to computing devices.

This can be expressed either as an infinite series in powers of $(t - s)$, or else as a finite series of n terms in powers of $(t - s)$ and $e^{\lambda_i(t-s)}$, where the λ_i are the eigenvalues of the matrix A. (We refer the reader to [10, Chap. V].) At any rate, the storage requirements for digital implementation are now reduced to nominal magnitudes, since functions such as $(t - s)$ and $e^{\lambda_i(t-s)}$ are very simply generated in an on-line fashion.

The criterion. In order to obtain concrete analytical results, now, we must be prepared to specify our system criterion of performance rather precisely. From the viewpoints of mathematical elegance and practical significance, the following quadratic "cost" or loss function is most suitable to our work:

$$(6) \qquad L_k = \int_{t_k}^{T} \| \mathbf{m}(t) \|_{M(t)}^2 \, dt + \| \mathbf{x}(T) - \mathbf{x}_d \|_X^2,$$

where $M(t)$ and X are $p \times p$ positive definite and $n \times n$ nonnegative definite matrices respectively, both symmetric, $M(t)$ being bounded and integrable in the same sense as the plant coefficient matrices; $\mathbf{x}_d$ is a desired-endpoint vector; $\| \mathbf{y} \|_Y = (\mathbf{y}^T Y \mathbf{y})^{1/2}$, the weighted Euclidean norm of any vector $\mathbf{y}$, $(\quad)^T$ denoting transpose; t_k is a point in time lying between t_0 and T, at which the state of the system is checked and at which certain calculations of optimization or prediction are to be carried out.

The minimization of this functional can be accomplished in a number of ways, such as, for example [6], via the classical formulation of Bolza in the calculus of variations. The optimal values for $\mathbf{m}$ and $\mathbf{x}(T)$ we denote by $\mathbf{m}^*$ and $\mathbf{x}^*(T)$, and these put into (6) produce the least cost L_k^*. Moreover, as is commonly the case, the optimal control algorithm of "law" provides $\mathbf{m}^*(t)$ as a function of state at the sampling times only, i.e., as a function of $\mathbf{x}(t_k)$.

Thus we can depict the function of the optimally controlled system by Fig. 1(a), and the next definition follows naturally.

Definition 1(a). A system optimized via the "exact" plant equations (2) under the (possibly false) assumption that the environment is disturbance-free, we will denote as an *ideal control system*, or just an *ideal system.* The associated cost

$$(7) \qquad L_k^* = \int_{t_k}^{T} \| \mathbf{m}^*(t) \|_{M(t)}^2 \, dt + \| \mathbf{x}^*(T) + \mathbf{x}_d \|_X^2$$

we will call the *ideal cost*, or *ideal performance.*

The same procedure can be carried through for the set of model equations and for a cost functional, say $\bar{L}_k^*$, written in terms of the barred model variables. Then Fig. 1(b) comes to the fore, and the extremization is performed as though the actual plant never existed.

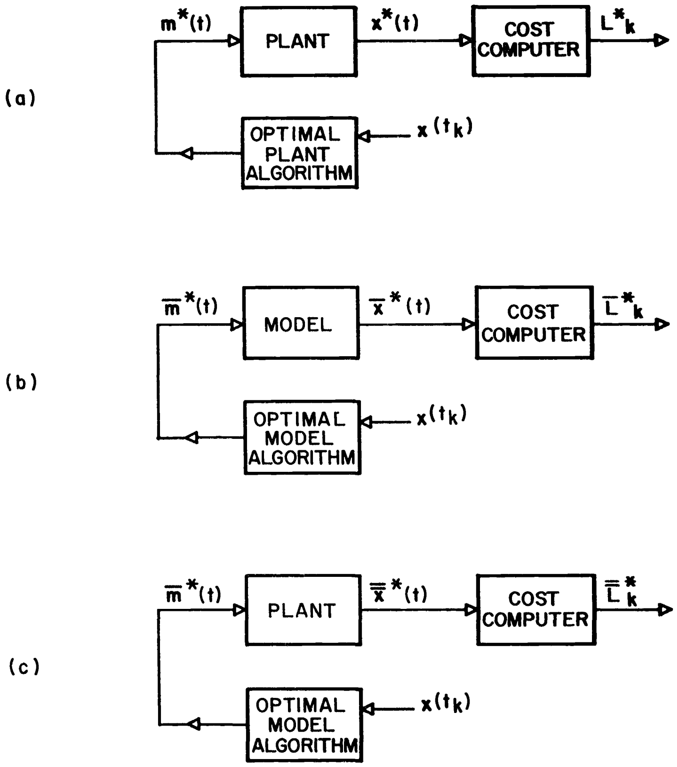


FIG. 1

DEFINITION 1(b). A strictly constant-coefficient model, optimized under the assumption that the environment is disturbance-free, we will denote as an *ideal model system*. The associated cost

$$(8) \qquad \bar{L}_k{}^* = \int_{t_k}^{T} \| \bar{\mathbf{m}}^*(t) \|^2_{M(t)}\, dt + \| \bar{\mathbf{x}}^*(T) - \mathbf{x}_d \|^2_X$$

will be known as the *estimated system cost* or *estimated performance.*

This latter term arises because if we were to use the model to predict the future cost of control, the computations would necessarily be performed according to the schematic of Fig. 1(b).

If we now examine part (c) of Fig. 1, we see the last plant-model configuration that will be of interest to us here. In essence, the plant is controlled *as though it were the model*: Namely, we force the plant with $\bar{\mathbf{m}}^*(t)$ calculated on the basis of the model equations, rather than with $\mathbf{m}^*(t)$, calculated on the basis of the actual plant equations. The plant response under these conditions we denote as $\bar{\bar{\mathbf{x}}}^*(t)$; it is not optimal, but rather the best we can do without employing the plant equations.

DEFINITION 1(c). A system consisting of the plant (2) driven by the model control law $\bar{\mathbf{m}}^*(t)$, which is obtained by optimizing (8) under the constraint of (3) and under the (possibly false) assumption that the environment is disturbance-free, we will denote as an *ideal model-controlled system*. The associated cost

$$(9) \qquad \bar{\bar{L}}_k^* = \int_{t_k}^{T} \| \bar{\mathbf{m}}^*(t) \|^2_{M(t)}\, dt + \| \bar{\bar{\mathbf{x}}}^*(T) - \mathbf{x}_d \|^2_X$$

will be called the *ideal model-controlled cost* or *performance.*

Modeling objectives. So far we have been discussing the model in a rather abstract light. In any practical circumstance, of course, the coefficient matrices A and B, or some equivalent parameters, must be evaluated. In terms of our stated interests, this determination might be performed in order that the model should accomplish one or more of the following three tasks:

Task 1. Of considerable importance can be the estimation of the "upper limit" on system performance: that is, the *ideal performance* of the system. In the context of an aerospace mission, for example, we can visualize the necessity of determining whether or not a projected mission can be accomplished with a given amount of available fuel. (Delivered fuel can be related to the integral part of the criterion.) If the simpler model calculations can precisely determine that a certain trajectory endpoint cannot be attained, even wi th a control based upon the "exact" plant equations in an undisturbed environment, then the mission would properly be aborted in favor of another objective which might still be achievable. In terms of (7) and (8), the prediction without error of the *ideal performance* entails setting

$$(10) \qquad L_k^* - \bar{L}_k^* = 0.$$

In terms of Fig. 1, we would be asking (b) to predict the performance of (a) exactly.

Task 2. If we intend to use our model to control the plant, then it is of obvious importance to estimate the performance of this model-control. It is of interest, then, to determine a model that can predict without error *its own* optimal control of the plant, at least in an ideal environment. This involves setting

$$(11) \qquad \bar{\bar{L}}_k^* - \bar{L}_k^* = 0,$$

whereby the configuration of Fig. 1(b) would be able to estimate exactly the performance of the configuration of Fig. 1(c).

Task 3. Both of the above tasks are concerned primarily with prediction, i.e., prediction of *ideal performance* or prediction of *ideal model-controlled performance*. In neither case is the associated model required to do a good job in controlling the plant. Yet it is clear that if we intend to use our model in a control capacity, then we want it to operate as efficiently as possible relative to the ideal control. This can be restated mathematically via (7) and (9):

$$\bar{\bar{L}}_k{}^* - L_k{}^* = \min, \tag{12}$$

which says, in reference to Fig. 1, that we would be choosing a model wherein (c) operates at as low a cost as possible relative to (a).

3. Basic restrictions and assumptions. In choosing our model analytically, there are certain constraints which must be met in order that the model be of maximal utility and practicality.

For example, we do not want the model coefficients to depend upon the measurements of the "present" state of the system, for if these measurements are in error we do not want our model to be any less valid, even though it must necessarily propagate these errors in its calculations.

In addition, as we model plants which are more and more nearly constant-coefficient, the general equations of modeling should yield models which are more and more like an exact representation of the plant, until finally, in the limit, a constant-coefficient plant should give rise to a model which matches it in every way.

In summary then, what we require of our model and assume of our environment is the following:

A. We assume for this paper that the environment is disturbance-free.

B. We assume, except as will be noted, that the weighting matrix X in (6) is nonnegative definite, but nonzero.

C. We require that the model have coefficient matrices which are state-independent: i.e., they are not functions of the state variables, past, present, or future. (These matrices may, however, be functions of the sampling time t_k .)

D. We require that if the plant becomes constant-coefficient over some interval $[t_k , T]$, then our modeling equations should yield a model which matches the plant exactly, and in every respect, over that interval.

4. The mathematical preliminaries. In the work which is to follow, certain sets of definitions prevail. The first is a generalization of controllability theory in Kalman's sense [12], [13], which is involved in many aspects of the modeling problem for minimum-energy systems.

Definition 2 (Controllability Matrices).

(a) The nonnegative definite symmetric matrix

$$\text{(13a)} \qquad H(T, t_k) = \int_{t_k}^{T} \Phi(T, s)Q(s)M^{-1}(s)Q^T(s)\Phi^T(T, s)\, ds$$

will be known as the *plant controllability matrix at time* t_k .

(b) The nonnegative definite symmetric matrix

$$\text{(13b)} \qquad \bar{H}(T, t_k) = \int_{t_k}^{T} \bar{\Phi}(T, s)BM^{-1}(s)B^T\bar{\Phi}^T(T, s)\, ds$$

will be known as the *model controllability matrix at time* t_k .

(c)

$$\text{(13c)} \qquad \bar{\bar{H}}(T, t_k) = \int_{t_k}^{T} \Phi(T, s)Q(s)M^{-1}(s)B^T\bar{\Phi}^T(T, s)\, ds$$

will be known as the *cross-controllability matrix at time* t_k .

The last part of the definition, (c), is denoted as such because it arises naturally in the model-control of the plant, involving both the plant variables (to the left of M^{-1}) and the model variables. Unlike H and $\bar{H}$ above, it is inherently neither nonnegative definite nor symmetric.

Since the concept of controllability—especially in the form of these matrices—is so fundamental to our work, we will introduce the following definition.

Definition 3 (See [12]). A state x_i is said to *controllable* at time t_k if there exists a control function $\mathbf{m}(t)$, depending on $x_i(t_k)$ and defined over the finite closed interval $[t_k , T]$, such that at time T,

$$x_i(T, x_i(t_k), \mathbf{m}) = x_{i,d} ,$$

where $x_{i,d}$ is any desired terminal value for that state. If this is true of *every* state $x_i(t_k)$, $i = 1, 2, \cdots , n$, then the plant is *completely controllable at time* t_k ; if this is true for every t_k , then the plant is *completely controllable*.

It is possible to generalize slightly a proposition of Kalman's [12, p. 107] and thus state the following assertion (the truth of which follows at once from Kalman's proof):

Assertion. *Given any symmetric positive definite matrix, say* $M(t)$, *then the plant* (2) *is completely controllable at time* t_k *if and only if the symmetric nonnegative definite matrix*

$$\text{(14)} \qquad W(t_k, T) = \int_{t_k}^{T} \Phi(t_k, s)Q(s)M^{-1}(s)Q^T(s)\Phi^T(t_k, s)\, ds$$

is positive definite.

As far as linear systems are concerned, a lack of complete controllability at any time t_k implies that the system is improperly formulated at that time, either because of redundancy in the state representation or because of inconsistencies in it. In order for our control algorithm to make sense, then, we must require the system to be completely controllable at every sampling instant: i.e., $W(t_k, T)$ must be positive definite for each t_k.

In comparing (13a) to (14), we observe the identity

$$H(T, t_k) = \Phi(T, t_k)W(t_k, T)\Phi^T(T, t_k), \tag{15}$$

which, since $\det \Phi(t, s) \neq 0$, implies that H is positive definite *if* and *only if* W is[4].

Birta [5], Kalman [12], [13], and Friedland [14] all use the W matrix in their work; for our purposes, however, the H matrix is more convenient, and it still has the controllability connotation in its definiteness property.

Since similar arguments apply to the model, we naturally require $\bar{H}$ to be positive definite as well.

DEFINITION 4. The unforced error, i.e., the difference between the desired endpoint $\mathbf{x}_d$ and the homogeneous response of the system at $t = T$, we denote as

$$\mathbf{u}(T, t_k) = \Phi(T, t_k)\mathbf{x}(t_k) - \mathbf{x}_d \tag{16a}$$

for the ideal system; analogously we define

$$\bar{\mathbf{u}}(T, t_k) = \bar{\Phi}(T, t_k)\mathbf{x}(t_k) - \mathbf{x}_d \tag{16b}$$

for the model system. Where there is no likelihood of confusion, we denote these simply as $\mathbf{u}$ and $\bar{\mathbf{u}}$ respectively.

To simplify our future notational problems further, we introduce the matrices

$$U = X(I + HX)^{-1}, \tag{17a}$$

$$\bar{U} = X(I + \bar{H}X)^{-1}, \tag{17b}$$

where H and $\bar{H}$ are short forms for (13a) and (13b), respectively. The important properties of the above we give in lemma form.

LEMMA 1. *Given that X is a symmetric, nonnegative definite matrix, and H is a symmetric, positive definite matrix, then*

(a) *$U = X(I + HX)^{-1}$ exists, is symmetric, and is nonnegative definite; moreover,*

(b) *U is positive definite or positive semi-definite with X.*

[4] This point depends on the ability to represent a symmetric positive definite matrix as K^TK, where K is some nonsingular matrix. A formal proof is given for general congruence transformations on symmetric matrices in [6, Appendix C].

Proof. The proof follows [6, Appendix C] once we note that[5]

$$U = (I + HX)^{-T}(X + XHX)(I + HX)^{-1}, \tag{18}$$

and then apply footnote 4 of this paper and a well-known theorem [15, p. 115].

Precisely the same sort of proof could, of course, be carried out for $\bar{U}$, with $\bar{H}$ replacing H everywhere.

5. The primary equations. The accomplishment of one or more of the Tasks 1 to 3 clearly involves the evaluation of the optimal performance indices (7), (8), and (9). These, in turn, contain the five quantities,

$$\mathbf{m}^*(t) = -M^{-1}(t)Q^T(t)\Phi^T(T, t)X(I + HX)^{-1}\mathbf{u}, \tag{19a}$$

$$\bar{\mathbf{m}}^*(t) = -M^{-1}(t)B^T\bar{\Phi}^T(T, t)X(I + \bar{H}X)^{-1}\bar{\mathbf{u}}, \tag{19b}$$

$$\mathbf{x}^*(T) = (I + HX)^{-1}\mathbf{u} + \mathbf{x}_d , \tag{20a}$$

$$\bar{\mathbf{x}}^*(T) = (I + \bar{H}X)^{-1}\bar{\mathbf{u}} + \mathbf{x}_d , \tag{20b}$$

$$\bar{\bar{\mathbf{x}}}^*(T) = \Phi(T, t_k)\mathbf{x}(t_k) - \bar{\bar{H}}X(I + \bar{H}X)^{-1}\bar{\mathbf{u}}, \tag{21}$$

which are derived from an application of the variational calculus to the requisite optimizations [6, Appendix A].

For the evaluation of the three performance functionals we can now begin by substituting (19a) into the integral part of (7) and (20a) into the nonintegral part. This produces

$$L_k^* = \| \mathbf{u} \|_{UHU}^2 + \| \mathbf{u} \|_{(I+HX)^{-T}X(I+HX)^{-1}}^2 ,$$

which, from the definition of the norm given in connection with (6), reduces to

$$L_k^* = \| \mathbf{u} \|_U^2. \tag{22}$$

With a similar line of reasonng, we can find $\bar{L}_k^*$ from (8), (19b), and (20b):

$$\bar{L}_k^* = \| \bar{\mathbf{u}} \|_{\bar{U}}^2 . \tag{23}$$

Finally we can evaluate (9) explicitly from the substitutions of (19b) and (21), from which there evolves

$$\bar{\bar{L}}_k^* = \| \bar{\mathbf{u}} \|_{\bar{U}\bar{\bar{H}}\bar{U}}^2 + \| \mathbf{u} - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}} \|_X^2 . \tag{24}$$

Now, Task 1 deals with the difference

$$L_k^* - \bar{L}_k^* = \| \mathbf{u} \|_U^2 - \| \bar{\mathbf{u}} \|_{\bar{U}}^2 , \tag{25}$$

as (22) and (23) attest.

[5] The notation $(\ )^{-T}$ signifies inverse transpose.

Task 2 deals with the difference

$$\bar{\bar{L}}_k{}^* - \bar{L}_k{}^* = \| \bar{\mathbf{u}} \|^2_{\bar{U}\bar{H}\bar{U}-\bar{U}} - \| \mathbf{u} - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}} \|^2_X , \tag{26}$$

as (23) and (24) attest.

Finally, Task 3 deals with the difference

$$\bar{\bar{L}}_k{}^* - L_k{}^* = \| \bar{\mathbf{u}} \|^2_{\bar{U}\bar{H}\bar{U}} + \| \mathbf{u} - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}} \|^2_X - \| \mathbf{u} \|^2_U , \tag{27}$$

as (22) and (24) attest.

This puts us in a position to present our main development.

6. Model prediction of ideal performance. In this section we are concerned with the attainment of Task 1, or synonymously, with obtaining

$$L_k{}^* - \bar{L}_k{}^* = 0. \tag{(10)}$$

Thereby we insure that given the "initial" state $\mathbf{x}(t_k)$ of the system, the model can predict exactly the *ideal* cost of controlling the plant with the most accurate equations available.

THEOREM 1. *Given that X is a positive definite matrix, then the necessary and sufficient conditions that Task* 1 *be accomplished are*

$$\bar{\Phi}(T, t_k) = \Phi(T, t_k), \tag{29}$$

$$\bar{H}(T, t_k) = H(T, t_k), \tag{30}$$

Proof. Sufficiency: To verify that (29) and (30) *do* actually satisfy (10), we need only set the right side of (25) to zero. Writing the variables in terms of their definitions, i.e., (16) and (17), we find that

$$\begin{aligned} 0 = {} & \| \Phi(T, t_k)\mathbf{x}(t_k) - \mathbf{x}_d \|^2_{X(I+HX)^{-1}} \\ & - \| \bar{\Phi}(T, t_k)\mathbf{x}(t_k) - \mathbf{x}_d \|^2_{X(I+\bar{H}X)^{-1}} . \end{aligned} \tag{31}$$

Clearly, the substitution of (29) and (30) into the right side of this expression causes it to vanish as required, thereby establishing the sufficiency.

Necessity: To demonstrate that there are *no other* solutions than (29) and (30), let us carry through this part of the proof in a local, or in-the-small, sense. What can be shown to be necessary in-the-small is certainly necessary in-the-large, and therefore necessary in general.

Thus let us seek the necessary conditions for (10) to hold in an arbitrarily small (but nonvanishing) neighborhood of a pertinent "operating point" $\mathbf{u}^0$, where we define $\mathbf{u}$ to be

$$\mathbf{u} = \mathbf{u}^0 + \Delta\mathbf{u}, \tag{32}$$

with $\| \Delta\mathbf{u} \| < \alpha$, α being a positive scalar constant sufficiently small. Similarly, the model has an operating point $\bar{\mathbf{u}}^0$ and a deviation $\Delta\bar{\mathbf{u}}$ such that

$$\bar{\mathbf{u}} = \bar{\mathbf{u}}^0 + \Delta\bar{\mathbf{u}}, \tag{33}$$

where the deviation is bounded in norm as $\Delta\mathbf{u}$.

Now, deviations from the nominal can arise in two ways: due to errors $\Delta\mathbf{x}(t_k)$ in our knowledge of the initial state, and due to changes $\Delta\mathbf{x}_d$ which occur in the target location. Obviously, then,

$$\Delta\mathbf{u} = \Phi(T, t_k)\Delta\mathbf{x}(t_k) - \Delta\mathbf{x}_d$$

and

$$\Delta\bar{\mathbf{u}} = \bar{\Phi}(T, t_k)\Delta\mathbf{x}(t_k) - \Delta\mathbf{x}_d ,$$

so that when the initial conditions are known with certainty ($\Delta\mathbf{x}(t_k) = 0$), we have

$$(34) \qquad \Delta\mathbf{u} = \Delta\bar{\mathbf{u}}.$$

In this special case, therefore, we can put (32) and (33) into (25), set the result to zero, and obtain

$$(35) \quad 0 = \| \mathbf{u}^0 \|_U^2 - \| \bar{\mathbf{u}}^0 \|_{\bar{U}}^2 + 2\Delta\mathbf{u}^T(U\mathbf{u}^0 - \bar{U}\bar{\mathbf{u}}^0) + \| \Delta\mathbf{u} \|_{U-\bar{U}}^2 ,$$

an expression which is to hold for *all* $\Delta\mathbf{u}$ small enough in norm. But such a requirement can be met only if the coefficients of the term in each "power" of $\Delta\mathbf{u}$ are separately zero. That is, it is required that

$$(36a) \qquad 0 = \| \mathbf{u}^0 \|_U^2 - \| \bar{\mathbf{u}}^0 \|_{\bar{U}}^2,$$

$$(36b) \qquad 0 = 2\Delta\mathbf{u}^T(U\mathbf{u}^0 - \bar{U}\bar{\mathbf{u}}^0),$$

$$(36c) \qquad 0 = \| \Delta\mathbf{u} \|_{U-\bar{U}}^2 .$$

Let us consider the last two of these expressions. Since $\Delta\mathbf{u}$ is arbitrary (although small) and since (as Lemma 1 implies) $U - \bar{U}$ is symmetric, the necessity clearly follows that

$$(37a) \qquad 0 = U\mathbf{u}^0 - \bar{U}\bar{\mathbf{u}}^0$$

and

$$(37b) \qquad 0 = U - \bar{U}$$

hold simultaneously. But according to (17), the second of the above expressions is simply

$$0 = X[(I + HX)^{-1} - (I + \bar{H}X)^{-1}].$$

Since X is positive definite (nonsingular), it follows that the bracketed term must vanish, from which we obtain the necessity of (30).

To show the necessity of (29), we put (37b) into (37a) and draw upon Lemma 1(b) to show that

$$(38) \qquad 0 = \mathbf{u}^0 - \bar{\mathbf{u}}^0.$$

If we refer to (16) and denote the nominal state-vector quantities there as $\mathbf{x}^0(t_k)$ and $\mathbf{x}_d^0$, it is clear that (38) reduces to the condition

$$(39) \qquad 0 = [\Phi(T, t_k) - \bar{\Phi}(T, t_k)]\mathbf{x}^0(t_k).$$

Although it is superfically true that (39) can be achieved in the small without the bracketed term being the null matrix—e.g., by having it singular with eigenvector $\mathbf{x}^0(t_k)$ corresponding to a zero eigenvalue—, still such a solution does not meet Requirement D of §3. For if the plant is constant-coefficient (or nearly so), then we want to obtain a model which is effective over many or all of the future checkpoints t_{k+1}, t_{k+2}, etc. That is, we expect to satisfy (or nearly satisfy) (39) when t_k is replaced by t_{k+1}, etc., without having to redetermine our model. But if the bracketed term in (39) is not set to zero, then a rapidly changing $\mathbf{m}(t)$ on $[t_k, t_{k+1}]$ may force $\mathbf{x}^0(t_{k+1})$ to be very much different than $\mathbf{x}^0(t_k)$, and this will result in

$$\| [\Phi(T, t_{k+1}) - \bar{\Phi}(T, t_{k+1})]\mathbf{x}^0(t_{k+1}) \| \gg 0,$$

even though the plant is constant-coefficient (or nearly so).

Thus it is necessary that (29) be satisfied, and the proof of the theorem is complete.

Example 1. Consider the plant

$$(40) \qquad \dot{\mathbf{x}} = \begin{bmatrix} -t & 0 \\ 0 & -2t \end{bmatrix} \mathbf{x} + \begin{bmatrix} te^{-t^2/2} & 0 \\ 0 & e^{-t^2} \end{bmatrix} \mathbf{m}$$

and the model

$$(40') \qquad \dot{\bar{\mathbf{x}}} = A\bar{\mathbf{x}} + B\bar{\mathbf{m}},$$

where A and B are 2×2 matrices, yet to be determined. Since the coefficient matrix of the $\mathbf{x}$ vector commutes with its integral we immediately have

$$(41) \qquad \Phi(t, s) = \begin{bmatrix} e^{-(t^2-s^2)/2} & 0 \\ 0 & e^{-(t^2-s^2)} \end{bmatrix},$$

whereas the fundamental matrix for (40′) is simply

$$(41') \qquad \bar{\Phi}(t, s) = e^{A(t-s)}.$$

Setting $\bar{\Phi}(T, t_k) = \Phi(T, t_k)$, we find that A must be diagonal. More specifically,

$$(42) \qquad A = \begin{bmatrix} -k/2 & 0 \\ 0 & -k \end{bmatrix}$$

and

$$(43) \qquad \bar{\Phi}(t, s) = \begin{bmatrix} e^{-k(t-s)/2} & 0 \\ 0 & e^{-k(t-s)} \end{bmatrix}.$$

Here we have employed or will employ the notation

$$(44) \qquad k = T + t_k, \quad k' = T - t_k, \quad k'' = T^2 + Tt_k + t_k^2.$$

To find the B matrix, we now have to solve (30). If, for simplicity, we take $M(t) = I$, relation (13a) for our example becomes explicitly

$$(45) \qquad H(T, t_k) = \begin{bmatrix} \frac{1}{3}k'k''e^{-T^2} & 0 \\ 0 & k'e^{-2T^2} \end{bmatrix}.$$

For (13b), now, we obtain

$$(46) \quad \bar{H}(T, t_k) = \begin{bmatrix} (b_{11}^2 + b_{12}^2)\dfrac{1 - e^{-kk'}}{k} & 2(b_{11}b_{21} + b_{12}b_{22})\dfrac{1 - e^{3kk'/2}}{3k} \\ \maltese & (b_{21}^2 + b_{22}^2)\dfrac{1 - e^{-2kk'}}{2k} \end{bmatrix},$$

where the mark in the lower left-hand corner of the latter matrix indicates that the entry there is the same as the entry in the upper right-hand corner; i.e., the matrix is symmetric.

Equating (46) to (45) and solving for the b_{ij}, we find that one possible set of solutions is

$$(47) \qquad \begin{aligned} b_{11} &= \pm\left(\frac{kk'k''e^{-T^2}}{3(1 - e^{-kk'})}\right)^{1/2}, \\ b_{12} &= b_{21} = 0, \\ b_{22} &= \pm\left(\frac{2kk'e^{-2T^2}}{1 - e^{2kk'}}\right)^{1/2}. \end{aligned}$$

Equations (42) and (47) suffice to specify the model so that $L_k^* - \bar{L}_k^* = 0$. Note, however, that in the predictive capacity for which the model is designed, b_{11} and b_{22} can be either (+) or (−). Thus, it is a questionable procedure to try to employ this model in a control capacity as well, for it is possible to satisfy Task 1 with a model whose *direction of control* is opposite that of the plant. In other words, an increase of $\mathbf{m}$ in (40) causes an *increase* in $\mathbf{x}$, whereas, if we pick the (−) signs in (47), an increase of $\bar{\mathbf{m}}$ in (40′) causes a decrease in $\bar{\mathbf{x}}$. So ends the example.

Having obtained a solution for a very simple example, we might wish to examine the solutions for the general case. Indeed, whether the equation

$$((29)) \qquad e^{A(T-t_k)} = \Phi(T, t_k)$$

is satisfied by a real matrix A is a fact not immediately obvious, for it is well known [10] that certain exponential matrix equations have no real solutions, whereas others have a continuum of such solutions.

What we must introduce to resolve these questions of existence and uniqueness are, in fact, certain complicated aspects of the theory of matrices and of linear transformations. Rather than cloud the framework of our modeling theory with these more abstract mathematical details, we will delay their presentation until a forthcoming paper. In the interim, we still have at our disposal sufficient power to develop many of the fundamental results of this study.

THEOREM 2 (*Fundamental Theorem of Structure*). *If there exists a unique real matrix solution A to (29) and if the plant (2) has a state vector $\mathbf{x}$ of dimension n and a control vector $\mathbf{m}$ of dimension p, then in $(\mathbf{x}, \mathbf{m})$ space Task 1 can be achieved for every pair of coefficient matrices $P(t)$ and $Q(t)$ only if the above vector dimensions are such that*

$$p \geqq \frac{n+1}{2}. \tag{48}$$

Proof. By hypothesis, the first condition of Theorem 1 is met. To obtain the second condition, (30), we would normally be able to choose both the elements of A and the elements of B (which appear on the left of (30)) is order to force the requisite equality. However, if A is uniquely determined, then we have at our disposal only the elements of B.

Both $\bar{H}$ and H are symmetric, nonnegative definite matrices, which, from controllability considerations, we require to be positive definite. Thus, setting $\bar{H} = H$ involves the solution of $n(n + 1)/2$ distinct simultaneous equations. Clearly, then, for arbitrary matrices H, we require a *minimum* of $n(n + 1)/2$ free parameters, which, as noted above, must be contributed by the B matrix of the model. Since this matrix is of dimension $n \times p$, where p is the dimension of the control vector, B contains np elements. Thus we require. $np \geqq n(n + 1)/2$, and the theorem is proved.

Remark A. Note that the above theorem presents a necessary condition only if we seek to achieve Task 1 for *all* $P(t)$ and $Q(t)$ matrices which are Riemann integrable and bounded in norm. For although the integrals H and $\bar{H}$ are positive definite, and thus of rank n, their respective integrands are at most of rank p, which is nearly always less than n. Thus, for that limited class of $P(t)$ and $Q(t)$ matrices which permit the integrand identity

$$\bar{\Phi}(T, t)BM^{-1}(t)B^T\bar{\Phi}^T(T, t) \equiv \Phi(T, t)Q(t)M^{-1}(t)Q^T(t)\Phi^T(T, t) \tag{49}$$

for all t on $[t_k, T]$, relation (48) no longer presents a necessary condition. If the *plant* is time-invariant, (49) can always be met.

7. Model estimation of model-control performance. In this section we are concerned with the satisfaction of Task 2. In short, it is desired to compute a model such that, given the "present" state $\mathbf{x}(t_k)$, the model can predict exactly the cost that will be incurred when the model-control law is applied to the plant.

Definition 5. The matrix difference between the model controllability matrix $\bar{H}(T, t_k)$ and the cross controllability matrix $\bar{\bar{H}}(T, t_k)$ we will denote as

$$D = \bar{\bar{H}}(T, t_k) - \bar{H}(T, t_k). \tag{50}$$

Theorem 3. *Given that X is a positive definite matrix, then the necessary and sufficient conditions that Task 2 be accomplished, i.e., that*

$$\bar{\bar{L}}_k{}^* - \bar{L}_k{}^* = 0, \tag*{((11))}$$

are

$$\bar{\Phi}(T, t_k) = \Phi(T, t_k), \tag*{((29))}$$

$$(I - DX)^T X(I - DX) = X. \tag{51}$$

Proof. Sufficiency: To verify that (29) and (51) do actually satisfy (11), we need only take (26) and set it to zero:

$$0 = \| \bar{\mathbf{u}} \|^2_{\bar{U}\bar{H}\bar{U}-\bar{U}} + \| \mathbf{u} - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}} \|^2_X . \tag{52}$$

Note that the weighting matrix of the first term is

$$\bar{U}[\bar{H}\bar{U} - I] = -(I + \bar{H}X)^{-T}X(I + \bar{H}X)^{-1} \tag{53}$$

Moreover, an examination of (16a) and (16b) indicates that (29) provides us with the result

$$\bar{\mathbf{u}} = \mathbf{u}, \tag{54}$$

so that (52) becomes

$$0 = \| \mathbf{u} \|^2_K , \tag{55}$$

where

$$K = (I - \bar{\bar{H}}\bar{U})^T X(I - \bar{\bar{H}}\bar{U}) - (I + \bar{H}X)^{-T}X(I + \bar{H}X)^{-1} \tag{56}$$

$$= (I + \bar{H}X)^{-T}\{(I - DX)^T X(I - DX) - X\}(I + \bar{H}X)^{-1}. \tag{57}$$

Since (51) causes the curly-bracketed term in (57) to vanish, it follows that (11) is satisfied and the sufficiency is established.

Necessity: To see that no other pair of solutions can obtain (11), let us attack the problem from its *local* aspect, as in the proof of Theorem 1. That is, in the small, (52) becomes

$$\begin{aligned} 0 = {} & \| \bar{\mathbf{u}}^0 \|^2_{\bar{U}\bar{H}\bar{U}-\bar{U}} + \| \mathbf{u}^0 - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}}^0 \|^2_X + \| \Delta\mathbf{u} \|^2_{\bar{U}\bar{H}\bar{U}-\bar{U}} \\ & + \| \Delta\mathbf{u} \|^2_{(\bar{I}-\bar{\bar{H}}\bar{U})^T X (I-\bar{\bar{H}}\bar{U})} + 2\Delta\mathbf{u}^T(\bar{U}\bar{H}\bar{U} - \bar{U})\bar{\mathbf{u}}^0 \\ & + 2\Delta\mathbf{u}^T(I - \bar{\bar{H}}\bar{U})^T X(\mathbf{u}^0 - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}}^0). \end{aligned} \tag{58}$$

Since if (58) holds at all, it must hold when $\Delta\mathbf{u} = 0$, we can remove the first two terms in the equation and rewrite the result as

$$0 = \| \Delta\mathbf{u} \|^2_{E-F^TXF} + 2\Delta\mathbf{u}^T[E\bar{\mathbf{u}}^0 - F^TX(\mathbf{u}^0 - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}}^0)], \tag{58$'$}$$

where for simplicity of notation we have taken

$$E \overset{\Delta}{=} \bar{U} - \bar{U}\bar{H}\bar{U}, \tag{59a}$$

$$F \overset{\Delta}{=} I - \bar{\bar{H}}\bar{U}. \tag{59b}$$

As argued in the proof of Theorem 1, the only way for (58′) to be satisfied for $\Delta\mathbf{u}$ small enough is for $E - F^TXF$ to be zero and for the coefficient of the second term to vanish. That is, it is necessary that

$$0 = E - F^TXF, \tag{60}$$

$$0 = E\bar{\mathbf{u}}^0 - F^TX(\mathbf{u}^0 - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}}^0). \tag{61}$$

Using (60) to replace E with F^TXF, we now find that (61) becomes

$$0 = F^TX[F\bar{\mathbf{u}}^0 - (\mathbf{u}^0 - \bar{\bar{H}}\bar{U}\bar{\mathbf{u}}^0)]. \tag{62}$$

Since E is clearly positive definite, it must be that F^TX is nonsingular, and therefore the bracketed term in (62) must vanish. Via (59b), then,

$$\bar{\mathbf{u}}^0 = \mathbf{u}^0,$$

and the necessity of (29) follows as in Theorem 1. Moreover, the right side of (60) is nothing but the K matrix defined by (56), which—as is clear from (57)—can be zero *if* and *only if*

$$(I - DX)^TX(I - DX) - X = 0.$$

The theorem is thus completely proved.

THEOREM 4. *Condition* (51) *of the above theorem is identical to the pair of conditions*

$$D^TXD = D + D^T \tag{63a}$$

and

$$D^TXD = DXD^T. \tag{63b}$$

Proof. (a) Expand (51) into

$$X - XD^TX - XDX + XD^TXDX = X,$$

cancel the free X's and pre- and post-multiply by X^{-1} (note that by the hypothesis of the theorem, X is positive definite). The first expression, (63a), then follows.

(b) Pre-multiply (51) by $(I - DX)^{-T}$ and post-multiply by $(I - DX)^{-1}$. These inverses exist because the right side of (51) is positive definite. Then we have

$$X = (I - DX)^{-T}X(I - DX)^{-1},$$

which we can invert and expand to obtain

$$0 = -XD^{T}X - XDX + XDXD^{T}X.$$

Cleared of the X's this expression is

$$DXD^{T} = D + D^{T},$$

which, when subtracted from (63a), leaves us with (63b). The theorem is therefore substantiated.

LEMMA 2. *Given that*

$$\bar{H}(T, t_k) = H(T, t_k), \tag{(30)}$$

then the matrix D defined above has diagonal terms d_{ii} all n of which are negative, except in the special case when we have on $[t_k, T]$,

$$Q(t) \equiv \Phi(t, T)\bar{\Phi}(T, t)B, \tag{64}$$

in which case D is the null matrix.

Proof. Let us generalize the inner product notation (,) of functional analysis [17, p. 80] to include square-integrable vectors and matrices as well. Then we will write

$$(Y, Z) = \int_{t_k}^{T} Y^{T}(t)Z(t)\,dt,$$

no matter whether Y and Z represent scalar, vector, or matrix quantities. In this notation,

$$H = (G, G), \quad \bar{H} = (\bar{G}, \bar{G}), \quad \bar{\bar{H}} = (G, \bar{G}), \tag{65}$$

where equations (13) identify the G matrices as

$$G^{T}(t) = \Phi(T, t)Q(t)N(t), \quad \bar{G}^{T}(t) = \bar{\Phi}(T, t)BN(t), \tag{66}$$

$N(t)$ being a nonsingular matrix (which must exist) satisfying the relation

$$M^{-1}(t) = N(t)N^{T}(t).$$

Let us now express the G matrices in terms of their n column vectors, e.g.,

$$G(t) = [\mathbf{g}_1 \cdots \mathbf{g}_n],$$

whereby the typical ijth elements of (65) become

(67) $$h_{ij} = (\mathbf{g}_i, \mathbf{g}_j), \quad \bar{h}_{ij} = (\bar{\mathbf{g}}_i, \bar{\mathbf{g}}_j), \quad \bar{\bar{h}}_{ij} = (\mathbf{g}_i, \bar{\mathbf{g}}_j).$$

Introducing the Cauchy-Schwartz inequality [10, p. 255], we see that

(68) $$\mathbf{g}_i^T \bar{\mathbf{g}}_j \leqq \{\mathbf{g}_i^T \mathbf{g}_i\}^{1/2} \{\bar{\mathbf{g}}_j^T \bar{\mathbf{g}}_j\}^{1/2},$$

which we can integrate from t_k to T to obtain

(69) $$(\mathbf{g}_i, \bar{\mathbf{g}}_j) \leqq (v, \bar{v}),$$

where

(70) $$v(t) = \{\mathbf{g}_i^T \mathbf{g}_j\}^{1/2}, \quad \bar{v}(t) = \{\bar{\mathbf{g}}_i^T \bar{\mathbf{g}}_j\}^{1/2}$$

are integrable scalar functions. To the right side of (69) we can now apply the Schwartz inequality [16, p. 19],

$$(v, \bar{v}) \leqq (v, v)^{1/2} (\bar{v}, \bar{v})^{1/2},$$

which, upon resubstitution of (69) and (70), leads to the result

(71) $$(\mathbf{g}_i, \bar{\mathbf{g}}_j)^2 \leqq (\mathbf{g}_i, \mathbf{g}_i)(\bar{\mathbf{g}}_j, \bar{\mathbf{g}}_j).$$

Thus, for $i = j$,

(72) $$\bar{\bar{h}}_{ii}^2 \leqq h_{ii} \bar{h}_{ii},$$

so that (30) yields the condition

(73) $$\bar{\bar{h}}_{ii}^2 - \bar{h}_{ii}^2 \leqq 0.$$

But since $\bar{H}$ is positive definite, $\bar{h}_{ii}$ must be positive [10], whereby it follows that (73) holds if and only if

(73′) $$d_{ii} = \bar{\bar{h}}_{ii} - \bar{h}_{ii} \leqq 0.$$

Moreover, for $i = j$, the equality sign in (71)—and therefore the equality sign in (73′)—holds *if* and *only if*

$$\mathbf{g}_i(t) = k_i \bar{\mathbf{g}}_i(t),$$

where k_i is a scalar constant, $i = 1, \cdots, n$. In terms of the G matrices, then,

(74) $$G(t) = \bar{G}(t) K,$$

where

$$K = \operatorname{diag} \{k_1, \cdots, k_n\}.$$

But by hypothesis $(\bar{G}, \bar{G}) = (G, G)$, so that it immediately follows that

$$K = \pm I. \tag{74'}$$

If the sign of (74′) is taken to be (−), then it is clear from (65) and (74) that

$$\bar{\bar{H}}(T, t_k) = -\bar{H}(T, t_k),$$

yielding $d_{ii} = -2h_{ii} \leqq 0$.

If, on the other hand, the (+) sign in (74′) is considered, it is equally clear that $D = 0$. Invoking (66), we see that (74′) with the (+) sign yields (64), and the proof of the lemma is complete.

THEOREM 5 (*The Interference Theorem*). *If X is positive definite, then* (64) *is a necessary condition for Tasks* 1 *and* 2 *to be achieved simultaneously. In other words, except when* (64) *can be satisfied, it is not possible to choose a time-invariant model that can predict exactly both the ideal performance and the model-control performance.*

Proof. To achieve Task 1, Theorem 1 tells us that (30) must hold. Thus the hypothesis of Lemma 2 is met, and *unless* (64) *is satisfied,*

$$d_{ii} < 0.$$

On the other hand, Theorem 4 proves that the second condition for the accomplishment of Task 2 is identical to the pair of relations given in (63). The first of these, (a), tells us that the symmetric part of D must be nonnegative definite, since $D^T X D$ clearly is. But a *necessary* condition for $D + D^T$ to be nonnegative definite is that

$$d_{ii} \geqq 0,$$

which is obviously contradicted by Lemma 2, and thus the theorem is proved.

Thus far, except in the above theorem, no mention has been made of the circumstances under which we can satisfy conditions (29) and (51) of Theorem 3. The first of these conditions will be discussed at length, and rather thoroughly, in a forthcoming paper. The second, (51), introduces complex considerations which we have not been able to resolve completely at the time of this writing, for we seek at least one *real* set of elements $\{b_{ij}\}$ (a) which satisfies (51), and (b) which yields a controllable model.

To illuminate some of the difficulties involved, here, it might be worthwhile to introduce a simple example.

Example 2. Consider the simple first order plant

$$\dot{x} = \frac{1}{t} x + \frac{1}{t} m, \tag{75}$$

to which we want to fit a model

$$\dot{\bar{x}} = a\bar{x} + b\bar{m} \tag{76}$$

in order to achieve Task 2. Suppose the performance criterion is

$$L_k = \int_{t_k}^{T} m^2\, dt + (x(T) - x_d)^2, \tag{77}$$

so that $M = X = 1$. The fundamental solution for (75) is

$$\Phi(t, s) = \Phi(t, 1)\Phi(1, s) = \frac{1}{t}\left(\frac{1}{s}\right)^{-1}$$

or

$$\Phi(t, s) = \frac{s}{t}, \tag{78}$$

whereas the fundamental solution for (76) is

$$\bar{\Phi}(t, s) = e^{-a(t-s)}. \tag{79}$$

With (79) and (78) substituted into (29), it is clear that

$$a = -\frac{\log q}{k'}, \tag{80}$$

where

$$q = \frac{t_k}{T}, \qquad k' = T - t_k. \tag{81}$$

Referring to (13b), we find that (89) provides

$$\bar{H}(T, t_k) = b^2 \frac{k'(q^2 - 1)}{2 \log q}. \tag{82}$$

Also, from (13c),

$$\bar{\bar{H}}(T, t_k) = b \frac{k'(q - 1)}{T \log q}. \tag{83}$$

Therefore

$$\begin{aligned} D &= \bar{\bar{H}} - \bar{H} \\ &= \frac{k'(q - 1)[b/T - (b^2/2)(q + 1)]}{\log q} \end{aligned} \tag{84}$$

which, in combination with the fact that $X = 1$, leaves (51) in the implicit form $(1 - D)^2 = 1$. Thus, $D = 0$ and $D = 2$ are potential solutions.

The case $D = 2$. Making use of (84) explicitly, now, the case $D = 2$ leaves us with the quadratic in b,

$$b^2 - 2\beta_1 b + 2\beta_2 = 0, \tag{85}$$

where

$$\beta_1 = \frac{1}{T(q+1)}, \qquad \beta_2 = 2\frac{\log q}{k'(q^2-1)}. \tag{86}$$

The solution to (85) is of course

$$b = \beta_1 \pm (\beta_1^2 - 2\beta_2)^{1/2}. \tag{87}$$

where the existence of a real b requires the discriminant be nonnegative.

If we take $T = 1$, for example, then equations (86) produce the condition

$$\beta_1^2 - 2\beta_2 = \frac{\delta}{q+1} \geqq 0, \tag{88}$$

where

$$\delta = \frac{1}{q+1} + \frac{4\log q}{(q-1)^2}. \tag{89}$$

From (81) it is clear that we must have $0 \leqq q \leqq 1$, over which interval δ can be shown to be everywhere negative, having, in fact, a stationary maximum of about -9. Thus relation (88) *cannot* be maintained, and $D = 2$ provides no real solution for b.

The case $D = 0$. Employing (84) when $D = 0$ we see that we have the homogeneous quadratic

$$b^2 - 2\beta_1 b = 0, \tag{90}$$

where (86) is again used for simplicity. Clearly, we now have the two real solutions (1) $b = 0$, and (2) $b = 2\beta_1$.

The first of these is not acceptable since it produces a system which is decoupled from the control input. In equivalent language, the control system would be uncontrollable since [13, Theorem 10]

$$\text{Rank } [\text{b}] = 0.$$

The second of these will do, however, leaving us with $b = 2/(T + t_k)$.

So ends the example, wherein we have seen three possible solutions for the B "matrix". One was acceptable; another was discarded because it contained an imaginary component; a third was discarded because it led to uncontrollability, i.e., it *led to a model which could predict exactly that it could do nothing about controlling the system.*

8. Model control of plant. In this section we concern ourselves with Task 3, wherein we seek to minimize the cost discrepancy between the ideal control and the model control: namely, we seek to obtain

$$\bar{\bar{L}}_k{}^* - L_k{}^* = \min.$$

THEOREM 6. *In a disturbance-free environment,*

$$\bar{\bar{L}}_k{}^* - L_k{}^* \geqq 0. \tag{91}$$

Proof. Expression (19a) is the plant input which produces the cost $L_k{}^*$. We see, in fact, that this input satisfies the Euler-Lagrange equations as well as the remaining necessary and sufficient conditions for a minimum [6, Appendices A, B]. Thus $\bar{\bar{L}}_k{}^*$ can *never* be less than $L_k{}^*$, and can equal it only when

$$\bar{\bar{\mathbf{m}}}^*(t) \equiv \mathbf{m}^*(t).$$

Perfect model control. Although one might suspect that the equality sign in (91) can hold only if plant and model are everywhere identical, such is *not* the case, as the following development indicates.

DEFINITION 7. We will define a *perfect model-controlled system* to be a model-controlled system which yields a performance *identical* to that obtained when the control is implemented according to the exact plant equations. That is, in a disturbance-free environment, a perfect model controlled system is one wherein

$$\bar{\bar{L}}_k{}^* - L_k{}^* = \min = 0. \tag{92}$$

THEOREM 7. *Given that X is positive definite, then the necessary and sufficient conditions for a model-controlled system to be perfect are the following:*

$$\bar{\bar{\Phi}}(T, t_k) = \Phi(T, t_k), \tag{93}$$

$$B \equiv \bar{\bar{\Phi}}(t, T)\Phi(T, t)Q(t). \tag{(64)}$$

Proof. According to Theorem 1, Task 1 can be accomplished *if* and *only if* (29) and (30) prevail, in which case

$$L_k{}^* - \bar{L}_k{}^* = 0. \tag{(10)}$$

According to Theorem 3, Task 2 can be accomplished *if* and *only if* (29) and (51) prevail, in which case

$$\bar{\bar{L}}_k{}^* - \bar{L}_k{}^* = 0. \tag{(11)}$$

Note that a sufficient condition for the satisfaction of (51) is $D = 0$, or, in other words,

$$\bar{\bar{H}}(T, t_k) = \bar{H}(T, t_k). \tag{94}$$

If we now subtract (10) from (11), it becomes evident that (92) is maintained *if* and *only if* these two equations are true simultaneously. But according to Lemma 2 and Theorem 5 which follows it, we cannot achieve (10) and (11) simultaneously unless (64) holds, and then (94) is the only possible solution of (51).

Let us now intersect the sets of necessary and/or sufficient conditions for (10) and (11) to be coincidently satisfied. The set (29), (30), and (51) is necessary and sufficient. Equation (64) is necessary and implies (but is not implied by) both (30) and (51). Therefore (29) and (64) are the necessary and sufficient conditions, and the theorem is proved.

Remark B. If (64) cannot be attained, then (29) is still *necessary* for a global minimum in $\bar{\bar{L}}_k^* - L_k^*$, although that minumum does not equal zero everywhere in $\mathbf{u}$ space.

For let us take (26) and examine it for the case where $\mathbf{u} = 0$. Then

$$\bar{\bar{L}}_k^* - L_k^* = \| \bar{\mathbf{u}} \|^2_{\bar{U}\bar{H}\bar{U}+\bar{U}\bar{\bar{H}}^T X\bar{\bar{H}}\bar{U}} , \tag{95}$$

and since $\bar{U}\bar{H}\bar{U}$ is positive definite, we can employ a result of Bellman's [15, p. 115] to tell us that the entire weighting matrix is positive definite. Clearly then, (95) achieves its minimum value when $\bar{\mathbf{u}} = 0$, from which we can conclude—as in the proof of Theorem 1—that $\bar{\Phi}(T, t_k)$ must equal $\Phi(T, t_k)$. So ends the remark.

Strengthening the interference theorem. Physically, the satisfaction of (29) insures us that when the plant decays naturally to $\mathbf{x}_d$ at $t = T$, thereby requiring no input, the model controller properly supplies it with no input. That is, *if* and *only if* (29) is satisfied will

$$\bar{\bar{L}}_k^* = \bar{L}_k^* = L_k^* = 0 \tag{96}$$

when the plant homogeneously hits the desired endpoint.

COROLLARY 7.1. *If* (64), *the second of the necessary and sufficient conditions for perfect model-control, cannot be met, then the equality*

$$L_k^* - \bar{L}_k^* = 0 \tag{(10)}$$

implies

$$\bar{\bar{L}}^* - \bar{L}_k^* \geqq 0, \tag{97}$$

with the equality sign holding in the latter only when the system incurs zero costs, i.e., when (96) *holds.*

Proof. If perfect model control cannot be obtained, then, according to Theorem 6,

$$\bar{\bar{L}}_k^* - L_k^* \geqq 0, \tag{98}$$

with strict inequality prevailing except at isolated points in **u**-space. Subtracting (10) from (98), we find

$$\bar{\bar{L}}_k{}^* - \bar{L}_k{}^* > 0,$$

except at isolated points such as correspond to

$$\mathbf{u} = \bar{\mathbf{u}} = 0.$$

In fact, this point is the *only* point at which the equality in (97) can manifest itself. For, given (29), $\bar{\mathbf{u}}$ is the same as **u** everywhere in **u**-space. Thus (24) is simply

$$\bar{\bar{L}}_k{}^* = \| \mathbf{u} \|^2_{\bar{U}\bar{\bar{H}}\bar{U}+(I-\bar{\bar{H}}\bar{U})^T X(I-\bar{\bar{H}}\bar{U})} ,$$

which is clearly a positive definite quadratic form that describes a hyperparaboloid symmetric about the range axis and zero at the origin. Since $\bar{L}_k{}^*$ is also such a paraboloid, it follows from analytic geometry that $\bar{\bar{L}}_k{}^*$ cannot equal $\bar{L}_k{}^*$ (except at the origin) unless $\bar{\bar{L}}_k{}^*$ is everywhere identical to $\bar{L}_k{}^*$, i.e., *unless we have perfect model-control.* This completes the proof.

Corollary 7.1 (Alternate Statement). *If* (64) *cannot be achieved, then, except when the system cost is zero, a model which is chosen to predict exactly the ideal cost of controlling the system will always underestimate its own cost of controlling the system.*

Corollary 7.2. *If* (64), *the second of the necessary and sufficient conditions for perfect model-control, cannot be met, then the equality*

$$\bar{\bar{L}}_k{}^* - \bar{L}_k{}^* = 0 \tag{(11)}$$

implies

$$L_k{}^* - \bar{L}_k{}^* \leqq 0, \tag{99}$$

with the equality sign holding in the latter only when the system incurs zero cost, i.e., when (96) *holds.*

Proof. The subtraction of (98) from (11) yields (99), from which the proof follows as in the previous corollary.

Corollary 7.2 (Alternate Statement). *If* (64) *cannot be achieved, then, except when the system cost is zero, a model which is chosen to predict exactly its own cost of controlling the system will always overestimate the cost of controlling the system with the most accurate equations available, i.e.*, (2).

9. Concluding remarks. If (64) can be met, we have shown that plant and model may be *identical in every performance aspect*, even though the plant is time-varying and the model is not. Essentially, then, we can use our performance functionals to partition the universe of linear systems into equivalence classes [16, p. 12], wherein the equality of the L_k^* norms defines a class of linear systems (time-varying or not) which are *performance-equivalent.* In fact, we can generalize Theorem 7 into the following.

THEOREM 8. *Given that X is positive definite, then the necessary and sufficient conditions for two systems to be performance-equivalent in the large are*

$$\bar{\Phi}(T, t_k) = \Phi(T, t_k),$$

$$\bar{Q}(t) \equiv \bar{\Phi}(t, T)\Phi(T, t)Q(t),$$

where the system plants are respectively

$$\dot{\mathbf{x}} = P(t)\mathbf{x} + Q(t)\mathbf{m},$$

and

$$\dot{\bar{\mathbf{x}}} = \bar{P}(t)\bar{\mathbf{x}} + \bar{Q}(t)\bar{\mathbf{m}},$$

and the barred fundamental matrix associates with the last plant.

Moreover, the control laws of the two plants are interchangeable.

Proof. The proof follows as for Theorem 7, except that Requirement D no longer applies since modeling is not involved here, for the moment. Thus, (39) need not result in the relation $\bar{\Phi}(T, t_k) = \Phi(T, t_k)$ unless the initial-condition vector that appears there is arbitrary: i.e., unless we are trying to obtain Tasks 1 and 2 in the large.

That the optimal control laws are interchangeable, now, follows from writing (19a) in terms of the variables for each of the above plants and then substituting the above conditions for performance-equivalence into one of the two control laws so obtained. It thereby becomes apparent that

$$\bar{\mathbf{m}}^*(t) \equiv \mathbf{m}^*(t),$$

and the theorem is proved.

Much of the work presented in this paper, like the above, has depended on the hypothesis that the X matrix in the criterion be positive definite. When this matrix is positive semi-definite instead—but not zero—and if $\mathbf{x}(T)$ is not fixed *a priori*, then all the *necessary* and *sufficient* conditions of this paper degenerate to *sufficient* ones only.

When X is null and $\mathbf{x}(T)$ is completely specified,[6] *all* of the necessary and sufficient conditions still hold, with minor modifications appearing only in the conditions of Theorem 3. For in all the theorems except that one, the proofs carry through equally well—and, in fact, more simply—if U is H^{-1} instead of $X(I + HX)^{-1}$, and $\bar{U}$ is $\bar{H}^{-1}$ instead of $X(I + \bar{H}X)^{-1}$.

In Theorem 3, however, the second of the necessary and sufficient conditions, i.e., (51), must be replaced with

[6] As is considered in much of the recent literature: for example, [12], [13], [14].

$$\bar{\bar{H}}(T, t_k) = \bar{H}(T, t_k), \tag{100}$$

since (52) reduces to

$$0 = \| \mathbf{u} - \bar{\bar{H}}\bar{H}^{-1}\bar{\mathbf{u}} \|^2.$$

For, if we replace the vectors $\mathbf{u}$ and $\bar{\mathbf{u}}$ with $(\mathbf{u}^0 + \Delta\mathbf{u})$ and $(\bar{\mathbf{u}}^0 + \Delta\mathbf{u})$, respectively, we can obtain (100) using the same approach as in the theorem.

In closing, we can say that much of our work should be considered more from an existence viewpoint than from a design viewpoint. Certainly, if we can assume that a greal deal is known about the plant, as is often the case with aerospace vehicles, then, with nominal storage requirements, models can be constructed from our formulas for on-line computation and control.

However, the primary purpose of a large portion of our work was to find out the basic limitations on modeling—what *could* and *could not* be done with a model, independent of the difficulty attached with the finding of that model. Thus, even when the plant is poorly known, as with many chemical, metallurgical, and biological processes, still our results may be of use in that they tell us what we should and should not expect of a model in an ideal environment.

10. Acknowledgments. The work included here would not have been possible without the direction of the author's principal faculty advisor at Case Institute of Technology, Dr. M. D. Mesarovic, who provided much of the conceptual basis for the research and all the freedom that the author desired. In addition, two of the author's colleagues, (now Drs.) R. W. Brockett and T. G. Windeknecht, provided innumerable criticisms to the details of the theory, and the results of their remarks appear virtually everywhere.

REFERENCES

[1] J. K. Lubbock, *The optimization of a class of non-linear filters*, Proc. Inst. Elec. Engrs. C., 107 (1960), pp. 60–74.

[2] R. B. Kerr and W. H. Surber, Jr., *Precision of impulse-response identification based on short, normal operating records*, I. R. E. Trans. on Automatic Control, AC-6 (1961), pp. 173–183.

[3] W. J. Culver and M. D. Mesarovic, *Dynamic statistical linearization*, IEEE Trans. Comm. and Electronics, 67 (1963), pp. 317–324.

[4] J. H. Laning, Jr., and R. H. Battin, *Random Processes in Automatic Control*, McGraw-Hill, New York, 1956.

[5] L. Birta, *The relative performance of interacting and non-interacting systems*, M. S. Thesis, Case Institute of Technology, Cleveland, Ohio, 1963.

[6] W. J. Culver, *A modeling theory for a class of optimal control systems*, Ph.D. Thesis, Case Institute of Technology, Cleveland, Ohio, 1964. Also published as Systems Research Center report SRC 38-A-63-15, September, 1963.

[7] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1956.

[8] I. P. Natanson, *Theory of Functions of a Real Variable*, vol. I, transl., L. F. Boron, Frederick Ungar, New York, 1961.

[9] I. J. Epstein, *Conditions for a matrix to commute with its integral*, Proc. Amer. Math. Soc., 14 (1963), pp. 266–270.

[10] F. R. Gantmacher, *Theory of Matrices*, vol. I, Chelsea, New York, 1959.

[11] ———, Ibid., vol. II.

[12] R. E. Kalman, *Contributions to the theory of optimal control*, Proc. Mexico City Conference on Ordinary Differential Equations, 1959; Bol. Soc. Mat. Mexicana, 1 (1960), pp. 102–119.

[13] R. E. Kalman, Y. C. Ho, and K. S. Narenda, *Controllability of linear dynamic systems*, Contributions to Differential Equations, 1 (1962), pp. 189–213.

[14] B. Friedland, *The design of optimal controllers for linear processes with energy limitations*, Trans. ASME, D (1963), pp. 181–196.

[15] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[16] A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, vol. I, (*Metric and Normed Spaces*), transl. L. F. Boron, Graylock Press, Rochester, New York, 1957.

[17] ———, Ibid., vol. II, (*Measure, The Lebesgue Integral, Hilbert Space*), Albany, New York, 1961.

[18] W. J. Culver, *On improving the performance of a class of optimal control systems by employing simple coarse controls*, to be presented at the Joint IEEE–Optical Society Symposium on Optimization Techniques, April 21–23, 1965, Pittsburgh, Pennsylvania.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

MINIMUM EFFORT CONTROL OF SEVERAL TERMINAL COMPONENTS*

J. V. BREAKWELL† AND F. TUNG‡

Abstract. The stochastic control problem of minimizing the total average velocity correction with several prescribed terminal variances in the presence of random injection and measurement errors is considered. It is shown that, for the case of linear feedback, this can be formulated as an optimization problem for an equivalent deterministic system whose states are the covariances of the predicted miss. However, the deterministic optimization problem is "degenerate" causing some difficulty in the computation of the feedback gain. It is shown that the optimum linear corrective strategy is, in general, discontinuous and consists of an initial period of no control, followed by a period of continuous control and finally a period of no control and possibly an impulse at the end. Equations are derived from which the variable feedback gain and the various time intervals can be computed. Two simple examples involving (1) the control of two terminal position components, and (2) the control of both the terminal position and the terminal velocity are considered in detail. Numerical results are given showing the comparison between this solution and that obtained by using the well known theory for the quadratic loss criterion. In particular, the computation includes, for the two position case, a gap in the information.

1. Introduction. The stochastic control problem of guiding a vehicle from its injection to prescribed rms terminal conditions in the presence of random injection and measurement errors with a minimum amount of fuel is of considerable interest in the field of interplanetary navigation [1], [2], [3]. It is often assumed that the vehicle dynamics are governed by known laws and that the departures of velocities and positions from the nominal trajectories are sufficiently small so that a linearized model evaluated along this nominal path may be used. One way of dealing with this kind of stochastic optimization problem is to define a meaningful average quantity and then formulate the problem in terms of an equivalent deterministic optimization problem using the average quantities as the states. This technique was used by Breakwell and Striebel [4] who recently developed a minimum effort theory when the variance of a single terminal component is specified. It was assumed that the control is linear and that the mechanization errors are negligible. The effort to be minimized is the expected value of the integral of the absolute value of the command acceleration and hence is simply related to the amount of fuel required in the case of chemi-

* Received by the editors February 16, 1964, and in revised form September 19, 1964.

† Lockheed Missiles and Space Company, Palo Alto, California, and Department of Aeronautics and Astronautics, Stanford University, Stanford, California.

‡ Research Laboratories, Lockheed Missiles and Space Company, 3251 Hanover Street, Palo Alto, California.

cal propulsion systems. An ingenious application of Green's Theorem [5] was used in obtaining the solution. The purpose of this paper is to extend the theory of Breakwell and Striebel to the case when the variances of more than one terminal component are specified. This arises, for instances, in interplanetary guidance when both the in-plane and the out-of-plane terminal positions are to be independently controlled and in problems of rendezvous when both the position and the velocity at the terminal time are specified. The novelty of the extension in this paper lies in the solution of the optimization problem by direct application of the maximum principle [5] since Green's Theorem cannot be readily applied in the multi-dimensional case. Also, a slightly modified criterion for the effort is used. This is because the expected amount of total velocity correction in the multi-dimensional case is expressible only in the form of an infinite series [6]. A reasonable criterion which we have adopted in this paper is the integral of the square root of the variance of the command acceleration. This loss function has the properties that (1) it reduces to the exact amount of total velocity requirements in the absence of random disturbances, (2) it reduces (except for an unimportant factor $\sqrt{2/\pi}$) to the same criterion as that used by Breakwell and Striebel in the case of controlling only one terminal miss, and (3) it sets an upper bound to the expected total velocity correction. The last statement can be easily verified by application of Schwarz's inequality.

The statement of the mathematical problem and the transformation to an equivalent optimization problem for a deterministic system are given in §2. It will be seen that the states of the equivalent deterministic system are the covariances of the predicted miss. Section 3 gives the necessary conditions for the optimal linear control and the forms of the optimal feedback coefficients. It is shown that, in general, the optimum linear corrective strategy consists of an initial period of no control while the information catches up. This is followed by a period of continuous control and finally a period of no control and possibly an impulse at the end. The section concludes with an outline of a computation procedure with which the optimal corrective strategy can be obtained. Two simple examples illustrating the techniques developed in this paper are considered in detail in §4. Numerical results are given and in order to get the "feel" of the solution, the results are compared with that obtained by using the well known theory of the quadratic loss criterion [7]. We include in this paper an appendix which specializes the results to the case when only one of the terminal variances is specified. It is included here for the purpose of establishing an equivalence between the results of this paper and that of Breakwell and Striebel [4].

Notation. Capital letters A, B, $\cdots$ denote matrices and small letters

a, b, $\cdots$ denote vectors. The elements of the matrix A are denoted by a_{ij} and the elements of the vector a are denoted by a_i.

2A. Statement of the problem. Given:
(1) The linearized equations of motion,

$$\dot{x}(t) = F(t)x(t) + G(t)(u(t) + \eta(t)), \tag{2.1}$$

and the observations*

$$y(t) = M(t)x(t) + \epsilon(t), \tag{2.2}$$

where $x(t)$ is a state n-vector, $u(t)$ a control m-vector, $y(t)$ an observable r-vector ($r \leqq n$), $\eta(t)$ a random m-vector accounting for the mechanization error (it will be assumed in this paper that the mechanization error is negligible), $\epsilon(t)$ a random r-vector accounting for the measurement error. It is assumed that $\epsilon(t)$ is normally distributed with zero mean and covariance

$$\operatorname{cov}(\epsilon(t), \epsilon(s)) = R(t)\delta(t - s), \tag{2.3}$$

where $\delta(\cdot)$ is the Dirac delta function.
(2) The covariance of the initial state

$$\operatorname{cov}(x(0)) = V(0). \tag{2.4}$$

We shall assume that $E(x(0)) = 0$.
(3) A $p \times n$ matrix H, where $p \leqq n$.

Find: The control $u(t)$, $t \in (0, T)$, as a linear functional of $y(s)$, $0 \leqq s \leqq t$, that minimizes

$$\int_0^T \sqrt{E \| u(t) \|^2}\, dt \tag{2.5}$$

for specified values of $\operatorname{cov}(Hx(T))_{ii}$, $i = 1, 2, \cdots, p$, where $\| u \|^2 = u'u$.

2B. Transformation to a deterministic optimization problem. We shall now show that by properly defining some average quantities, the solution of the stochastic optimization problem posed above can be obtained from solving a deterministic optimization problem using these average quantities as the states. Let $\hat{x}(t)$ be the estimate of $x(t)$ given by

$$\hat{x}(t) = E(x(t) \mid y(s), \quad 0 \leqq s \leqq t), \tag{2.6}$$

* Equations (2.1) and (2.2) are the perturbation equations along a nominal trajectory which is assumed to have been pre-computed. The elements of the matrices are therefore the partial derivatives evaluated along this nominal path.

and let $V(t)$ be the covariance of the estimation error

$$V(t) = \operatorname{cov}\,(x(t) - \hat{x}(t)). \tag{2.7}$$

It has been shown by Kalman [8] that

$$\operatorname{cov}\,(\hat{x}(t), x(t) - \hat{x}(t)) = 0, \tag{2.8}$$

$$\dot{\hat{x}}(t) = F(t)\hat{x}(t) + G(t)u(t) + \beta(t), \quad \hat{x}(0) = 0, \tag{2.9}$$

and

$$\dot{V}(t) = F(t)V(t) + V(t)F'(t) - V(t)\dot{I}(t)V(t), \tag{2.10}$$

where

$$\dot{I}(t) = M'(t)R^{-1}(t)M(t) \tag{2.11}$$

is the information rate matrix relative to the state $x(t)$, and

$$\beta(t) = V(t)M'(t)R^{-1}(t)(y(t) - M(t)\hat{x}(t)), \tag{2.12}$$

which may be considered as a white noise with covariance matrix

$$\operatorname{cov}\,(\beta(t), \beta(s)) = V(t)\dot{I}(t)V(t)\delta(t - s). \tag{2.13}$$

Let $\Phi(T, t)$ be the transition matrix satisfying the matrix differential equation

$$\dot{\Phi}(T, t) = -\Phi(T, t)F(t), \quad \Phi(T, T) = I, \tag{2.14}$$

and let

$$\hat{x}(T, t) = \Phi(T, t)\hat{x}(t), \tag{2.15}$$

which is the predicted miss of the state at the final time based on all the data up to time t under the assumption that no additional control is applied over the interval (t, T). Using (2.9) and (2.14), it is seen that those predicted miss components whose terminal rms values are prescribed satisfy the differential equation

$$H\dot{\hat{x}}(T, t) = H\Phi(T, t)G(t)u(t) + H\Phi(T, t)\beta(t). \tag{2.16}$$

It has been shown by Striebel [9] that the optimal linear corrective strategy (i.e., the class of controls which is restricted to be a linear functional of the past observations and which is the class to be considered in this paper) depends only linearly on $H\hat{x}(T, t)$. Hence, without loss of generality, we may let the optimal linear control be given by

$$u(t) = -S(t)H\hat{x}(T, t), \tag{2.17}$$

where $S(t)$ is an $m \times p$ matrix whose elements are to be determined such that (2.5) is minimized for specified values of $\operatorname{cov}(Hx(T))_{ii}$, $i = 1, 2, \cdots, p$. One method of doing this is to formulate this stochastic optimization problem in terms of an optimization problem for a deterministic system using the elements of the covariance matrix of $H\hat{x}(T, t)$ as the states. To do this, we define

$$P(t) = E(H\hat{x}(T,t)\hat{x}'(T,t)H'), \tag{2.18}$$

which is equivalent to cov $(H\hat{x}(T, t))$ since $\hat{x}(T, t)$ is a zero mean process. Using (2.7), (2.8), (2.13), (2.16) and (2.17), it is seen that

$$\begin{aligned}\dot{P}(t) = -H\Phi(T, t)G(t)S(t)P(t) - P(t)S'(t)G'(t)\Phi'(T, t)H' \\ + H\Phi(T, t)V(t)\dot{I}(t)V(t)\Phi'(T, t)H',\end{aligned} \tag{2.19}$$

$$\operatorname{cov}(Hx(T)) = P(T) + HV(T)H', \tag{2.20}$$

and

$$E\| u(t) \|^2 = E\| -S(t)H\hat{x}(T, t) \|^2 = \operatorname{tr} P(t)S'(t)S(t). \tag{2.21}$$

Since the last term of (2.20) is independent of the control, it follows that specification of $\operatorname{cov}(Hx(T))_{ii}$ is the same as specifying $P_{ii}(T)$ and the determination of $S(t)$ is equivalent to solving the following deterministic optimization problem:

Given: the dynamic system (2.19) *with* $P(0) = 0$. *Find* $S(t)$ *which minimizes*

$$\int_0^T \sqrt{\operatorname{tr} P(t)S'(t)S(t)}\, dt \tag{2.22}$$

for specified values of $P_{ii}(T)$, $i = 1, 2, \cdots, p$.

Inspection of (2.19) and (2.22) shows that both are linear in S in so far as the magnitude is concerned. This is a "degenerate" (or singular) problem in the calculus of variations and special techniques are usually necessary for the method of solution. In general, the optimal solution will consist of different subarcs connected at a finite number of points, called the corner points. At the corner points, the adjoint variables (to be defined in the next section) must be continuous. We shall obtain the solution by application of the maximum principle. This is done in the next section.

3. Equations for optimality and computation procedure. To put in evidence the "singular" nature of the problem, we define*

$$\phi(t) = \sqrt{\operatorname{tr} PS'S}, \quad \phi(t) \geqq 0, \tag{3.1}$$

* For convenience, we shall, hereafter, omit the argument t.

and let the matrix of feedback gains be written as†

$$S = \phi(t)B, \tag{3.2}$$

where B is an undetermined $m \times p$ matrix (undefined when $\phi = 0$) such that

$$\operatorname{tr} PB'B = 1. \tag{3.3}$$

Substituting (3.2) into (2.19) shows

$$\dot{P} = -\phi(t)(H\Phi GBP + PB'G'\Phi'H') + Q, \tag{3.4}$$

where

$$Q = H\Phi V\dot{I}V\Phi'H' \tag{3.5}$$

is a known function of time. The problem now reduces to that of finding B and $\phi(t) \geqq 0$ which minimizes $\int_0^T \phi\, dt$ subject to the constraint (3.3) and specified values of $P_{ii}(T)$.

Let the Hamiltonian be given by

$$2\phi(t) + \operatorname{tr} \Lambda\dot{P}, \tag{3.6}$$

where the elements of the $p \times p$ symmetric matrix Λ are the adjoint variables. For a given $\phi(t) \neq 0$, minimizing this Hamiltonian with respect to B subject to the constraint (3.3) is a simple nondegenerate problem in calculus of variation. The necessary equations for optimality are

$$B = \frac{G'\Phi'H'\Lambda}{\sqrt{\operatorname{tr} P\Lambda Z\Lambda}}, \tag{3.7}$$

$$\dot{P} = -\frac{\phi(t)(Z\Lambda P + P\Lambda Z)}{\sqrt{\operatorname{tr} P\Lambda Z\Lambda}} + Q, \tag{3.8}$$

and

$$\dot{\Lambda} = \phi(t)\frac{\Lambda Z\Lambda}{\sqrt{\operatorname{tr} P\Lambda Z\Lambda}}, \tag{3.9}$$

where $Z = H'\Phi GG'\Phi'H'$ and is a given function of time. The transversality conditions are $\lambda_{ij}(T) = 0$, $i \neq j$; $\lambda_{ii}(T) = c_i$, $i = 1, 2, \cdots, p$, where c_i are to be adjusted such that $P_{ii}(T)$ meet the prescribed values.

The Hamiltonian now becomes linear in $\phi(t)$ and can be written as

$$2\phi(t)(1 - \sqrt{\operatorname{tr} P\Lambda Z\Lambda}) + \operatorname{tr} \Lambda Q. \tag{3.10}$$

† This substitution essentially converts a control problem potentially singular in $m \times p$ variables into a problem which is singular in only one variable.

It only remains to minimize this Hamiltonian with respect to $\phi(t)$. Since $\phi(t) \geqq 0$, it follows that $\phi = 0$ if $\operatorname{tr} P\Lambda Z\Lambda < 1$, is undetermined if $\operatorname{tr} P\Lambda Z\Lambda = 1$, and is infinite if $\operatorname{tr} P\Lambda Z\Lambda > 1$. The last case cannot occur over any finite interval since otherwise $\int_0^T \phi(t)\,dt$ will diverge. Now $\operatorname{tr} P\Lambda Z\Lambda = 0$ at $t = 0$ and can be shown to be continuous for any $\phi(t) \geqq 0$ including impulses (i.e., $\phi(t)$ are Dirac delta functions). Hence the case $\operatorname{tr} P\Lambda Z\Lambda > 1$ cannot occur and we are left with either $\phi = 0$ (when $\operatorname{tr} P\Lambda Z\Lambda < 1$), or $\phi \neq 0$, in which case $\operatorname{tr} P\Lambda Z\Lambda = 1$.

It turns out that the optimal gain S consists of (in general, but not always) three portions; an initial period of no control where $S = 0$, followed by a period of continuous control, and finally a period of no control and possibly an impulse at the end. Let us now consider the two cases.

(1) $\phi(t) = 0$. Equations (3.8) and (3.9) reduce to

$$\dot{\Lambda} = 0 \tag{3.11}$$

and

$$\dot{P} = Q, \tag{3.12}$$

which show that the adjoint variables remain unchanged during this period.

(2) $\phi(t) \neq 0$. Then

$$\operatorname{tr} P\Lambda Z\Lambda = 1. \tag{3.13}$$

This defines a surface which must contain the solution whenever $\phi \neq 0$. We now note that in order to integrate the set of equations (3.8) and (3.9) along this surface, it is necessary to express $\phi(t)$ in terms of P and Λ. This is done by twice differentiating (3.13). It is of interest to note that along this surface, $\phi(t)$ is also given by

$$\phi(t) = \operatorname{tr} P\dot{\Lambda}, \tag{3.14}$$

which can be verified by combining (3.9) and (3.13). It is a measure of the average "acceleration" and vanishes only when $S = 0$, or equivalently, $\Lambda = \text{constant}$.

Differentiating (3.13) once, using (3.8), (3.9) and the commutative properties of the trace operations, we find

$$\operatorname{tr}(P\Lambda\dot{Z}\Lambda + Q\Lambda Z\Lambda) = 0. \tag{3.15}$$

Differentiating (3.15) once more yields a relation between P, Λ, and $\phi(t)$ which after suitable reduction can be written as

$$\phi(t) = \frac{-\operatorname{tr}(2Q\Lambda\dot{Z}\Lambda + \dot{Q}\Lambda Z\Lambda + P\Lambda\ddot{Z}\Lambda)}{2\operatorname{tr}(Q\Lambda Z\Lambda Z\Lambda)}. \tag{3.16}$$

We now have the necessary equations, namely (3.8), (3.9), (3.13), (3.15) and (3.16), for computing the optimal feedback gains. It is noted that the denominator in (3.16) is the trace of the product of two positive semidefinite symmetric matrices and hence is always $\geqq 0$. It will be assumed to be > 0 in this paper. In other words, the matrix $Q\Lambda Z\Lambda Z\Lambda$ is not identically zero.

Since $P(0) = 0$, it follows that (3.13) cannot be satisfied at $t = 0$. Hence, $\phi(0) = 0$ and there will be an initial period of no control. The time at which the control is first turned on depends on (1) the information rate which is imbedded in Q, and (2) the initial values of Λ. Mathematically, the exact time of turning on is determined by simultaneously satisfying (3.13) and (3.15). It should be noted that satisfaction of (3.15) determines the time. The common multiplicative constant of the adjoint variables is determined by the normalizing equation (3.13).

Computation starts by guessing an initial set of $\Lambda(0)$ and integrating the dynamic equation (3.12) forward until (3.15) is satisfied. This determines t_{on}. Use is then made of (3.13) to compute the normalizing constant which determines the adjoint variables at the time of turning on. We are now on the surface such that $\phi \neq 0$. To proceed along this surface, we use (3.16) to find $\phi(t)$. This is then used in (3.8) and (3.9) to integrate the equations for P and Λ forward. The optimal feedback gain can be obtained by using ϕ and (3.7). Assume that the control is turned off at some time t, say $t_{\text{off}} \geqq t_{\text{on}}$. Then $S(t) = 0$ for $t > t_{\text{off}}$. The total average velocity correction required is given by

$$\int_{t_{\text{on}}}^{t_{\text{off}}} \phi(t)\, dt, \tag{3.17}$$

and

$$\Lambda(T) = \Lambda(t_{\text{off}}), \tag{3.18}$$

$$P(T) = P(t_{\text{off}}) + \int_{t_{\text{off}}}^{T} Q(t)\, dt. \tag{3.19}$$

The computational procedure we have proposed gives a parametric study of $p(p + 1)/2$ elements consisting of the ratio of the initial adjoint variables and t_{off} as functions of the $p(p + 1)/2$ elements of $P(T)$. Let

$$A(t) = P(t) + H\Phi(T, t)V(t)\Phi'(T, t)H'. \tag{3.20}$$

Then $A(t)$ is the covariance of the actual terminal miss when the control is turned off at t. Hence, without loss of generality, we may consider that the parametric study is between the $p(p + 1)/2$ elements consisting of the ratio of the initial adjoint variables and t_{off} and the $p(p + 1)/2$

elements of $A(t_{\text{off}})$. If $A_{ii}(t_{\text{off}})$ for all $t_{\text{off}} \in (t_{\text{on}}, T)$ do not meet the specified values, the computation is repeated again with an improved estimate of $\Lambda(0)$.

It should be noted that the computation procedure we have outlined assumes that the computed $\phi(t) > 0$. In the event that $\phi(t)$ becomes negative for some $t \in (t_{\text{on}}, t_{\text{off}})$, then there exist periods of no control in the interval $(t_{\text{on}}, t_{\text{off}})$. Physically, this implies that it is not possible to follow the critical surface defined by (3.13). Assume t_1 is the first time such that $\phi(t_1) < 0$; then the control must be turned off at some time t before t_1. The problem here is to determine the exact times of leaving the surface and intercepting the surface again. This can be done by using the criterion that the adjoint variables must remain constant during the time that the control is off. It is equivalent to the searching of a normalization constant which must remain the same at the two points. An iterative scheme taking care of this can be easily implemented on the digital computer. This is illustrated in one of the numerical examples given in the next section.

So far, we have avoided the possibility of impulsive corrections, i.e., S or $\phi(t)$ are impulses. Impulsive corrections give rise to discontinuities in P and Λ. Let de be the incremental effort. Then

$$de = \phi(t)\,dt, \tag{3.21}$$

so that the effort due to this impulsive correction is

$$e = \int_{t^-}^{t^+} \phi(t)\,dt. \tag{3.22}$$

Using the effort as the independent variable, (3.8) and (3.9) can be written as

$$\frac{d\Lambda}{de} = \Lambda Z \Lambda \tag{3.23}$$

and

$$\frac{dP}{de} = -(Z\Lambda P + P\Lambda Z). \tag{3.24}$$

The relation between the amount of the impulsive effort and the jump (or drop) in Λ (or P) can therefore be obtained by directly integrating (3.23) and (3.24) with respect to the effort. Using (3.23) and (3.24), we find

$$\frac{d\,\mathrm{tr}\,(P\Lambda Z\Lambda)}{de} = 0, \tag{3.25}$$

which implies that impulsive corrections leave tr $(P\Lambda Z\Lambda)$ invariant. In fact, (3.25) is true for any initial values of tr $(P\Lambda Z\Lambda)$. This, incidently, is necessary for establishing the fact that tr $(P\Lambda Z\Lambda)$ is continuous. We shall now show that impulsive corrections can be applied at t_0 if and only if $Q(t)$ is discontinuous at t_0.

Assume that an impulse is applied at t_0 and $Q(t)$ is continuous at t_0. The time derivative of tr $(P\Lambda Z\Lambda)$ is tr $(P\Lambda \dot{Z}\Lambda + Q\Lambda Z\Lambda)$ which, immediately after the impulse of area E, is given by

$$\text{(3.26)} \qquad \operatorname{tr}(P\Lambda\dot{Z}\Lambda + Q\Lambda Z\Lambda)\Big|_{t_0^-} + \int_0^E \frac{d\operatorname{tr}(P\Lambda\dot{Z}\Lambda + Q\Lambda Z\Lambda)}{de}\,de.$$

Now, the first term in (3.26) is zero since we were on the singular surface at t_0^-. Using (3.23) and (3.24), we see that the second term in (3.26) can be written as

$$2\int_0^E \operatorname{tr}(Q\Lambda Z\Lambda Z\Lambda)\,de,$$

which is greater than 0 in view of our assumption that $Q\Lambda Z\Lambda Z\Lambda$ is not identically zero. This implies that tr $P\Lambda Z\Lambda$ will be greater than 1 for $t > t_0$, which is not permissible. Hence, impulsive corrections cannot be applied at any time when $Q(t)$ is continuous. (This is the same as requiring that the Hamiltonian be continuous.) On the other hand, assume $Q(t)$ is discontinuous at t_0. Inspection of (3.15) shows that it can be satisfied only if P and Λ are discontinuous at t_0. Hence, impulsive corrections are allowed to occur when $Q(t)$ is discontinuous or at the final time since our argument does not apply there.

Remark 1. In most cases, the optimal corrective strategy consists of an initial period of no control, followed by a period of continuous control, and finally a period of no control and possibly an impulse at the end. Corresponding to that $\Lambda(0)$, the possibility of periods of no control between t_{on} and t_{off} when $\phi(t) > 0$ for all $t \in (t_{\text{on}}, t_{\text{off}})$ can be established easily by computing the quantity (tr $P(t')\Lambda(t)Z(t')\Lambda(t) - 1$) for all $t' > t$, $t \in (t_{\text{on}}, t_{\text{off}})$. If it differs from zero, then it can be concluded that there do not exist periods of no control between t_{on} and t_{off}.

Remark 2. It is not clear whether or not there exist different initial values of the adjoint variables which will give rise to the same terminal conditions. This is a problem involving the uniqueness of our solution and as such has not been solved.

Remark 3. It is shown in the appendix that in the case of controlling only one terminal component, the solution we have obtained is unique and that there exist no periods of no control between t_{on} and t_{off} if $\phi(t)$ is positive over this period. This was first solved by Breakwell and Striebel [4] using Green's Theorem.

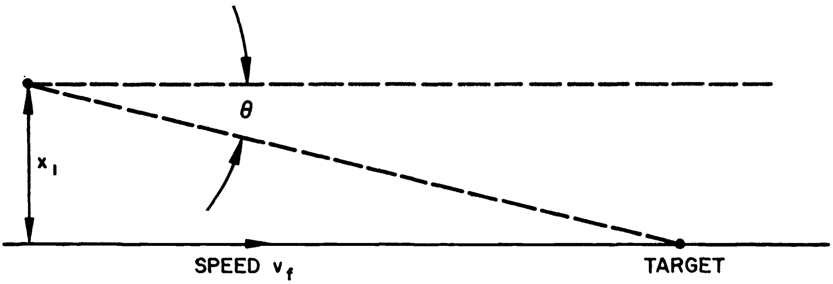


FIG. 4.1. *The one dimensional model*

4. Two simple examples and the computer results.

4.1. *Controlling the position and the velocity of a one-dimensional model.* Consider a space ship which is "homing" with constant velocity v_f on a massless planet. Let x_1 and x_2 be the transverse position and velocity deviations from a nominal orbit (see Fig. 4.1), and let

$$\dot{x}_1 = x_2, \tag{4.1}$$

$$\dot{x}_2 = u, \tag{4.2}$$

so that the free motion is uniform. It will be assumed that the variances of both the position and the velocity at the final time are specified. The initial error is to be only in velocity. However, for computational purposes, a small positional error is included. Hence

$$V(0) = \begin{bmatrix} v_{11}(0) & 0 \\ 0 & v_{22}(0) \end{bmatrix}, \tag{4.3}$$

where $v_{11}(0) \ll T^2 v_{22}(0)$. It is assumed that the information rate is purely positional and that the estimates of the transverse position are obtained by angle measurements at frequent intervals Δt with constant accuracy σ_ϵ. Hence

$$y_1 = \theta = \frac{x_1}{v_f(T - t)} + \epsilon_1(t), \tag{4.4}$$

and the information rate matrix is

$$\dot{I}(t) = \begin{bmatrix} \dfrac{1}{v_f^2 \sigma_\epsilon^2 \Delta t (T - t)^2} & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.5}$$

The product $v_f^2 \sigma_\epsilon^2 \Delta t$ may be related to a dimensionless *information rate* *parameter* k defined for this problem by

$$k = \frac{10^{-4} v_{22}(0) T}{v_f^2 \Delta t \sigma_\epsilon^2}.$$

This parameter compares the incoming information with the a priori information $(v_{22}(0))^{-1}$ about the initial velocity error. For the examples used in this paper, we have let $(v_{22}(0))^{1/2} = 100$ m/sec., $T = 10^6$ sec. and $k = 1$. Realistic values of k would be much higher and lead to earlier reduction of the predicted miss. For example, if $v_f = 3$ km/sec. and $\Delta t = 1$ hour, then $k = 1$ implies $\sigma_\epsilon = 0.32$ degree.

Using the information rate matrix given by (4.5), it is found that an analytical expression may be obtained for the covariance matrix $V(t)$. It can be readily verified that

$$V(t) = \frac{1}{\det W} \begin{bmatrix} w_{22}(t) & -w_{12}(t) \\ -w_{12}(t) & w_{11}(t) \end{bmatrix}, \tag{4.6}$$

where

$$\begin{aligned} w_{11}(t) &= \frac{1}{v_{11}(0)} + \frac{at}{T(T-t)}, \\ w_{12}(t) &= -\frac{t}{v_{11}(0)} + \frac{at}{T} + a \log \frac{T-t}{T}, \\ w_{22}(t) &= \frac{1}{v_{22}(0)} + \frac{t^2}{v_{11}(0)} + 2at - \frac{at^2}{T} + 2a(T-t) \log \frac{T-t}{T}, \end{aligned} \tag{4.7}$$

$$a = \frac{1}{v_f^2 \sigma_\epsilon^2 \Delta t},$$

and

$$\det W = w_{11}(t) w_{22}(t) - w_{12}^2(t). \tag{4.8}$$

Moreover,

$$V(T) = 0, \tag{4.9}$$

and

$$Z(t) = \begin{bmatrix} (T-t)^2 & T-t \\ T-t & 1 \end{bmatrix}. \tag{4.10}$$

Computation procedure and the numerical results. It can be shown that the adjoint variables are monotonically increasing functions of time if $\lambda_{12} > 0$. (λ_{11} and λ_{22} are always positive.) Since $\lambda_{12}(T) = 0$, we must let

$\lambda_{12}(0) < 0$ so that λ_{12} is negative at the time of turning on the control. Moreover, the control must be turned off at the time when λ_{12} reaches zero and not turned on again until possibly at the terminal time. It was shown in the previous section that an impulse may be applied at the final time if (3.13) is satisfied. In our case, this implies

$$\lambda_{22}^2(T)p_{22}(T) = 1. \tag{4.11}$$

It should be noted that an impulse at T brings down $p_{22}(T)$ and cannot change the values of λ_{12}, λ_{11}, and p_{11}. Using (3.23) and (3.24), we find, at time T,

$$\frac{dp_{22}}{de} = -2p_{22}\lambda_{22} \tag{4.12}$$

and

$$\frac{d\lambda_{22}}{de} = \lambda_{22}^2, \tag{4.13}$$

where de is the incremental effort due to the impulse. Using (4.12), (4.13) and the fact that (4.11) must be satisfied before and after application of the impulse, we find

$$\text{effort due to the impulse} = \frac{1}{\overline{\lambda_{22}}}\left(1 - \sqrt{\frac{\overset{+}{p_{22}}}{\overline{p_{22}}}}\right), \tag{4.14}$$

where $\overline{p_{22}}$ and $\overset{+}{p_{22}}$ denote the values of $E(x_2^2(T))$ immediately before and after the impulse respectively. Hence, if $\overset{+}{p_{22}} = 0$ (corresponding to perfect velocity control), then the additional effort required is $\sqrt{\overline{p_{22}}(T)}$. We shall assume that the desired $p_{22}(T) = 0$.

The actual computation proceeds as follows:

(1) Let $\lambda_{12}(0) = -1$ and guess $\lambda_{11}(0)$ and $\lambda_{22}(0)$.

(2) Integrate (3.12) until (3.15) is satisfied. This determines t_{on}.

(3) Use (3.13) to determine the value of Λ at t_{on}.

(4) Integrate along the surface by using (3.8), (3.9), and (3.16) until $\lambda_{12} = 0$.

(5) Turn off the control until T. This determines $P(T)$ and is a possible solution. But $p_{22}(T)$, in general, will not be zero. Note that $\Lambda(T)$ remains the same as at the time that the control was turned off.

(6) If (4.11) is satisfied, an impulse is applied at T to bring $p_{22}(T)$ to zero. The additional velocity required is $\sqrt{\overline{p_{22}}(T)}$.

(7) If (4.11) is not satisfied, we repeat the procedure again with a different guess of $\lambda_{11}(0)$ and $\lambda_{22}(0)$.

The results are given in Figs. 4.2–4.4 with the corresponding curves identified by the symbol ME2. Fig. 4.2 gives the plot of $\sqrt{p_{11}(T)}$ (which is

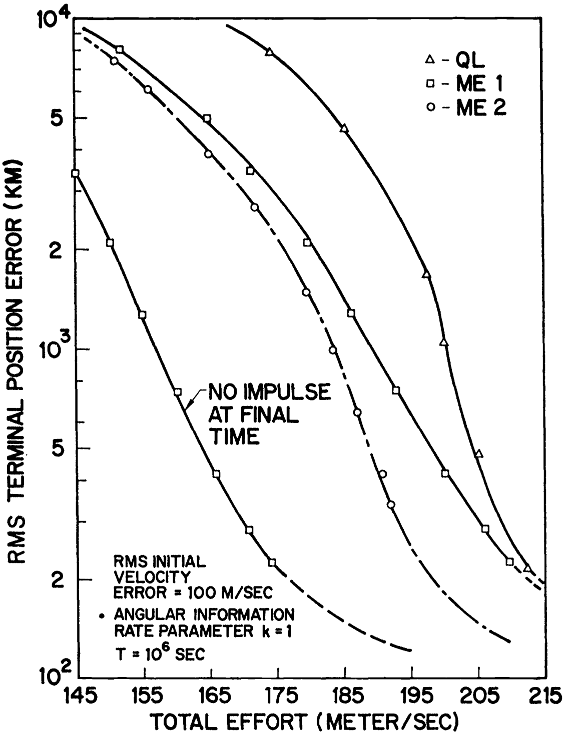


FIG. 4.2. *RMS terminal position error vs. total effort: position and velocity control*

the same as $\sqrt{a_{11}(T)}$ since $V(T) = 0$) versus the total effort. It is seen that most of the expended effort appears near the beginning of the trip and near the end of the trip when very high terminal accuracy is required. A typical plot of the history of $\sqrt{a_{22}(t)}$ versus time to go is given in Fig. 4.3 for the case where $\sqrt{a_{11}(T)} = 1530$ km. Note the period of no control and the impulse at the end. The corresponding total velocity required as a function of the time to go is shown in Fig. 4.4. The jump at T is due to the impulsive correction.

In order to get a "feeling" for these numbers, we include, in the same graph, some typical values obtained from other solutions. The two solutions we have used are the *quadratic loss* (to be denoted by QL) and the *minimum effort* for controlling only the final position (to be denoted by ME1).

QL: This is the problem of minimizing

$$E \int_0^T \| u(t) \|^2 dt = \int_0^T \operatorname{tr} PS'S\, dt \tag{4.15}$$

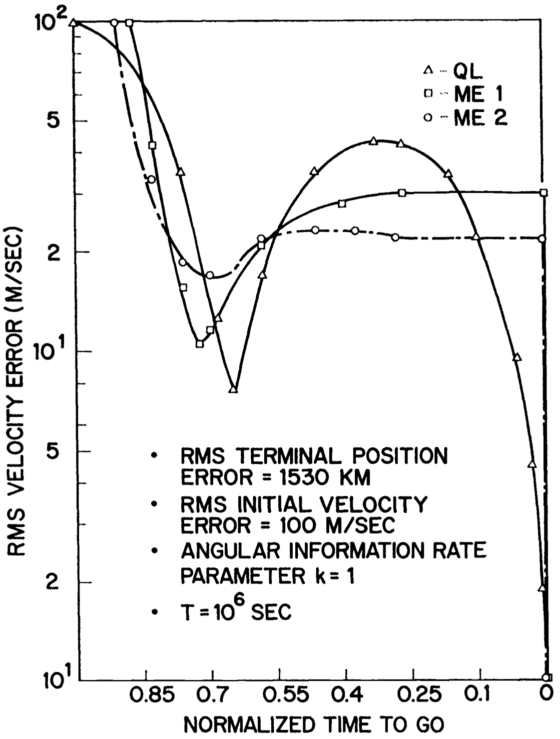


FIG. 4.3. *History of remaining velocity error vs. time to go; position and velocity control*

for a specified $P(T)$. The solution of this problem is well known (for example, see [7]). Let the solution be denoted by *. Then

$$S^* = G'\Phi'\Lambda^*, \tag{4.16}$$

$$\dot{\Lambda}^* = \Lambda^* Z \Lambda^*, \tag{4.17}$$

$$\dot{P}^* = -(Z\Lambda^* P^* + P^*\Lambda^* Z) + Q. \tag{4.18}$$

With the exception of $\phi(t)$, we see that this set of equations is the same as that given by (3.7)–(3.9) with $\operatorname{tr} P\Lambda Z\Lambda = 1$. However, here the problem is not singular. The solution can be obtained easily by integrating the adjoint equations backwards with an estimated value of $\Lambda^*(T)$. The off diagonal elements of $\Lambda^*(T)$ are zero and the diagonal elements of $\Lambda^*(T)$ are to be adjusted so that the prescribed values of $P_{ii}(T)$ are satisfied. To obtain the solution corresponding to the case that $p_{22}^*(T) = 0$, we let $\lambda_{22}^*(T) = \infty$. The results are also plotted in Figs. 4.2–4.4. The numerical values indicate that the difference between this solution and the optimal

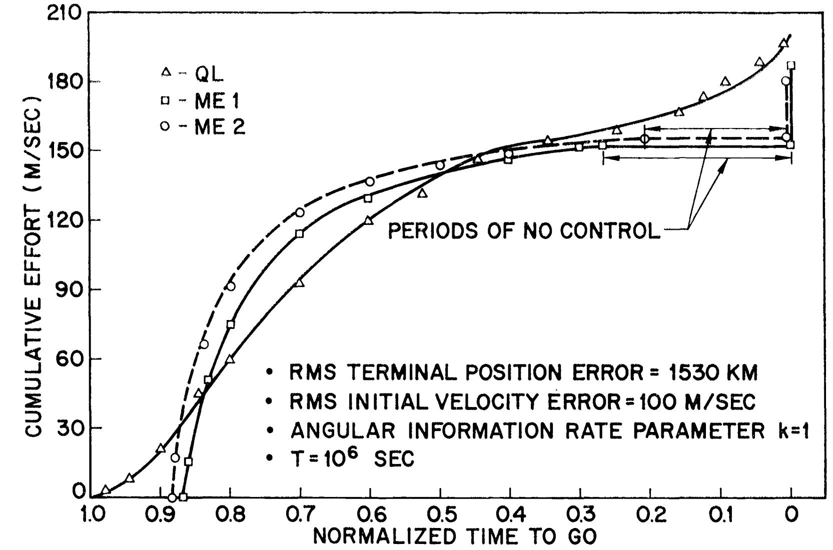


FIG. 4.4. *Cumulative effort vs. normalized time to go; position and velocity control*

solution developed in this paper in the total velocity requirement is about 10%.

ME1: This is the problem of minimizing the effort when only $p_{11}(T)$ is specified. It corresponds to the case of letting $\lambda_{12}(0) = \lambda_{22}(0) = 0$. In other words, we control the position to the specified rms value and turn off the control until T. An impulse is then added to bring $p_{22}(T)$ down to zero. In Fig. 4.2 we plot the results of $\sqrt{a_{11}(T)}$ versus the total effort with or without the final impulse. The amount of the additional velocity correction due to the impulse is, of course, $\sqrt{p_{22}(T)}$. Similar plots are given in Figs. 4.3 and 4.4. As expected, for the same terminal rms values, this design requires a little more effort than that required by controlling both components starting at $t = 0$.

4.2. *Controlling the two positions of two one-dimensional models.* This example considers the terminal phase of an interplanetary trip where both the in-plane and the out-of-plane terminal position components are to be independently controlled. It is assumed that the perturbed motions are decoupled and that each one moves in a uniform motion. We shall use the same information rate matrix as that used in the previous example and it will be further assumed that the information rate with respect to the two positions are independent.

The differential equations governing the adjoint variables are

$$\dot{\lambda}_{11}(t) = (T - t)^2 \lambda_{11}^2(t) \phi(t), \tag{4.19}$$

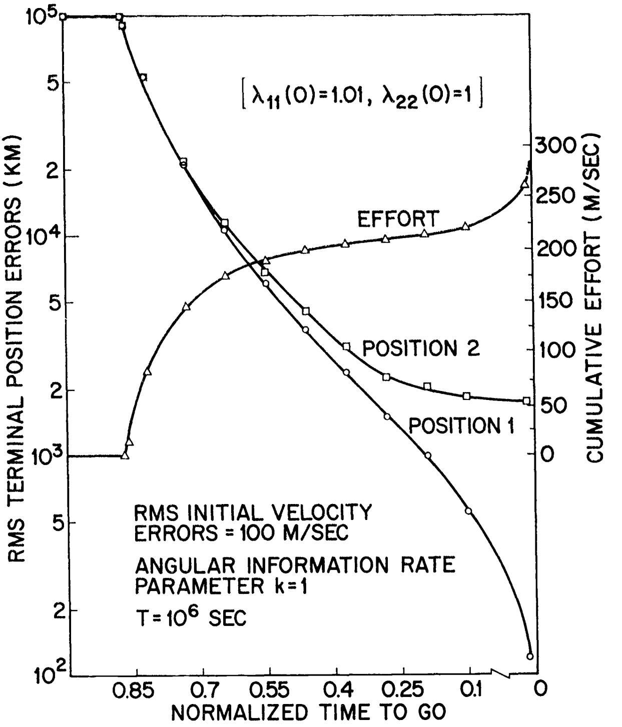


FIG. 4.5. *History of remaining position errors and cumulative effort vs. time to go; two position control*

$$\dot{\lambda}_{22}(t) = (T - t)^2\lambda_{22}^2(t)\phi(t), \tag{4.20}$$

and

$$\lambda_{12}(t) = p_{12}(t) = 0. \tag{4.21}$$

Equations (4.19) and (4.20) do not imply that the equations are decoupled. The coupling is introduced by the function $\phi(t)$. By letting $\lambda_{22}(0) = 1$, a family of solutions can be obtained for different values of $\lambda_{11}(0)$. A typical one corresponding to $\lambda_{11}(0) = 1.01$ is given in Fig. 4.5. It shows the plot of the history of $\sqrt{a_{11}(T)}$, $\sqrt{a_{22}(T)}$ and the effort versus the time to go. It is seen that the solution consists of an initial period of no control, followed by a period of continuous control and finally a period of no control at the end. The last statement is true since the control may be turned off when sufficient terminal accuracies have been obtained.

Case involving a gap in information rate. It was stated in the previous section that in the event that the computed $\phi(t) < 0$, $t \in (t_{\text{on}}, t_{\text{off}})$, then there will exist intervals within $(t_{\text{on}}, t_{\text{off}})$ such that the control is turned

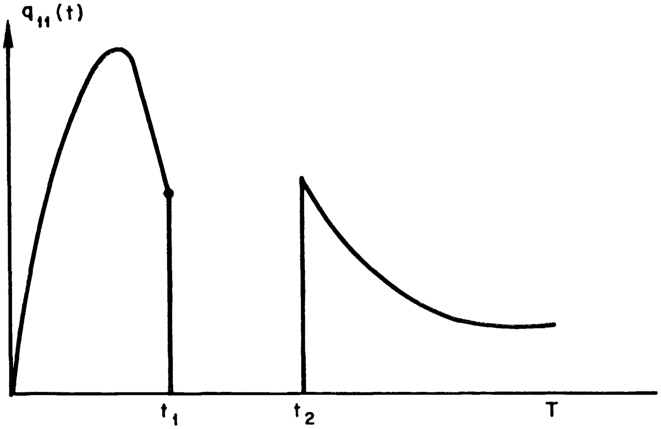


FIG. 4.6 *Uncertainty improvement rate with information gap*

off. This occurs, for instance, when the information rate suddenly increases. A computation procedure was described in the previous section by which the intervals of no control can be found. For purpose of illustration, we assume that the information vanishes over the interval (t_1, t_2) and suddenly increases at t_2. In particular, we choose $t_1 = 0.27T$ and $t_2 = 0.45T$.

Now the two elements of Q (q_{11} and q_{22}) are equal and have the general shape as shown in Fig. 4.6. It is clear that the control can not follow the sharp rise of q_{ii} at t_2, i.e., $\phi(t_2) < 0$. Therefore, the control must be turned off before or immediately after t_1. Numerical solution from a trial run indicates that the control is to be turned off immediately after t_1. Since Q is discontinuous at t_1, it follows from the reasoning given in the previous section that an impulse may be applied at t_1. This is indeed the case. The amount of the impulse (which is not a full correction) is determined by the condition that the adjoint variables after the correction must be the same as at the time when the control is turned on again. The amount of the drop (or jump) in P (or Λ) can be determined by integrating with respect to the effort at t_1 using (3.23) and (3.24) which in our case can be written as

$$\frac{dp_{ii}}{de} = -2(T - t_1)^2\lambda_{ii}p_{ii}, \tag{4.22}$$

$$\frac{d\lambda_{ii}}{de} = (T - t_1)^2\lambda_{ii}^2, \qquad i = 1, 2. \tag{4.23}$$

Let the superscripts $^-$ and $^+$ denote the times immediately before and after the impulse respectively. Direct integration of (4.23) yields

$$\text{effort due to the impulse} = \frac{1}{(T - t_1)^2}\left(\frac{1}{\lambda_{ii}^-} - \frac{1}{\lambda_{ii}^+}\right). \tag{4.24}$$

Dividing (4.23) and (4.22) shows

$$\frac{dp_{ii}}{2p_{ii}} = -\frac{d\lambda_{ii}}{\lambda_{ii}}, \tag{4.25}$$

which can be integrated to give

$$(p_{ii}\lambda_{ii}^2)^- = (p_{ii}\lambda_{ii}^2)^+. \tag{4.26}$$

Equation (4.26) shows, as expected, that (3.13) is satisfied during the impulse.

Prior to t_1, the computation remains the same as before. At t_1, we proceed as follows. Let $d = \frac{\lambda_{11}}{\lambda_{22}}$.

(1) Assume an effort due to the impulse and compute $\overset{+}{\lambda_{11}}$, $\overset{+}{\lambda_{22}}$, and $d(\overset{+}{t_1})$ from (4.24).

(2) Use (4.26) to determine $\overset{+}{p_{11}}$ and $\overset{+}{p_{22}}$.

(3) Integrate the equations for p_{ii} with $S = 0$ until (3.15) is satisfied.

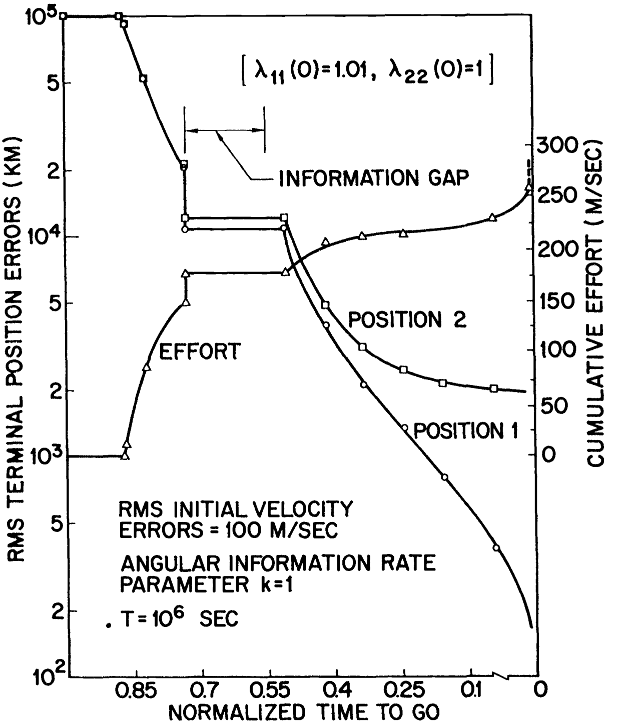


FIG. 4.7. *History of remaining position errors and cumulative effort vs. time to go; two position control*

This determines t_1'. Use is then made of (3.13) to determine λ_{ii} and d at t_1'.

(4) If $d(t_1^+) \neq d(t_1')$, we repeat the procedure again by assuming a different effort.

The results are given in Fig. 4.7 for the case $\lambda_{11}(0) = 1.01$. The discontinuities at t_1 correspond to the impulsive correction. It is seen that t_1' is greater than t_2 which agrees with the intuitive reasoning that it is necessary to let the information catch up after an interval of no observation.

It is of interest to note that the quadratic loss solution corresponding to this particular example is completely decoupled. In other words, specification of the variance of the terminal in-plane position does not effect the solution of the out-of-plane component and vice versa. The coupling, in our case, is introduced by the loss function.

Acknowledgement. The authors wish to thank Dr. Y. C. Ho for his many helpful suggestions. The work was supported by NASA under contract NAS-1-3777 and in part by the Lockheed Independent Research Program.

Appendix. This section specializes the results derived in §3 to the control of only one terminal miss. Without loss of generality, it will be assumed that the particular terminal miss we wish to control is the final uncertainty in the position. It will be shown that the solution in this case is unique. This result was first obtained by Breakwell and Striebel [4] by applying Green's Theorem.

For the case of one terminal miss, $p = 1$ and H is a row matrix of $1 \times n$. Let the scalar $H\Phi GG'\Phi'H'$ be denoted by D_v^2, which is the sensitivity of the miss distance to a change of velocity in the direction of the correction. Equations (3.13) and (3.15) become

$$p_{11}\lambda_{11}^2 D_v^2 = 1 \tag{A.1}$$

and

$$\lambda_{11}^2(2D_v\dot{D}_v p_{11} + q_{11}D_v^2) = 0, \tag{A.2}$$

respectively. It follows from (A.2) that t_{on} is determined by the equation

$$\int_0^{t_{\text{on}}} q_{11}(t)\,dt = p_{11}^*(t_{\text{on}}), \tag{A.3}$$

where[†]

$$p_{11}^*(t) = -\frac{D_v q_{11}}{2\dot{D}_v}. \tag{A.4}$$

[†] This is the critical curve defined in [4].

Also, for a given $p_{11}(T)$, t_{off} is determined by

$$p_{11}^*(t_{\text{off}}) + \int_{t_{\text{off}}}^{T} q_{11}(t)\,dt = p_{11}(T). \tag{A.5}$$

Moreover, the optimal solution must follow the critical curve defined by (A.3) if

$$\phi(t) = p_{11}^*\dot{\lambda}_{11} > 0, \quad t \in (t_{\text{on}}, t_{\text{off}}). \tag{A.6}$$

This can be seen as follows. Assume (A.6) is true. It can be easily verified that this implies

$$\dot{p}_{11}^* < q_{11}, \quad t \in (t_{\text{on}}, t_{\text{off}}). \tag{A.7}$$

Suppose for some t' where $t_{\text{on}} < t' < t_{\text{off}}$, we leave the critical curve. Then the control must be turned off and for $t > t'$,

$$p_{11}(t) = p_{11}^*(t') + \int_{t'}^{t} q_{11}(s)\,ds, \tag{A.8}$$

which by (A.7) is greater than $p_{11}^*(t)$. Hence, the given terminal $p_{11}(T)$ can not be satisfied. In other words, we cannot come back to the critical curve after leaving it. This establishes our assertion.

Suppose (A.6) is not satisfied. Then there exists an interval within $(t_{\text{on}}, t_{\text{off}})$ such that the control must be turned off. This corresponds to the case of an unusual increase in the information rate considered in [4]. Let t_a and t_b be the times of turning off and on respectively. Since the adjoint variable must remain constant during the time that the control is off, we see from (A.1) that

$$p_{11}(t_a)D_v^2(t_a) = p_{11}(t_b)D_v^2(t_b). \tag{A.9}$$

Moreover,

$$a_{11}(t_a) = a_{11}(t_b), \tag{A.10}$$

which is obvious since $a_{11}(t)$ is the actual terminal miss if no control is applied after t. Equations (A.9) and (A.10) provide sufficient conditions for determining the times t_a and t_b. It is of interest to note that in the case of the control of only the terminal velocity, the optimal solution, according to our theory, is an impulse at the final time. This solution is reasonable since the effort necessary to nullify the velocity error remains constant in $(0, T)$ and hence the optimal solution is the one in which all the information is collected before applying the control.

REFERENCES

[1] R. H. Battin, *A statistical optimizing navigation procedure for space flights*, ARS J., (1962), pp. 1681–1697.

[2] J. V. Breakwell, *The spacing of corrective thrust in interplanetary navigation*, Advances in Astronautical Sciences, vol. 7, 1961, pp. 219–235.

[3] G. L. Smith, *Multivariable linear filter theory applied to space vehicle guidance*, presented at the SIAM Meeting on Control, Cambridge, Massachusetts, October, 1962.

[4] J. V. Breakwell and C. T. Striebel, *Minimum effort control in interplanetary guidance*, this Journal, 1965, to appear; also presented at IAS Meeting, New York, 1963, Preprint 63–80.

[5] G. Leitmann, *Optimization Techniques with Applications to Aerospace Systems*, Academic Press, New York, 1962, Chaps. 3 and 4.

[6] H. Robbins and J. Pitman, *Application of the method of mixtures to quadratic forms in normal variates*, Ann. Math. Statist., 20 (1949), pp. 552–560.

[7] F. Tung, *Linear control theory applied to interplanetary guidance*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 82–89.

[8] R. E. Kalman, *New methods and results in linear prediction and filtering theory*, RIAS Technical Report 61–1, RIAS, Baltimore.

[9] C. T. Striebel, *Sufficient statistics in the optimum control of stochastic systems*, (private communication), to be published.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

SOME MATHEMATICAL THEORY OF THE PENALTY METHOD FOR SOLVING OPTIMUM CONTROL PROBLEMS*

KIYOHISA OKAMURA†

Abstract. The penalty method is a powerful technique for solving the optimum control problems involving systems subject to holonomic side constraints. In the usual calculus of variations, the above problems are formulated in consideration of the Weierstrass-Erdmann corner conditions which add considerable complexity in practice. In the penalty method, however, the side constraints are eliminated by introducing a sequence of approximate formulations. Thus the Weierstrass-Erdmann corner conditions need not be checked.

When the penalty method is applied in the ordinary calculus the sequence of approximate formulations is proved to be equivalent to the original formulation in the limiting case. However, no mathematical rigor has been claimed when the penalty method is applied to the variational problems.

The author establishes, in this paper, some mathematical basis for the penalty method applied in the calculus of variations, particularly optimum control problems.

Introduction. The state of a physical system in control problems may be represented by a real n-dimensional vector $x(t) \equiv (x_1(t), \cdots, x_n(t))$, called the *state vector*. This vector can also be considered as a point in the Euclidean n-space, called the *state space*, at the time t. As this point moves from one point to another, during the time interval $0 \leqq t \leqq T$, it moves according to the following differential equations[1]

$$(1) \qquad \frac{dx_i}{dt} = f_i(x, u), \qquad i = 1, \cdots, n,$$

where $u(t) \equiv (u_1(t), \cdots, u_r(t))$. The plot of this moving point is called the *trajectory* of the system. Here, the r-dimensional real vector $u(t)$ is called the *control vector*, or simply the *control*. The control u can be considered as a point in the Euclidean r-space, called the *control space*. The function $u(t)$ is a *well-defined* piecewise continuous function of t (Appendix 1). Each f_i is a real function of x and u, and[2] $f_i \in \mathrm{Lip}(x, u)$. The initial and final point of the trajectory, respectively denoted by x^0 and x^T, generally lie on the specified manifolds, called the *initial* and *final* manifolds, and designated as S_0 and S_T. The time T may be either specified or unspecified.[3] The set of all controls which enable the state to move

* Received by the editors August 10, 1964, and in revised form October 21, 1964.

† Allison Division, General Motors Corporation, Indianapolis, Indiana.

[1] Non-autonomous cases are included in this representation [1, p. 59].

[2] This means that f_i satisfies the Lipschitz conditions with respect to the arguments in the parenthesis in the domain under consideration.

[3] A problem with the specified final time may be formulated as a problem with unspecified final time [2].

from the initial point to the final point is called the *admissible control set in wide sense* and designated by Ω. A control u such that

$$u \in \Omega \tag{2}$$

is called an *admissible control.* Throughout this paper the existence of an admissible control is assumed and only an admissible control is considered. If T is not specified, then the time when some condition $x \in S_T$ is attained determines T itself. The system is considered to be subject to the following constraints, called the *side constraints*:

$$g_k(x, u) \leqq 0, \qquad k = 1, \cdots, m, \tag{3}$$

where the functions g_k are real functions of x and u, and $g_k \in \mathrm{Lip}(x, u)$. The set of all admissible controls, which always enable the system to satisfy the above constraints, is called the *admissible control set in narrow sense* and denoted by Ω^*, i.e.,

$$\Omega^* \triangleq \Omega \cap \{u : g_k(x, u) \leqq 0, k = 1, \cdots, m\}. \tag{4}$$

It is assumed that the set Ω^* is not empty.

The problem of the optimum control treated in this paper is to minimize a cost functional described by

$$J \triangleq \int_0^T h(x, u)\, dt, \quad \text{for} \quad u \in \Omega^*, \tag{5}$$

where $h(x, u)$ is a bounded real function of x and u, and $h \in \mathrm{Lip}(x, u)$.

This type of problem was extensively treated by Berkovitz [3]. Some special cases were treated by Pontryagin et al. [1] and Chang [4]. From the viewpoint of applications of the theory to practical problems however, these formulations and solutions are too complicated, since the logic in programming these theories is quite involved. Special mathematical difficulties in these methods lie in the fact that the Weierstrass-Erdmann corner conditions must be satisfied when the trajectory reaches or leaves the boundary of the constraint inequalities.

Recently an approximation method, the *penalty method,* has been found which alleviates the above difficulty. In this method the original problem is replaced by an approximate one which eliminates an explicit evaluation of the constraints (3). Thus, the Weierstrass-Erdmann corner conditions need not be checked. One may anticipate that an approximate solution may be obtained as close to the exact solution as desired.

The original formulation of the penalty method is due to Courant [5]. Moser [5] gave a mathematical basis to this method as applied to the minimization problem in the ordinary calculus. Extensions of the penalty method to the field of the calculus of variations have also been made.

Among them are the works of Kelley [6], Ostrovskii [7], and others. However, none of them guarantee the mathematical rigor in the penalty method when applied to the calculus of variations problems.

In this paper the mathematical rigor necessary to establish the penalty method applied to the calculus of variations is developed, with special emphasis being placed on the convergence problem.

2. The penalty method in the ordinary calculus. The penalty method in the ordinary calculus was introduced by Courant [5]. This method is briefly explained below.

Let the problem A be defined as follows:

Find a point P at which a given real function of P, designated by $\Phi(P)$, is minimum consistent with the side constraints

$$\Psi(P) = 0, \tag{6}$$

with $\Psi(P)$ a given nonnegative real function of P.

Let the problem A_k be defined as follows:

Find a point P_k at which the function $\Phi_k(P_k)$, represented by the relation

$$\Phi_k(P_k) \overset{\Delta}{=} \Phi(P_k) + k\Psi(P_k), \tag{7}$$

where k is a real positive quantity, is minimum. The penalty method is based on the following theorem.

THE APPROXIMATION THEOREM IN THE PENALTY METHOD[4].

$$A_k \to A \quad as \quad k \to \infty. \tag{8}$$

The proof is given by Moser [5]. Here, only the intuitive explanation will be presented.

The second term on the right hand side of (7) is an index of the violation to constraint (6), since $k\Psi(P_k)$ is zero when (6) holds and is positive when (6) does not hold. As the constant k increases the index of violation also increases without bound. Hence the effort to minimize the function $\Phi_k(P_k)$ is primarily focused on minimization of this index. Having minimized this function, the first term is minimized. As k becomes greater the function $\Psi(P_k)$ must approach zero so that the term $k\Psi(P_k)$ might be finite. Thus it is seen, in the solution of the problem A_k, that as k tends to infinity constraint (6) is satisfied and the function $\Phi(P_k)$ is minimized, i.e., the problem A_k is equivalent to the problem A in the limiting case.

3. Reformulation of the original problem in terms of the penalty method. In this section, the original problem is reformulated by employing and ex-

[4] Actually several assumptions are made for this theorem but not listed here for the sake of brevity. The reader is referred to [5] for details.

tending the penalty method explained in the previous section. The *penalty functions*, $p_k(g_k)$, $k = 1, \cdots, m$, are defined by the following:

(a) $p_k(g_k)$ is a real, continous and nondecreasing function and defined for $g_k(x, u) \in (-\infty, \infty)$;

(b) $$p_k(g_k)\begin{cases} > 0 & \text{for} \quad g_k(x, u) > 0, \\ = 0 & \text{for} \quad g_k(x, u) \leqq 0. \end{cases}$$

From properties (a) and (b) together with (5) the following property is derived:

(c) $$p_k(g_k) \equiv 0 \quad \text{if and only if} \quad u \in \Omega^*.$$

An example of such function is:

$$(9)\quad \begin{array}{lll} \text{(a)} & p_k(g_k) = (g_k)^2 & \text{for} \quad g_k(x, u) > 0, \\ \text{(b)} & p_k(g_k) = 0 & \text{for} \quad g_k(x, u) \leqq 0. \end{array}$$

The above example is the one used by Kelley [6]. As seen above, if u is determined for $t \geqq 0$, then $x(t)$, $g_k(x, u)$, and $p_k(g_k)$ are successively determined. For this reason the penalty functions sometimes are written as $p_k(x, u)$, $p_k(u)$, or $p_k(t)$. Real nonnegative constants σ_k, $k = 1, \cdots, m$, called the *penalty weighting coefficients* are introduced. The inner product of the two vectors $\sigma \equiv (\sigma_1, \cdots, \sigma_m)$ and $p \equiv (p_1, \cdots, p_m)$ is defined by

$$(10)\qquad \pi(\sigma, u) = \sum_{k=1}^{m} \sigma_k p_k(u).$$

This representation is called the *penalty*. We consider a sequence $\{\sigma^\mu\}$ for $\mu = 1, 2, 3, \cdots$, such that

$$(11)\qquad \sigma_k^\mu < \sigma_k^\nu \quad \text{for} \quad \mu < \nu, \qquad k = 1, \cdots, m,$$

with

$$(12)\qquad \sigma_k^\mu \to \infty \quad \text{as} \quad \mu \to \infty, \qquad k = 1, \cdots, m;$$

and a sequence of functionals $\{I_\mu\}$,

$$(13)\qquad I_\mu \triangleq \int_0^T \left\{ h(x, u) + \sum_{k=1}^{m} \sigma_k^\mu p_k(u) \right\} dt$$

$$(14)\qquad = J(u) + \sum_{k=1}^{m} \int_0^T \sigma_k^\mu p_k(u)\, dt, \qquad \mu = 1, 2, 3, \cdots,$$

where constraints (3) have been eliminated. We call I_μ the *charged cost functional*. The revised optimum control problem is stated as:

Minimize the charged cost functional I_μ for $u \in \Omega$.

In the succeeding sections we prove that the revised optimum control problem is, under some conditions, equivalent to the original one as $\mu \rightarrow \infty$.

4. Preliminary remarks and the first convergence theorem. In this section we discuss the mathematical properties of the control, the penalty functions and the charged cost functional. First, notation and several definitions will be introduced.

$$J^* \triangleq J(u^*) \triangleq \inf_{u \in \Omega^*} J(u). \tag{15}$$

$$I_\mu^* \triangleq I_\mu(u^{\mu *}) \triangleq \inf_{u \in \Omega} I_\mu(u), \qquad \mu = 1, 2, 3, \cdots . \tag{16}$$

The existence of u^*, $u^{\mu *}$, J^*, I_μ^* and the corresponding final times T^* and T_μ^* is always assumed. As seen above, an asterisk $(^*)$ stands for *optimal.*

Next, the norm in the control space is given by

$$\| u(t) \|_T \triangleq \sup_{j \in [1, \cdots, r]} \int_0^T | u_j(t) | \, dt. \tag{17}$$

Lemma 1. *$\{I_\mu^*\}$ is a nondecreasing sequence, i.e.,*

$$I_\mu^* \leqq I_\nu^* \quad \textit{for} \quad \mu < \nu. \tag{18}$$

Proof. The proof is accomplished by contradiction. Suppose the contrary to the lemma; then there exist positive integers μ and ν such that

$$I_\mu^* > I_\nu^* \quad \text{for} \quad \mu < \nu. \tag{19}$$

Substituting (14) and (16) into the right hand side of the above inequality we obtain

$$I_\mu^* > J(u^{\nu *}) + \sum_{k=1}^{m} \int_0^{T_\nu^*} \sigma_k^{\nu} p_k(u^{\nu *}) \, dt, \tag{20}$$

where T_ν^* is the final time determined by the control $u^{\nu *}$. Equation (11) and the property (b) of the penalty function yield

$$\sum_{k=1}^{m} \int_0^{T_\nu^*} \sigma_k^{\mu} p_k(u^{\nu *}) \, dt < \sum_{k=1}^{m} \int_0^{T_\nu^*} \sigma_k^{\nu} p_k(u^{\nu *}) \, dt. \tag{21}$$

Substituting this inequality into (19) results in

$$I_\mu^* > J(u^{\nu *}) + \sum_{k=1}^{m} \int_0^{T_\nu^*} \sigma_k^{\mu} p_k(u^{\nu *}) \, dt. \tag{22}$$

From the definition of $I_\mu(u)$, the right hand side of (22) is $I_\mu(u^{\nu *})$, i.e.,

$$I_\mu^* > I_\mu(u^{\nu *}). \tag{23}$$

However, this result violates the definition of $I_\mu{}^*$ represented by (16). Thus the lemma has been proved. As before, it is sufficient to consider only the case where the final time is unspecified.

LEMMA 2.

$$I_\mu{}^* \leqq J^*, \qquad \mu = 1, 2, 3, \cdots . \tag{24}$$

Proof. From the property (c) of the penalty function and (14) and (16), the following equation is obtained.

$$J^* = J(u^*) + \sum_{k=1}^{m} \int_0^{T^*} \sigma_k{}^\mu p_k(u^*)\, dt, \qquad \mu = 1, 2, 3, \cdots . \tag{25}$$

The right hand side of (25) is, by definition, $I_\mu(u^*)$. Hence it follows that

$$J^* = I_\mu(u^*) \geqq I_\mu{}^*, \qquad \mu = 1, 2, 3, \cdots , \tag{26}$$

where the equal sign is taken when $u^* \equiv u^{\mu *}$. Thus the lemma has been proved.

From the above lemma the following is derived.

COROLLARY 1.

$$J^* \geqq J(u^{\mu *}) \quad for \quad \mu = 1, 2, 3, \cdots . \tag{27}$$

Proof. By definition,

$$I_\mu{}^* = J(u^{\mu *}) + \sum_{k=1}^{m} \int_0^{T_\mu{}^*} \sigma_k{}^\mu p_k(u^{\mu *})\, dt, \qquad \mu = 1, 2, 3, \cdots .$$

Since the summation in the above relation is nonnegative it follows that

$$I_\mu{}^* \geqq J(u^{\mu *}), \qquad \mu = 1, 2, 3, \cdots . \tag{28}$$

Inequalities (24) and (28) yield the corollary.

By Lemmas 1 and 2 the following theorem directly follows.

THE FIRST CONVERGENCE THEOREM.

$$I_\mu{}^* \textit{ converges as } \mu \longrightarrow \infty. \tag{29}$$

We denote this limit by $I_\infty{}^*$, i.e.,

$$I_\infty{}^* = \lim_{\mu \to \infty} I_\mu{}^*. \tag{30}$$

In the next section we prove that the charged cost functional, as well as the cost functional, approaches the optimum functional for the original problem in the limiting case.

5. The second convergence theorem and the optimal control. In the last section it was proved that the limit of the sequence of the optimum charged cost functionals exists. In this section we shall further investigate the problem of this convergence. Here we adopt the following hypothesis.

Hypothesis. There exist a control $u^{\infty *} \in \Omega$, an initial state $x^{\infty *}(0) \in S_0$, and the corresponding final time T_∞^* satisfying the following conditions: For an arbitrary $\epsilon > 0$ an $N > 0$ exists such that

$$(31)\quad \begin{aligned} &(a) \qquad | x^{\mu *}(0) - x^{\infty *}(0) | < \epsilon, \\ &(b) \qquad | T_\mu^* - T_\infty^* | < \epsilon, \end{aligned}$$

and

$$(32)\qquad \left\| u^{\infty *}(t) - u^{\mu *}\left(\frac{T_\mu^*}{T_\infty^*} t\right) \right\|_{T_\infty *} < \epsilon,$$

for all $\mu > N$.

Remark 1. The first part of the hypothesis represented by (31) means that the initial state $x^{\mu *}(0) \in S_0$ and the final time T_μ^* converge as $\mu \rightarrow \infty$. For the case where the initial state is completely specified, (31a) is not required. Similarly, (31b) is not necessary when the final time is specified. The meaning of the second part of the hypothesis represented by (32) is explained below with the help of Fig. 1. An example of $u_j^{\infty *}(t)$ is shown by $A_0'A_1'A_3' - A_3' + A_T'$ in Fig. 1(a) and an example of $u_j^{\mu *}(t)$ by $B_0B_1B_2 - B_2 + B_T$ in Fig. 1(b). The domain $[0, T_\mu^*]$ of $u_j^{\mu *}(t)$ is mapped into the domain $[0, T_\infty^*]$ by a linear transformation $(T_\mu^*/T_\infty^*)t$. By this transformation the control $u_j^{\mu *}(t)$ becomes $u_j^{\mu *}(T_\mu^* t/T_\infty^*)$ which is plotted as $A_0A_1A_3 - A_3 + A_T$ in Fig. 1(b). Thus we can compare the controls $u_j^{\mu *}$ and $u_j^{\infty *}$ in the same domain. The number and locations of discontinuities for $u_j^{\infty *}(t)$ and $u_j^{\mu *}(T_\mu^* t/T_\infty^*)$ do not necessarily coincide. The shaded area in Fig. 1(b) is given by

$$\int_0^{T_\infty *} \left| u_j^{\mu *}\left(\frac{T_\mu^*}{T_\infty^*} t\right) - u_j^{\infty *}(t) \right| dt.$$

Therefore, by definition, the norm represented by the left hand side of (32) is the maximum of the shaded area. Hence (32) means that the shaded area for each j approaches zero as $\mu \rightarrow \infty$.

Remark 2. The relationships between the introduced quantities, $u^{\infty *}(t)$ and $x^{\infty *}(0)$, and the charged cost functional I_∞^* have not been stated. These relationships shall be investigated in this section.

Remark 3. In the following, $u^{\mu *}$ stands for $u^{\mu *}(t)$, not for $u^{\mu *}(T_\mu^* t/T_\infty^*)$.

First, we prove that the control $u^{\infty *}$ is admissible in narrow sense.

LEMMA 3.

$$(33)\qquad u^{\infty *} \in \Omega^*.$$

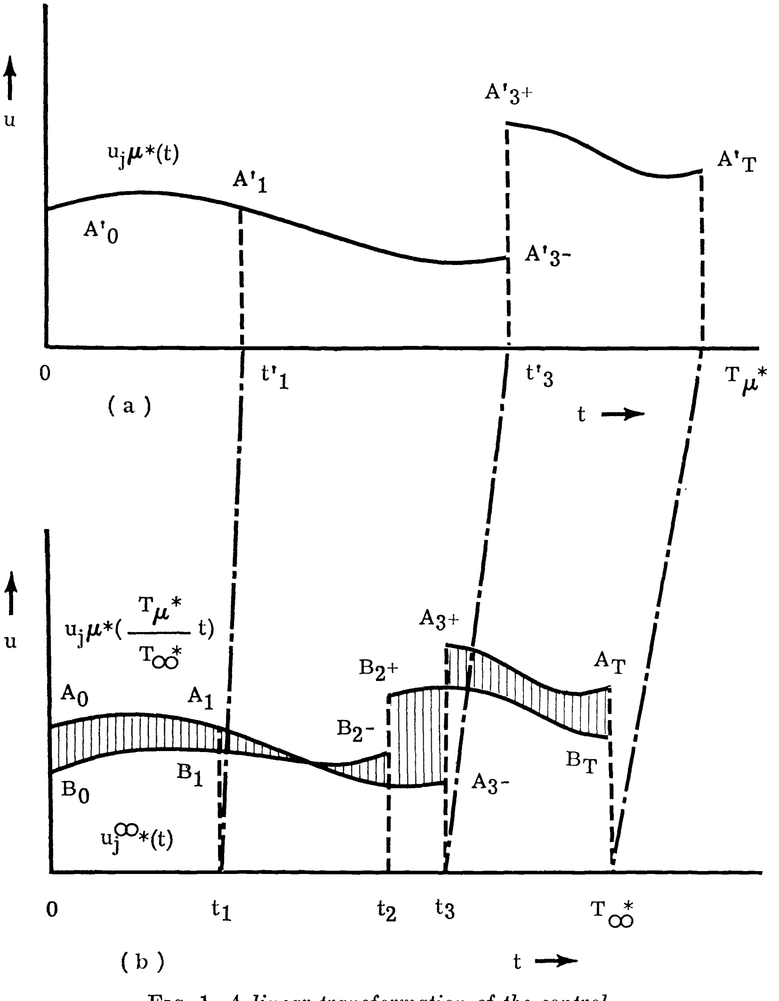


FIG. 1. *A linear transformation of the control*

Proof. Since the function $h(x, u)$ is bounded and the time T_μ^* is finite, $J(u^{\mu*})$ is bounded. Hence, referring to Lemma 2, we find that the corresponding integrated penalty is bounded, i.e.,

$$(34) \qquad 0 \leqq \sum_{k=1}^{m} \int_0^{T_\mu^*} \sigma_k^\mu p_k(x^{\mu*}, u^{\mu*})\, dt \leqq M, \qquad \mu = 1, 2, 3, \cdots,$$

with M a positive constant. Since for each k the integrand in (34) is nonnegative the inequalities (34) reduce to

$$(35)\quad 0 \leqq \sigma_k^{\mu} \int_0^{T_\mu *} p_k(x^{\mu *}, u^{\mu *})\,dt \leqq M, \qquad k = 1, \cdots, m;\ \mu = 1, 2, 3, \cdots.$$

Dividing all sides of the inequalities (35) by σ_k^{μ} and using the definitions (64) and (65) under the linear transformation in Appendix 2, we can rewrite (35) as

$$(36)\quad 0 \leqq \int_0^{T_\infty *} p_k(w^{\mu}, v^{\mu})\,dt \leqq \frac{M}{\sigma_k^{\mu}}, \qquad k = 1, \cdots, m;\ \mu = 1, 2, 3, \cdots.$$

Since the penalty functions are nonnegative there exist some constants ρ_k such that

$$(37)\qquad \int_0^{T_\infty *} p_k(x^{\infty *}, u^{\infty *})\,dt = \rho_k \geqq 0, \qquad k = 1, \cdots, m.$$

From (36) and (37) it follows that

$$(38)\quad \rho_k \leqq \frac{M}{\sigma_k^{\mu}} + \int_0^{T_\infty *} \{p_k(x^{\infty *}, u^{\infty *}) - p_k(w^{\mu}, v^{\mu})\}\,dt, \qquad k = 1, \cdots, m.$$

Strengthening each of the above inequalities by taking the absolute value of the integrand and changing the order of two terms in the integrand, we have

$$(39)\quad \rho_k \leqq \frac{M}{\sigma_k^{\mu}} + \int_0^{T_\infty *} |\,p_k(w^{\mu}, v^{\mu}) - p_k(x^{\infty *}, u^{\infty *})\,|\,dt, \qquad k = 1, \cdots, m.$$

Substituting (73) of Appendix 3 in the above integral we obtain

$$(40)\qquad 0 \leqq \rho_k \leqq \frac{M}{\sigma_k^{\mu}} + K_2\epsilon, \qquad k = 1, \cdots, m,$$

where the quantities μ and ϵ are the same as specified in the hypothesis. Since the above inequalities are valid for all $\mu > N$ consider a μ such that

$$(41)\qquad \frac{1}{\sigma_k^{\mu}} < \epsilon, \qquad k = 1, \cdots, m.$$

The existence of such σ_k^{μ} is guaranteed by the definition of σ_k^{μ}. Substituting (41) into (40) we obtain

$$(42)\qquad 0 \leqq \rho_k < (M + K_2)\epsilon, \qquad k = 1, \cdots, m.$$

Since the constants M and K_2 are positive finite and the positive quantity ϵ is arbitrarily chosen, (42) holds if and only if

$$(43)\qquad \rho_k = 0, \qquad k = 1, \cdots, m.$$

Substituting (43) into (37) we have

$$\int_0^{T_\infty^*} p_k(x^{\infty *}, u^{\infty *})\, dt = 0, \qquad k = 1, \cdots, m. \tag{44}$$

Since p_k are continuous with respect to $x^{\infty *}$ and $u^{\infty *}$, and since $u^{\infty *}$ is a well-defined piecewise continuous function of t, p_k are also well-defined piecewise continuous functions of t. Hence (44) implies that

$$p_k(x^{\infty *}, u^{\infty *}) \equiv 0, \qquad 0 \leqq t \leqq T_\infty^*, \qquad k = 1, \cdots, m. \tag{45}$$

This further implies (33) because of the property (c) for the penalty functions.

Corollary 2.

$$I_\mu(u^{\infty *}) \equiv J(u^{\infty *}), \qquad \mu = 1, 2, 3, \cdots . \tag{46}$$

Proof. The definition of I_μ and (44) directly yield this corollary.

Next we prove the convergence of the cost functional from the control that optimizes the charged cost functional.

Lemma 4.

$$J(u^{\mu *}) \rightarrow J(u^{\infty *}) \quad \textit{as} \quad \mu \rightarrow \infty. \tag{47}$$

Proof. The proof is similar to that for Lemma 3. Using the definition of cost functional we obtain

$$| J(u^{\mu *}) - J(u^{\infty *}) | \equiv \left| \int_0^{T_\mu^*} h(x^{\mu *}, u^{\mu *})\, dt - \int_0^{T_\infty^*} h(x^{\infty *}, u^{\infty *})\, dt \right|,$$

and again applying the linear transformation in Appendix 2,

$$\begin{aligned} | J(u^{\mu *}) - J(u^{\infty *}) | &= \left| \int_0^{T_\infty^*} h(w^\mu, v^\mu) - h(x^{\infty *}, u^{\infty *})\, dt \right| \\ &\leqq \int_0^{T_\infty^*} | h(w^\mu, v^\mu) - h(x^{\infty *}, u^{\infty *}) |\, dt. \end{aligned} \tag{48}$$

Employing the Hypothesis, and the notations used in it, and referring to (75), we find that

$$| J(u^{\mu *}) - J(u^{\infty *}) | < K_4 \epsilon. \tag{49}$$

Thus we have shown that for an $\epsilon > 0$ there exists $N > 0$ such that (49) holds. Since K_4 is a finite positive constant the above inequality is enough to prove the lemma.

The second convergence theorem.

$$J(u^{\mu *}) \rightarrow J^*, \tag{50}$$

and

$$I_\mu^* \to J^* \tag{51}$$

as $\mu \to \infty$.

Proof. We prove (50) first. Consider the following:

$$J(u^{\mu *}) - J^* = \{J(u^{\mu *}) - J(u^{\infty *})\} + \{J(u^{\infty *}) - J^*\}. \tag{52}$$

By Lemma 3 and the definition of J^* the difference in the second brace in (52) is nonnegative. On the other hand, employing the Hypothesis at the beginning of §5 and using (49), we find that the value in the first brace is greater than $-K_4\epsilon$. Hence it follows that

$$J(u^{\mu *}) - J^* > -K_4\epsilon, \tag{53}$$

where the quantities μ and ϵ are specified in the Hypothesis. By Corollary 1 we have

$$J(u^{\mu *}) - J^* \leqq 0 < K_4\epsilon. \tag{54}$$

The above two inequalities yield

$$| J(u^{\mu *}) - J^* | < K_4\epsilon. \tag{55}$$

Thus we have shown that for an $\epsilon > 0$ there exists an $N > 0$ such that (55) holds. Since the positive constant K_4 that is given by (76) in Appendix 3 is finite, (55) is enough to prove (50).

Next we prove (51). By definition,

$$I_\mu^* = J(u^{\mu *}) + \int_0^{T_\mu *} \pi(\sigma^\mu, u^{\mu *})\, dt, \tag{56}$$

where the integrand $\pi(\sigma^\mu, u^{\mu *})$ is the penalty, previously defined, corresponding to σ^μ and $u^{\mu *}$. From (56) and Lemma 2 it follows that

$$J^* \geqq J(u^{\mu *}) + \int_0^{T_\mu *} \pi(\sigma^\mu, u^{\mu *})\, dt,$$

or

$$J^* - J(u^{\mu *}) \geqq \int_0^{T_\mu *} \pi(\sigma^\mu, u^{\mu *})\, dt. \tag{57}$$

By (50), the left hand side of (57) approaches zero as $\mu \to \infty$. This implies that the right hand side also approaches zero as $\mu \to \infty$, since the penalty π is nonnegative. Thus we have proved that the first term in the right hand side of (56) approaches J^* and that the second term approaches zero. This is equivalent to (51).

6. Practical considerations and suboptimal control problems. In the preceding sections it was proved that a sequence of approximate formulations employing the penalty functions solves the original problem under the Hypothesis. If the optimal control which minimizes the charged cost functional, in any finite sequence of formulations developed by the penalty method, belongs to the admissible set in narrow sense, then this control itself is the optimal control for the original problem. Generally speaking, however, the admissible control in narrow sense will not be found in the process of minimizing a sequence of the charged cost functionals. After solving a finite number of the above sequential problems one may ask how far the solutions obtained differ from the exact one. This problem often arises in practice. Another practical consideration arises when the side constraints must be strictly satisfied from the physical viewpoint (the satisfaction of the side constraints needs to precede the minimization of the cost functional).

In this section we shall give consideration to the above situations from the viewpoint of applications.

Consider the following side constraints instead of (3):

$$g_k(x, u) + d_k \leqq 0, \qquad k = 1, \cdots, m, \tag{58}$$

with positive real numbers d_k. If these inequalities are satisfied, then (3) is also satisfied. However, the converse does not always hold. Thus (58) restricts the system *more strongly* than does (3). We apply the penalty method to the system using constraints (58). As we solve the sequential problems in the penalty method there may exist some large penalty weighting coefficients such that (3) holds even though (58) is violated. If this occurs, then let the corresponding control and cost functional be respectively $\tilde{u}$ and $\tilde{J}$. We call $\tilde{u}$ the *admissible suboptimal control in narrow sense* or simply the *suboptimal control*, and $\tilde{J}$ the *suboptimal cost functional.* From the definitions of $\tilde{u}$, $\tilde{J}$ and J^* it follows that

$$\tilde{J} = J(\tilde{u}) \geqq J^*. \tag{59}$$

This relation together with Corollary 1 yields

$$J(\tilde{u}) \geqq J^* \geqq J(u^{\mu*}). \tag{60}$$

Now we can estimate the difference between the optimal cost functional and the suboptimal cost functional as

$$0 \leqq J(\tilde{u}) - J^* \leqq J(\tilde{u}) - J(u^{\mu*}), \tag{61}$$

where the value of the right hand side of the above inequality is known. One may anticipate that this difference will decrease as each d_k decreases and μ increases.

Presented above is a technique to find the admissible control in narrow sense which yields an approximate optimal cost functional. The method of estimating the deviation of the approximate cost functional from the optimal one has also been considered.

Conclusion. A mathematical basis has been established for the penalty method applied to the optimum control problems subject to holonomic side constraints. The penalty method simplifies the logic involved in an optimum control problem from the standpoint of programming—since the corner conditions need not be checked. It was proved that the optimal solution of a sequence of problems employing the penalty method is equivalent to the optimal solution of the original problem in the limiting case. Since an infinite sequence is not physically realizable, the finite sequence of problems leading to a suboptimal control was considered. For brevity, only the minimization problems were treated. However, the theory is directly applicable to the maximization problems.

Acknowledgment. The author wishes to thank Professor R. A. Gambill of Purdue University for his reviewing this paper and Mr. D. E. Lovelace for his assistance during the preparation of this paper.

The constructive advice of the referee is also acknowledged.

Appendix 1. A *well-defined* piecewise continuous function $u(t)$ is a function of t such that each element $u_i(t)$ is

(a) piecewise continuous in the ordinary mathematical sense, and

(b) defined everywhere in the time domain considered and, for $\epsilon > 0$, either

$$u_i(t) = \lim_{\epsilon \to 0} u_i(t + \epsilon), \tag{62}$$

or

$$u_i(t) = \lim_{\epsilon \to 0} u_i(t - \epsilon). \tag{63}$$

Thus a piecewise continuous function $u(t)$ with an element $u_i(t)$, which has a singular point represented by A in Fig. 2, is not a well-defined piecewise continuous function mentioned above.

Appendix 2. As seen in the Introduction the definition of Ω assumes the existence of the corresponding trajectory $x(t)$. Thus $x^{\mu*}(t)$ and $x^{\infty*}(t)$ may be found corresponding to $x^{\mu*}(0)$ and $u^{\mu*}(t)$, and $x^{\infty*}(0)$ and $u^{\infty*}(t)$ respectively.

Let

$$v^{\mu}(t) \overset{\Delta}{=} u^{\mu*}\left(\frac{T_{\mu}^{*}}{T_{\infty}^{*}} t\right), \qquad 0 \leqq t \leqq T_{\infty}^{*}, \tag{64}$$

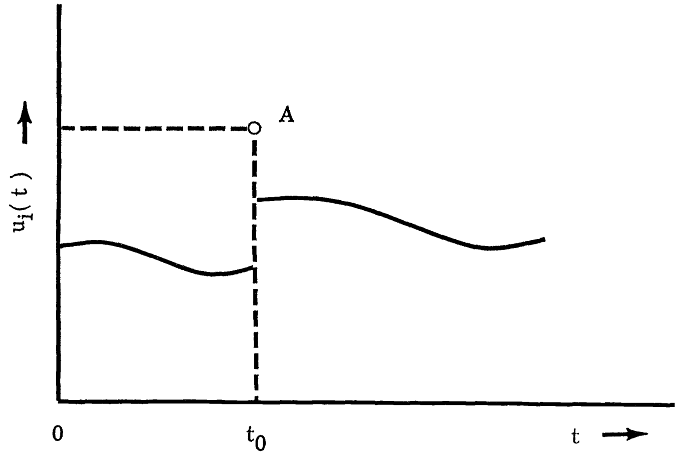


FIG. 2. *A piecewise continuous, but not "well-defined" function of time. A*: $(t_0, u_i(t_0))$

$$w^\mu(t) \overset{\Delta}{=} x^{\mu*}\left(\frac{T_\mu^*}{T_\infty^*}t\right), \qquad 0 \leqq t \leqq T_\infty^*. \tag{65}$$

We consider the variations

$$\delta u = v^\mu(t) - u^{\infty*}(t), \tag{66}$$

$$\delta x = w^\mu(t) - x^{\infty*}(t). \tag{67}$$

Modifying Rozonoer's technique [8], we obtain the inequalities

$$\begin{aligned} |\delta x_i(t)| \leqq K\Bigg\{\left|\frac{T_\mu^*}{T_\infty^*} - 1\right| &+ \int_0^{T_\infty^*}\sum_{j=1}^{r}|\delta u_j(t)|\,dt \\ &+ \sum_{i=1}^{n}|x_i^{\mu*}(0) - x_i^{\infty*}(0)|\Bigg\}, \qquad i = 1, \cdots, n, \end{aligned} \tag{68}$$

where K is a positive constant which has been introduced in considering the Lipschitz conditions and does not depend on $\delta u(t)$. Using the Hypothesis in §5 and the definition of the norm of control we have

$$\int_0^{T_\infty^*}|\delta u_j(t)|\,dt < \epsilon, \qquad j = 1, \cdots, r. \tag{69}$$

From (69) together with (31), inequalities (68) reduce to

$$|\delta x_i(t)| < K\left(\frac{n+r+1}{T_\infty^*}\right)\epsilon. \tag{70}$$

Appendix 3. By the properties of $p_k(g_k)$ and $g_k(x, u)$ it is clear that

$$p_k(x, u) \in \text{Lip }(x, u). \tag{71}$$

Hence it follows that

$$\tag{72} \begin{aligned} &\int_0^{T_\infty^*} | p_k(w^\mu, v^\mu) - p_k(x^{\infty *}, u^{\infty *}) | \, dt \\ &\qquad \leqq K_1 \int_0^{T_\infty^*} \left\{ \sum_{j=1}^{r} | \delta u_j(t) | + \sum_{i=1}^{n} | \delta x_i(t) | \right\} dt, \end{aligned}$$

where K_1 is the maximum of the Lipschitz constants for p_k. The existence of such a maximum constant can readily be assured. From (68) and (70), the above relation reduces to

$$\tag{73} \int_0^{T_\infty^*} | p_k(w^\mu, v^\mu) - p_k(x^{\infty *}, u^{\infty *}) | \, dt < K_2 \epsilon,$$

where

$$\tag{74} K_2 \equiv K_1\{1 + KT_\infty^*(n + r)\} > 0.$$

Similarly we can show that

$$\tag{75} \int_0^{T_\infty^*} | h(w^\mu, v^\mu) - h(x^{\infty *}, u^{\infty *}) | \, dt < K_4 \epsilon,$$

where

$$\tag{76} K_4 \equiv K_2\{1 + KT_\infty^*(n + r)\} > 0$$

and K_2 is the maximum of the Lipschitz constants for $h(x, u)$.

REFERENCES

[1] L. S. Pontryagin, V. D. Boltyanskii, R. D. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[2] C. D. Johnson and J. E. Gibson, *Singular solutions in problems of optimal control*, IEEE Trans. Automatic Control, (1963), pp. 4–15.

[3] L. D. Berkovitz, *Variational methods in problems of control programming*, J. Math. Anal. Appl., 3 (1961), pp. 145.

[4] S. S. L. Chang, *Minimal time control and multiple saturation limits*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 35–42.

[5] R. Courant, *Calculus of Variations and Supplementary Notes and Exercises*, 1945–1946, revised and amended by J. Moser, New York University Institute of Mathematical Sciences, New York, 1956–1957.

[6] G. Leitmann, *Optimization Techniques with Applications to Aerospace Systems*, Academic Press, New York, 1962, pp. 212–216.

[7] G. M. Ostrovskii, *On a method of solving variational problems*, Automat. Remote Control, 23 (1962), pp. 1284–1289.

[8] L. I. Rozonoer, *L. S. Pontryagin maximum principle in the theory of optimum systems*, Automat. Remote Control, 20 (1959), pp. 1298–1299.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

OPTIMAL CONTROL OF APERIODIC DISCRETE-TIME SYSTEMS*

B. W. JORDAN† and E. POLAK‡

1. Introduction. Much of previous work on the optimal control of discrete-time processes was concerned with sampled-data systems in which the sampling instants are fixed and equi-spaced [1]–[6] and in which the optimization is carried out over the amplitudes of the piecewise constant controls.

This paper is devoted to establishing necessary conditions for optimal control of a class of fixed duration, discrete-time processes with aperiodically modulated inputs, which are suggested by engineering considerations. The plants of the systems under consideration are described by nonlinear differential equations and the inputs are piecewise constant, suitably restricted in amplitude, and required to have K or fewer discontinuities whose position is not restricted. (K is a fixed, positive integer.) Using techniques analogous to the ones used in establishing the Pontryagin maximum principle, it is shown that for an admissible control to be optimal, it is necessary that a Hamiltonian-like functional be either locally maximum or stationary with respect to the admissible controls, and that a set of transversality conditions be satisfied.

Computational and engineering aspects of this problem are dealt with in a separate paper to be published soon.

2. Formulation of the optimal control problem.

a. *System equations.* We shall consider a system described by the vector differential equation

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

where $\mathbf{x} = (x_1, \cdots, x_n) \in E^n$ describes the state of the system, $\mathbf{u} = (u_1, \cdots, u_r) \in U \subset E^r$ is the control (or input), and $\mathbf{f} = (f_1, \cdots, f_n)$, where $f_i \in C^1$ on $E^n \times U$, $i = 1, 2, \cdots, n$. The set U will be defined below.

b. *Admissible controls.* Let U be a subset of E^r with the following properties.

(i) For every $\mathbf{v} \in U$ there exists at least one $\delta\mathbf{v} \in E^r$, $\delta\mathbf{v} \neq \mathbf{0}$, and a constant $\epsilon_1(\mathbf{v}, \delta\mathbf{v}) > 0$ such that $(\mathbf{v} + \epsilon\delta\mathbf{v}) \in U$ for all ϵ with $0 \leqq \epsilon \leqq \epsilon_1(\mathbf{v}, \delta\mathbf{v})$.

* Received by the editors March 19, 1964, and in final revised form November 9, 1964. This research has been made possible through support received from the Departments of Army, Navy, and Air Force under grant AF-AFOSR-139-63; and the National Aeronautics and Space Administration grant NSG 354.

† Aerospace Corporation, P. O. Box 95085, Los Angeles, California.

‡ Department of Electrical Engineering, University of California, Berkeley, California.

(ii) For a given $\mathbf{v} \in U$, let the set of all $\delta\mathbf{v}$ that possess the property (i) be denoted by $\Lambda(\mathbf{v})$. Then it is also assumed that $\Lambda(\mathbf{v})$ is convex. (It is clear that $\Lambda(\mathbf{v})$ is a cone.)

Let K be a fixed positive integer, and let M be the set of all vectors $\boldsymbol{\mu} = (\mathbf{v}_0, \mathbf{v}_1, \cdots, \mathbf{v}_{K-1}) \in E^{rK}$ such that $\mathbf{v}_i \in U$, $i = 0, 1, \cdots, K - 1$.

Let $I = (t_0, t_K)$, $t_0 < t_K$, be a fixed open interval. Let W be the set of vectors $\boldsymbol{\tau} = (t_1, \cdots, t_{K-1}) \in E^{K-1}$ such that $t_0 < t_1 \leqq t_2 \leqq \cdots \leqq t_{K-1} < t_K$.

For any $\boldsymbol{\tau} = (t_1, \cdots, t_{K-1}) \in W$ and $\boldsymbol{\mu} = (\mathbf{v}_0, \cdots, \mathbf{v}_{K-1}) \in M$, let $\mathbf{u}(t; \boldsymbol{\tau}, \boldsymbol{\mu})$ be the function from $[t_0, t_K]$ to E^r defined by

$$\begin{aligned} &\mathbf{u}(t; \boldsymbol{\tau}, \boldsymbol{\mu}) = \mathbf{v}_i, \qquad \text{for} \quad t_i \leqq t < t_{i+1}, \quad i = 0, 1, \cdots, K - 1, \\ &\mathbf{u}(t_K; \boldsymbol{\tau}, \boldsymbol{\mu}) = \mathbf{v}_{K-1}. \end{aligned} \tag{2}$$

All such controls will be referred to as admissible controls. When convenient, we shall simply write $\mathbf{u}(t)$ for $\mathbf{u}(t; \boldsymbol{\tau}, \boldsymbol{\mu})$.

c. *Constraints on the terminal state.* If $\mathbf{u}(\cdot)$ is any piecewise continuous function from $[t', \infty)$ to E^r, let $\mathbf{x}(t; \mathbf{x}', t', \mathbf{u})$ denote the solution of (1) which corresponds to the "control" $\mathbf{u}(t)$ and which satisfies the initial condition $\mathbf{x}(t') = \mathbf{x}'$. When there is no possibility of confusion, $\mathbf{x}(t; \mathbf{x}^0, t_0, \mathbf{u})$ will simply be denoted by $\mathbf{x}(t)$.

The terminal state $\mathbf{x}(t_K)$ will be required to belong to a set $S \subset E^n$. Two distinct cases for the set S will be considered.

Case (i). The set S is an $(n - l)$-dimensional manifold described by a set of l scalar equations; i.e.,

$$S = \{\mathbf{x} \mid g_i(\mathbf{x}) = 0, \mathbf{x} \in E^n, i = 1, 2, \cdots, l < n\}, \tag{3}$$

where the g_i, $i = 1, 2, \cdots, l$, are continuously differentiable functions from some region in E^n to E^1, and the vectors $\nabla g_i(\mathbf{x})$, $i = 1, 2 \cdots, l$, are assumed to be linearly independent for each $\mathbf{x} \in S$.

Case (ii). The set S is a closed convex subset of E^n. If S is a proper subset of E^n consisting of more than one point, it will be assumed that at each boundary point of S there exists a unique support hyperplane. Clearly, S may also consist of a single point or be the entire space E^n.

d. *Cost functional.* If for an admissible control $\mathbf{u}$, $\mathbf{x}(t''; \mathbf{x}', t', \mathbf{u}) = \mathbf{x}''$, we shall say that the cost of the transition from the state $\mathbf{x}'$ to the state $\mathbf{x}''$ in the interval $[t', t'']$, caused by the control $\mathbf{u}(t)$, is

$$\int_{t'}^{t''} f_0(\mathbf{x}(t; \mathbf{x}', t', \mathbf{u}), \mathbf{u}(t))\, dt,$$

where f_0 is a given function in C^1 on $E^n \times U$. Let the cost of a transition from the initial state $\mathbf{x}^0$ to the terminal state $\mathbf{x}^f$ caused by a control $\mathbf{u}(t)$ in the interval $[t_0, t_K]$ be denoted by $J(\mathbf{u}, \mathbf{x}^0)$. Then clearly, $J(\mathbf{u}, \mathbf{x}^0) = x_0(t_K; \mathbf{x}^0, t_0, \mathbf{u})$ where

$$\begin{aligned} \dot{x}_0(t; \mathbf{x}^0, t_0, \mathbf{u}) &= f_0(\mathbf{x}(t; \mathbf{x}^0, t_0, \mathbf{u}), \mathbf{u}(t)), \qquad t_0 \leqq t \leqq t_K, \\ x_0(t_0, \mathbf{x}^0, t_0, \mathbf{u}) &= 0. \end{aligned} \tag{4}$$

e. *Augmented system equations.* Now, for convenience, the system equations will be augmented to include the cost variable by defining the $(n + 1)$-dimensional vector $\tilde{\mathbf{x}} = (x_0, \mathbf{x})$, and the $(n + 1)$-dimensional vector function $\tilde{\mathbf{f}} = (f_0, \mathbf{f})$. From (1) and (4), the differential equation of the augmented system is

$$\dot{\tilde{\mathbf{x}}} = \tilde{\mathbf{f}}(\mathbf{x}, \mathbf{u}). \tag{5}$$

We shall also denote $(x_0(t; \mathbf{x}^0, t_0, \mathbf{u}), \mathbf{x}(t; \mathbf{x}^0, t_0, \mathbf{u}))$ by $\tilde{\mathbf{x}}(t; \mathbf{x}^0, t_0, \mathbf{u})$.

f. *Statement of the problem.* The optimal control problem for the systems under consideration can be stated as follows. (P): Given the initial time t_0, the final time t_K, the initial state $\mathbf{x}^0$, and the terminal state constraint set S, for the system described by (5), find an admissible control $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$, $(t \in \bar{I} = [t_0, t_K]$, $\boldsymbol{\tau}^* = (t_1^*, \cdots, t_{K-1}^*) \in W$, $\boldsymbol{\mu}^* = (\mathbf{v}_0^*, \cdots, \mathbf{v}_{K-1}^*) \in M)$, such that (i) $\mathbf{x}(t_K; \mathbf{x}^0, t_0, \mathbf{u}^*) \in S$, (ii) for all admissible controls $\mathbf{u}(t) = \mathbf{u}(t; \boldsymbol{\tau}, \boldsymbol{\mu})$ such that $\mathbf{x}(t_K; \mathbf{x}^0, t_0, \mathbf{u}) \in S$,

$$x_0(t_K; \mathbf{x}^0, t_0, \mathbf{u}^*) \leqq x_0(t_K; \mathbf{x}^0, t_0, \mathbf{u}).$$

An admissible control $\mathbf{u}^*$ which satisfies (i) and (ii) above will be called an optimal control and the corresponding trajectory $\tilde{\mathbf{x}}^*(t) = (x_0^*(t), \mathbf{x}^*(t)) = \tilde{\mathbf{x}}(t; \mathbf{x}^0, t_0, \mathbf{u}^*) = (x_0(t; \mathbf{x}^0, t_0, \mathbf{u}^*), \mathbf{x}(t; \mathbf{x}^0, t_0, \mathbf{u}^*))$, $t_0 \leqq t \leqq t_K$, will be called an optimal trajectory.

3. Necessary conditions for an optimal control.

a. *The adjoint system.* For a given admissible control $\mathbf{u}$, let $\tilde{\mathbf{p}} = (p_0, p_1, \cdots, p_n)$ be a solution of the differential equation

$$\dot{\tilde{\mathbf{p}}}(t) = -\left[\frac{\partial \tilde{\mathbf{f}}(\mathbf{x}(t; \mathbf{x}^0, t_0, \mathbf{u}), \mathbf{u}(t))}{\partial \tilde{\mathbf{x}}}\right]^T \tilde{\mathbf{p}}(t), \qquad t_0 \leqq t \leqq t_K, \tag{6}$$

where the superscript T denotes transposition, and $(\partial \tilde{\mathbf{f}}/\partial \tilde{\mathbf{x}})$ is the $(n + 1) \times (n + 1)$ matrix whose i, jth element is $\partial f_i/\partial x_j$, $(i, j = 0, 1, \cdots, n)$.

b. *The Hamiltonian.* Let the "Hamiltonian", H, be defined by

$$H(\tilde{\mathbf{p}}, \mathbf{x}', \mathbf{v}, t', t'') = \left(\tilde{\mathbf{p}}, \int_{t'}^{t''} \tilde{\mathbf{f}}(\mathbf{x}(t; \mathbf{x}', t', \mathbf{v}), \mathbf{v})\, dt\right), \tag{7}$$

for $\tilde{\mathbf{p}} \in E^{n+1}$, $\mathbf{x}' \in E^n$, $\mathbf{v} \in U$, $t_0 \leqq t' \leqq t'' \leqq t_K$. Note that, in this formulation, $\mathbf{v}$ is used to denote both a vector and the constant function that takes on the value $\mathbf{v}$ for all t, $t' \leqq t \leqq t''$.

THEOREM 1. *If* $\mathbf{u}^*(t) = \mathbf{u}(t;\, \boldsymbol{\tau}^*,\, \boldsymbol{\mu}^*)$, $(\boldsymbol{\tau}^* = (t_1{}^*, \cdots, t_{K-1}^*)$, $\boldsymbol{\mu}^* = (\mathbf{v}_0{}^*, \cdots, \mathbf{v}_{K-1}^*))$, *is an optimal control, and* $\tilde{\mathbf{x}}^*(t) = \tilde{\mathbf{x}}(t;\, \mathbf{x}^0,\, t_0,\, \mathbf{u}^*)$, $t_0 \leqq t \leqq t_K$, *the corresponding optimal trajectory for* (P), *then there exists a nonzero function,* $\tilde{\mathbf{p}}^*(t)$, $t_0 \leqq t \leqq t_K$, *satisfying* (6) *with* $\mathbf{u}(t) = \mathbf{u}^*(t)$ *for* $t_0 \leqq t \leqq t_K$ *such that*

(i) *for each* $k = 1, 2, \cdots K$, *the function* $H(\tilde{\mathbf{p}}^*(t_k{}^*),\, \mathbf{x}^*(t_{k-1}^*),\, \mathbf{v},\, t_{k-1}^*,\, t_k{}^*)$ *of the variable* $\mathbf{v} \in U$ *has either a local maximum or a stationary value at the point* $\mathbf{v} = \mathbf{v}_{k-1}^*$†;

(ii) $(\tilde{\mathbf{p}}^*(t_k{}^*),\, \tilde{\mathbf{f}}(\mathbf{x}^*(t_k{}^*),\, \mathbf{v}_k{}^*)) = (\tilde{\mathbf{p}}^*(t_k{}^*),\, \tilde{\mathbf{f}}(\mathbf{x}^*(t_k{}^*),\, \mathbf{v}_{k-1}^*))$, $k = 1, 2, \cdots, K - 1$;

(iii) $p_0{}^*(t_K) \leqq 0$.

Discussion. This theorem is basic and holds regardless of the form of the terminal constraint set S.

The problem (P) has been reduced to a two point boundary value problem. There are $(n + 1 + rK + K - 1)$ unknowns in the problem: the components of the vectors $\boldsymbol{\tau}$, $\boldsymbol{\mu}$ and $\tilde{\mathbf{p}}(t)$ for some t, $t_0 \leqq t \leqq t_K$. Condition (i) of Theorem 1 gives rK necessary conditions, while condition (ii) gives $K - 1$ necessary conditions for an optimal control for (P). There remain $n + 1$ necessary conditions to be found. These $n + 1$ necessary conditions will be the transversality conditions which $\tilde{\mathbf{p}}^*(t_K)$ must satisfy and which depend on the form of the constraint set, S.

Since conditions (i) and (ii) of Theorem 1 are only necessary conditions, they may not uniquely determine the optimal control. Indeed, (P) may not have a unique solution.

Theorem 1 will be proven by examining each type of terminal constraint set in turn and establishing the transversality conditions for each case. These transversality conditions are developed in Theorems 2–3. It will be shown that for each type of terminal constraint, the conditions of Theorem 1 are necessary.

The basic technique to be used will be to assume that the optimal control and trajectory are known. The control will then be perturbed so as to affect the trajectory only slightly. The necessary conditions which the optimal control must satisfy will then arise from the realization that any trajectory resulting from an admissible perturbed control and satisfying the constraints on the terminal state must not give a lower cost.

The first item, then, to be considered is the effect upon the trajectory of small perturbations in the control. Since $\tilde{\mathbf{x}}(t_0)$ is given, no perturbations of its value need be considered. Only perturbations in the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ defining the control must be considered. It is at this point that the basic difference between this discrete time problem and the similar one for continuous time problems occurs. It is required that any perturbation must be

† The exact meaning of this condition is given later in (22), §3d.

such that the perturbed control is admissible, and that it affect the trajectory only slightly. In the continuous time problem, the control is only assumed to be measurable. Consequently, the perturbed control can vary from the original control by any finite amount, provided the length of time, over which the perturbations are large, is made arbitrarily small. Such perturbations will affect the trajectory only slightly. This allows one to search out all of the control space, at each time, and leads to the requirement that the Hamiltonian be an absolute maximum at each instant of time.

In our problem, however, the admissible controls are piecewise constant. Thus, the only perturbations of the control which will affect the trajectory only slightly are small perturbations of $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$. Consequently, only local conditions can be obtained.

c. *Variation of the control and the trajectory.* Let an optimal control, $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$, $t_0 \leqq t \leqq t_K$, together with its resulting optimal trajectory, $\tilde{\mathbf{x}}^*(t)$, $t_0 \leqq t \leqq t_K$, be given.

We now consider perturbations of the optimal control $\mathbf{u}^*(t)$ defined as follows. Let δt_i $(i = 1, \cdots, K - 1)$ be real numbers such that $\delta t_j = \delta t_K$ if $t_j^* = t_k^*$, but otherwise arbitrary. Let $\delta \mathbf{v}_i$ $(i = 0, 1, \cdots, K - 1)$ be vectors in E^r such that $\delta \mathbf{v}_i \in \Lambda(\mathbf{v}_i^*)$ for each i. Let $\delta\boldsymbol{\tau} = (\delta t_1, \cdots, \delta t_{K-1})$ and $\delta\boldsymbol{\mu} = (\delta \mathbf{v}_0, \cdots, \delta \mathbf{v}_{K-1})$. Then if $\epsilon > 0$ is sufficiently small, $(\boldsymbol{\tau}^* + \epsilon\delta\boldsymbol{\tau}) \in W$ and $(\boldsymbol{\mu}^* + \epsilon\delta\boldsymbol{\mu}) \in M$. We denote the control $\mathbf{u}(t; \boldsymbol{\tau}^* + \epsilon\delta\boldsymbol{\tau}, \boldsymbol{\mu}^* + \epsilon\delta\boldsymbol{\mu})$ by $\mathbf{u}(t; \epsilon)$. Let $\tilde{\mathbf{x}}(t; \epsilon) = \tilde{\mathbf{x}}(t; \mathbf{x}^0, t_0, \mathbf{u}(t; \epsilon))$ and $\delta\tilde{\mathbf{x}}(t; \epsilon) = \tilde{\mathbf{x}}(t; \epsilon) - \tilde{\mathbf{x}}^*(t)$.

First, suppose that $\delta\boldsymbol{\tau} = 0$, $\delta\mathbf{v}_i = 0$, $i = 0, 1, \cdots, K - 1$, $i \neq \nu - 1$. Then $\partial\mathbf{u}(t; \epsilon)/\partial\epsilon$ exists and

$$\left(\frac{\partial \mathbf{u}(t; \epsilon)}{\partial \epsilon}\right)_{\epsilon=0} = 0, \qquad t_0 \leqq t < t_{\nu-1}^*, \quad t_\nu^* \leqq t \leqq t_K,$$

$$\left(\frac{\partial \mathbf{u}(t; \epsilon)}{\partial \epsilon}\right)_{\epsilon=0} = \delta\mathbf{v}_{\nu-1}, \quad t_{\nu-1}^* \leqq t < t_\nu^*.$$

Further, $\tilde{\mathbf{y}}(t) = (\partial\tilde{\mathbf{x}}(t; \epsilon)/\partial\epsilon)_{\epsilon=0}$, for $t_0 \leqq t \leqq t_K$, exists and satisfies the differential equation

$$(8) \quad \frac{d\tilde{\mathbf{y}}}{dt} = \frac{\partial\tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{u}^*(t))}{\partial\tilde{\mathbf{x}}}\,\tilde{\mathbf{y}} + \frac{\partial\tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{u}^*(t))}{\partial\mathbf{u}}\left(\frac{\partial\mathbf{u}(t; \epsilon)}{\partial\epsilon}\right)_{\epsilon=0}, \quad \tilde{\mathbf{y}}(t_0) = 0,$$

where $\partial\tilde{\mathbf{f}}(\mathbf{x}, \mathbf{u})/\partial\tilde{\mathbf{x}}$ is the $(n + 1) \times (n + 1)$ matrix whose j, kth element is $\partial f_j(\mathbf{x}, \mathbf{u})/\partial x_k$ and $\partial\tilde{\mathbf{f}}(\mathbf{x}, \mathbf{u})/\partial\mathbf{u}$ is the $(n + 1) \times r$ matrix whose j, kth element is $\partial f_j(\mathbf{x}, \mathbf{u})/\partial u_k$.

The solution of (8) is given by

$$\tilde{\mathbf{y}}(t) = \int_{t_0}^{t} \boldsymbol{\Phi}(t, s)\,\frac{\partial\tilde{\mathbf{f}}(\mathbf{x}^*(s), \mathbf{u}^*(s))}{\partial\mathbf{u}}\left(\frac{\partial\mathbf{u}(s; \epsilon)}{\partial\epsilon}\right)_{\epsilon=0} ds, \quad t_0 \leqq t \leqq t_K,$$

where $\boldsymbol{\Phi}(t, s)$ is the $(n + 1) \times (n + 1)$ matrix function satisfying the homogeneous differential equation

$$(9) \quad \frac{\partial \boldsymbol{\Phi}(t, s)}{\partial t} = \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{u}^*(t))}{\partial \tilde{\mathbf{x}}} \boldsymbol{\Phi}(t, s), \quad \boldsymbol{\Phi}(s, s) = \mathbf{I}, \qquad t_0 \leqq t, s \leqq t_K .$$

Therefore, since $\boldsymbol{\Phi}(t_1, t_2)\boldsymbol{\Phi}(t_2, t_3) = \boldsymbol{\Phi}(t_1, t_3)$,

$$\tilde{\mathbf{y}}(t_K) = \boldsymbol{\Phi}(t_K, t_\nu^*) \int_{t_{\nu-1}^*}^{t_\nu^*} \boldsymbol{\Phi}(t_\nu^*, t) \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)}{\partial \mathbf{u}} \delta \mathbf{v}_{\nu-1}\, dt,$$

and

$$\delta \tilde{\mathbf{x}}(t_K ; \epsilon) = \epsilon \left(\frac{\partial \tilde{\mathbf{x}}(t_K ; \epsilon)}{\partial \epsilon} \right)_{\epsilon=0} + \tilde{\mathbf{o}}(\epsilon) = \epsilon \tilde{\mathbf{y}}(t_K) + \tilde{\mathbf{o}}(\epsilon).$$

The vector $\tilde{\mathbf{o}}(\epsilon) = (o_0(\epsilon), \cdots, o_n(\epsilon))$, where $\lim_{\epsilon \to 0} o_i(\epsilon)/\epsilon = 0$, $i = 0, \cdots, n$.

Now suppose that $\delta \boldsymbol{\mu} = 0$, $\delta t_i = 0$, $i \neq \nu$, $i = 1, \cdots, K - 1$, and that $\delta t_\nu \leqq 0$. Then

$$\tilde{\mathbf{x}}(t_\nu^*; \epsilon) = \tilde{\mathbf{x}}^*(t_\nu^*) + \int_{t_\nu^* + \epsilon \delta t_\nu}^{t_\nu^*} [\tilde{\mathbf{f}}(\mathbf{x}(t; \epsilon), \mathbf{v}_\nu^*) - \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)]\, dt.$$

It is easily seen that $[\partial \tilde{\mathbf{x}}(t; \epsilon)/\partial \epsilon]_{\epsilon=0}$ exists for $t_\nu^* < t \leqq t_K$ in this case, and that, if we again introduce the notation $\tilde{\mathbf{y}}(t)$ for this derivative, then

$$\tilde{\mathbf{y}}(t_\nu^{*+}) = \left(\frac{\partial \tilde{\mathbf{x}}(t_\nu^{*+}; \epsilon)}{\partial \epsilon} \right)_{\epsilon=0} = [\tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_{\nu-1}^*) - \tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_\nu^*)]\delta t_\nu ,$$

$$\tilde{\mathbf{y}}(t) = \boldsymbol{\Phi}(t, t_\nu^*)\tilde{\mathbf{y}}(t_\nu^{*+}), \qquad t_\nu^* \leqq t \leqq t_K ,$$

so that, in particular,

$$\tilde{\mathbf{y}}(t_K) = \boldsymbol{\Phi}(t_K, t_\nu^*)[\tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_{\nu-1}^*) - \tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_\nu^*)]\delta t_\nu .$$

If $\delta t_\nu \geqq 0$, an identical formula can be derived. It can be shown that for arbitrary $\delta \boldsymbol{\tau}$, $\delta \boldsymbol{\mu}$, the effects due to the δt_i and $\delta \mathbf{v}_i$ are additive so that, in general,

$$\delta \tilde{\mathbf{x}}(t_K ; \epsilon) = \epsilon \tilde{\mathbf{y}} + \tilde{\mathbf{o}}(\epsilon),$$

where

$$(10) \quad \begin{aligned} \tilde{\mathbf{y}} &= \sum_{i=0}^{K-1} \boldsymbol{\Phi}(t_K, t_{i+1}^*) \int_{t_i^*}^{t_{i+1}^*} \boldsymbol{\Phi}(t_{i+1}^*, t) \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_i^*)}{\partial \mathbf{u}}\, dt\, \delta \mathbf{v}_i \\ &\quad + \sum_{i=1}^{K-1} \boldsymbol{\Phi}(t_K, t_i^*)[\tilde{\mathbf{f}}(\mathbf{x}^*(t_i^*), \mathbf{v}_{i-1}^*) - \tilde{\mathbf{f}}(\mathbf{x}^*(t_i^*), \mathbf{v}_i^*)]\delta t_i . \end{aligned}$$

Let $\mathcal{K} = \{\tilde{\mathbf{y}} + \tilde{\mathbf{x}}^*(t_K) | \tilde{\mathbf{y}}$ of the form (10), $\delta \mathbf{v}_i \in \Lambda(\mathbf{v}_i^*)$, $i = 0, \cdots, K - 1$,

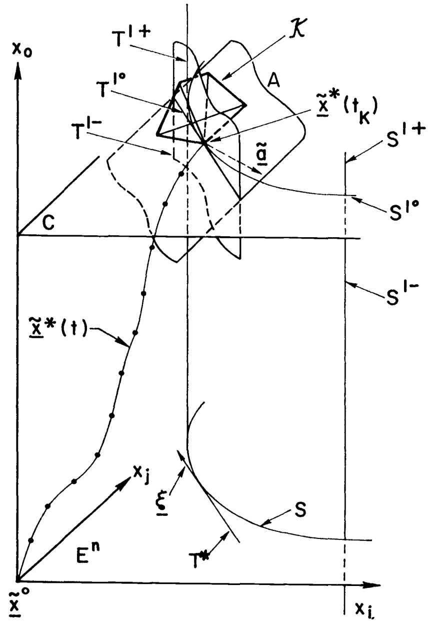


FIG. 1. *Illustration for Theorem 2*

δt_i arbitrary real numbers except that if $t_i = t_j$, $\delta t_i = \delta t_j$, $i, j = 1, \cdots, K - 1\}$. Since $\Lambda(\mathbf{v})$ is a covex cone for each $\mathbf{v} \in U$, $\mathcal{K}$ is a convex cone.

d. *Case* 1. *Right end constrained to lie on a smooth surface.* Let the terminal constraint set S be an $(n - l)$-dimensional manifold as given by (3). Because of the assumptions on the g_j, there is an $(n - l)$-dimensional plane, T, tangent to S at each $\mathbf{x}' \in S$ described by

$$T = \{\mathbf{x}'' \mid \left((\mathbf{x}'' - \mathbf{x}'), \frac{\partial g_j(\mathbf{x}')}{\partial \mathbf{x}}\right) = 0, j = 1, 2, \cdots, l\}.$$

Let S^1 denote the $(n + 1 - l)$-dimensional cylinder in E^{n+1} defined by

$$S^1 = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} = (\eta, \mathbf{x}), \mathbf{x} \in S, \eta \text{ an arbitrary real number}\}$$

(see Fig. 1). Since $\tilde{\mathbf{x}}^*(t_K) \in S^1$ by hypothesis, there is an $(n + 1 - l)$-dimensional plane, T^1, tangent to S^1 at $\tilde{\mathbf{x}}^*(t_K)$ which is given by

$$T^1 = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} = (\eta, \mathbf{x}), \mathbf{x} \in T^*, \eta \text{ an arbitrary real number}\},$$

where T^* is the tangent plane to S at $\mathbf{x}^*(t_K)$.

Let C be the n-dimensional hyperplane passing through $\tilde{\mathbf{x}}^*(t_K)$ and perpendicular to the x_0-axis, i.e.,

$$C = \{\tilde{\mathbf{x}} = (x_0, x_1, \cdots, x_n) \mid x_0 = x_0^*(t_K)\}.$$

The hyperplane C divides the plane T^1 into two semi-infinite planes

$$(11) \qquad T^{1+} = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} \in T^1, x_0 \geqq x_0^*(t_K)\},$$

$$(12) \qquad T^{1-} = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} \in T^1, x_0 \leqq x_0^*(t_K)\},$$

with the common boundary

$$(13) \qquad T^{1,0} = C \cap T^1.$$

The hyperplane C also divides S^1 into two semi-infinite cylinders

$$S^{1+} = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} \in S^1, x_0 \geqq x_0^*(t_K)\},$$

$$S^{1-} = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} \in S^1, x_0 \leqq x_0^*(t_K)\},$$

with the common boundary

$$S^{1,0} = C \cap S^1.$$

Let $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_n)$ be an arbitrary n-dimensional vector lying in T^*. Then for this case, the following theorem holds;

Theorem 2. *Consider the problem* (P) *when the constraint set S is an $(n - l)$-dimensional smooth manifold defined by* (3). *Then, necessary conditions for $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$, $t_0 \leqq t \leqq t_K$, to be an optimal control are that conditions* (i), (ii) *and* (iii) *of Theorem* 1 *be satisfied and, in addition, that for every vector $\boldsymbol{\xi}$ lying in T^*,*

$$(\text{iv}) \qquad (\mathbf{p}^*(t_K), \boldsymbol{\xi}) = 0,$$

where

$$\tilde{\mathbf{p}}^*(t_K) = (p_0^*(t_K), \mathbf{p}^*(t_K)).$$

Proof. Since $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$ is an optimal control, it is necessary that any admissible perturbed control whose corresponding trajectory satisfies the terminal conditions not give a lower cost. For this requirement to be fulfilled, it is necessary that there exist a hyperplane separating $\mathcal{K}$ and T^{1-}. This is shown by establishing Lemma 1.

Lemma 1. *Let $\tilde{\mathbf{x}}^*(t)$, $t_0 \leqq t \leqq t_K$, be an optimal trajectory corresponding to an optimal control $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$, $t_0 \leqq t \leqq t_K$. Let G be an l-dimensional ($l \leqq n$) smooth manifold with an edge, G_e, in E^{n+1}, such that $\tilde{\mathbf{x}}^*(t_K) \in G_e$. Let L be the halfplane tangent to G* at $\tilde{\mathbf{x}}^*(t_K)$.

If the cones, $\mathcal{K}$ and L, having a common vertex at $\tilde{\mathbf{x}}^*(t_K)$, are not separated,† *then there exists an admissible control $\mathbf{u}(t) = \mathbf{u}(t; \boldsymbol{\tau}, \boldsymbol{\mu})$,*

† Two convex cones M_1 and M_2 in E^{n+1} with a common vertex $\tilde{\mathbf{x}}'$ are separated if there exists a hyperplane such that M_1 is entirely contained in one closed halfspace defined by this hyperplane, and M_2 is entirely contained in the other. Also, see [7, p. 94].

$t_0 \leqq t \leqq t_K$, with a corresponding trajectory, $\tilde{\mathbf{x}}(t; \mathbf{x}^0, t_0, \mathbf{u})$, $t_0 \leqq t \leqq t_K$, such that $\tilde{\mathbf{x}}(t_K) \in G$ but $\tilde{\mathbf{x}}(t_K) \notin G_e$.

Proof of Lemma 1. The proof of this lemma is very similar to the one given for [7, Lemma 10] and is therefore omitted.

Proof of Theorem 2. It follows from Lemma 1 that if the cones $\mathcal{K}$ and T^{1-}, having the common vertex $\tilde{\mathbf{x}}^*(t_K)$, are not separated, then there exists an admissible control $\mathbf{u}(t) = \mathbf{u}(t; \boldsymbol{\tau}, \boldsymbol{\mu})$, $t_0 \leqq t \leqq t_K$, such that $\tilde{\mathbf{x}}(t_K; \mathbf{x}^0, t_0, \mathbf{u})$ lies in S^{1-} but not on the edge of S^{1-}, and consequently will satisfy the constraints on the terminal state and have a lower cost. This is a contradiction. Therefore, for the admissible control $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$, $t_0 \leqq t \leqq t_K$, and the corresponding trajectory $\tilde{\mathbf{x}}^*(t)$, $t_0 \leqq t \leqq t_K$, to be optimal, it is necessary that there exist an n-dimensional hyperplane separating $\mathcal{K}$ and T^{1-}. We shall denote this hyperplane by A.

Let the $(n + 1)$-dimensional vector $\tilde{\mathbf{a}} = (a_0, a_1, \cdots, a_{n+1})$ be a normal to A directed so that

$$(14) \qquad \begin{aligned} ((\tilde{\mathbf{x}} - \tilde{\mathbf{x}}^*(t_K)), \tilde{\mathbf{a}}) &\leqq 0 \quad \text{if} \quad \tilde{\mathbf{x}} \in \mathcal{K}, \\ \text{and} \quad ((\tilde{\mathbf{x}} - \tilde{\mathbf{x}}^*(t_K)), \tilde{\mathbf{a}}) &\geqq 0 \quad \text{if} \quad \tilde{\mathbf{x}} \in T^{1-}. \end{aligned}$$

Clearly, the hyperplane A contains $\tilde{\mathbf{x}}^*(t_K)$ and $T^{1,0}$. Let

$$\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)$$

be any vector lying in T^*. Then $\tilde{\boldsymbol{\xi}} = (0, \boldsymbol{\xi})$ will be parallel to $T^{1,0}$. Since $T^{1,0} \subset A$, $(\tilde{\mathbf{a}}, \tilde{\boldsymbol{\xi}}) = 0$. But $\xi_0 = 0$. Therefore

$$(\tilde{\mathbf{a}}, \tilde{\boldsymbol{\xi}}) = \sum_{i=1}^{n} a_i \xi_i = 0.$$

Also, from the definition of T^{1-}, it is clear that the vector $[\tilde{\mathbf{x}}^*(t_K) + (-1, 0, 0, \cdots, 0)] \in T^{1-}$. Hence, by virtue of (14) $a_0 \leqq 0$.

Since $\tilde{\mathbf{y}} \in \mathcal{K}$ it follows from the necessary condition (14) that

$$(15) \qquad (\tilde{\mathbf{a}}, \tilde{\mathbf{y}}) \leqq 0.$$

Let $\tilde{\mathbf{p}}^*(t)$ be the solution of (6), with $\mathbf{u} = \mathbf{u}^*$, that satisfies the boundary condition $\tilde{\mathbf{p}}^*(t_K) = \tilde{\mathbf{a}}$. Then (15) becomes

$$(16) \qquad (\tilde{\mathbf{p}}^*(t_K), \tilde{\mathbf{y}}) \leqq 0.$$

Consider the element of $\mathcal{K}$ for which $\delta t_i = 0$ for each i, $\delta \mathbf{v}_i = 0$ for $i \neq \nu - 1$, and $\delta \mathbf{v}_{\nu-1}$ is an arbitrary element of $\Lambda(\mathbf{v}_{\nu-1}^*)$.

Then,

$$\tilde{\mathbf{y}} = \boldsymbol{\Phi}(t_K, t_\nu^*) \int_{t_{\nu-1}^*}^{t_\nu^*} \boldsymbol{\Phi}(t_\nu^*, t) \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)}{\partial \mathbf{u}} \, dt \, \delta \mathbf{v}_{\nu-1},$$

and

$$(17)\qquad (\tilde{\mathbf{p}}^*(t_K), \tilde{\mathbf{y}}) = \left(\mathbf{\Phi}^T(t_K, t_\nu^*)\tilde{\mathbf{p}}^*(t_K), \int_{t_{\nu-1}^*}^{t_\nu^*} \mathbf{\Phi}(t_\nu^*, t) \cdot \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)}{\partial \mathbf{u}}\, dt\, \delta\mathbf{v}_{\nu-1}\right).$$

It is well known that the general solution of (6), with $\mathbf{u} = \mathbf{u}^*$, is $\tilde{\mathbf{p}}(t) = \mathbf{\Phi}^T(t_K, t)\tilde{\mathbf{q}}$, where $\tilde{\mathbf{q}}$ is an arbitrary vector in E^{n+1}. Consequently,

$$(18)\qquad \mathbf{\Phi}^T(t_K, t_\nu^*)\tilde{\mathbf{p}}^*(t_K) = \tilde{\mathbf{p}}^*(t_\nu^*).$$

Using (17) and (18) in (16), the necessary condition (16) becomes

$$(19)\qquad \left(\tilde{\mathbf{p}}^*(t_\nu^*), \int_{t_{\nu-1}^*}^{t_\nu^*} \mathbf{\Phi}(t_\nu^*, t)\, \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)}{\partial \mathbf{u}}\, dt\, \delta\mathbf{v}_{\nu-1}\right) \leqq 0.$$

Now the "Hamiltonian" (7) becomes for $t' = t_{\nu-1}^*$, $t'' = t_\nu^*$,

$$H(\tilde{\mathbf{p}}^*(t_\nu^*), \mathbf{x}^*(t_{\nu-1}^*), \mathbf{v}_{\nu-1}^*, t_{\nu-1}^*, t_\nu^*) = \left[\int_{t_{\nu-1}^*}^{t_\nu^*} \tilde{\mathbf{f}}(\mathbf{x}(t; \mathbf{x}^*(t_{\nu-1}^*), t_{\nu-1}^*, \mathbf{v}_{\nu-1}^*), \mathbf{v}_{\nu-1}^*)\, dt\right]^T \tilde{\mathbf{p}}^*(t_\nu^*),$$

and

$$(20)\qquad \nabla_{\mathbf{u}} H(\tilde{\mathbf{p}}^*(t_\nu^*), \mathbf{x}^*(t_{\nu-1}^*), \mathbf{v}_{\nu-1}^*, t_{\nu-1}^*, t_\nu^*) = \left[\frac{\partial}{\partial \mathbf{u}} \int_{t_{\nu-1}^*}^{t_\nu^*} \tilde{\mathbf{f}}(\mathbf{x}(t; \mathbf{x}^*(t_{\nu-1}^*), t_{\nu-1}^*, \mathbf{v}_{\nu-1}^*), \mathbf{v}_{\nu-1}^*)\, dt\right]^T \tilde{\mathbf{p}}^*(t_\nu^*),$$

where

$$\frac{\partial}{\partial \mathbf{u}} \int_{t_{\nu-1}^*}^{t_\nu^*} \tilde{\mathbf{f}}(\mathbf{x}(t; \mathbf{x}^*(t_{\nu-1}^*), t_{\nu-1}^*, \mathbf{v}_{\nu-1}^*), \mathbf{v}_{\nu-1}^*)\, dt$$

is an $(n + 1) \times r$ matrix whose i, jth element is

$$\frac{\partial}{\partial \mathbf{u}_j} \int_{t_{\nu-1}^*}^{t_\nu^*} f_i(\mathbf{x}(t; \mathbf{x}^*(t_{\nu-1}^*), t_{\nu-1}^*, \mathbf{u}), \mathbf{u})\, dt\, \big|_{\mathbf{u}=\mathbf{v}_{\nu-1}^*}.$$

Above, as in the remainder of the proof of this theoem, $\mathbf{u}$ denotes a *vector* in U (as well as the constant function that takes on the value $\mathbf{u}$). Clearly,

$$\tilde{\mathbf{x}}(t; \tilde{\mathbf{x}}^*(t_{\nu-1}^*), t_{\nu-1}^*, \mathbf{u}) = \tilde{\mathbf{x}}(t_{\nu-1}^*) + \int_{t_{\nu-1}^*}^{t} \tilde{\mathbf{f}}(\mathbf{x}(t; \mathbf{x}^*(t_{\nu-1}^*), t_{\nu-1}^*\, \mathbf{u}), \mathbf{u})\, dt,$$

$$t_{\nu-1}^* \leqq t \leqq t_\nu^*.$$

Let

$$\mathbf{Z}(t) = \frac{\partial \tilde{\mathbf{x}}(t; \tilde{\mathbf{x}}^*(t_{\nu-1}^*), t_{\nu-1}^*, \mathbf{u})}{\partial \mathbf{u}} \bigg|_{\mathbf{u}=\mathbf{v}_{\nu-1}^*}, \quad \text{for } t_{\nu-1}^* \leqq t \leqq t_\nu^*.$$

Then it is well-known that $\mathbf{Z}(t)$ satisfies the differential equation

$$\frac{d\mathbf{Z}(t)}{dt} = \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)}{\partial \tilde{\mathbf{x}}} \mathbf{Z} + \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)}{\partial \mathbf{u}}, \quad \mathbf{Z}(t_{\nu-1}^*) = 0.$$

Therefore, using the variation of parameters formula for solutions of this equation,

$$\begin{aligned} \mathbf{Z}(t_\nu^*) &= \frac{\partial}{\partial \mathbf{u}} \int_{t_{\nu-1}^*}^{t_\nu^*} \tilde{\mathbf{f}}(\mathbf{x}(t; \mathbf{x}^*(t_{\nu-1}^*), t_{\nu-1}^*, \mathbf{v}_{\nu-1}^*), \mathbf{v}_{\nu-1}^*)\, dt \\ &= \int_{t_{\nu-1}^*}^{t_\nu^*} \mathbf{\Phi}(t_\nu^*, t)\, \frac{\partial \tilde{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{v}_{\nu-1}^*)}{\partial \mathbf{u}}\, dt. \end{aligned} \tag{21}$$

Using (20) and (21), the necessary condition (19) becomes

$$(\nabla_{\mathbf{u}} H(\tilde{\mathbf{p}}^*(t_\nu^*), \mathbf{x}^*(t_{\nu-1}^*), \mathbf{v}_{\nu-1}^*, t_{\nu-1}^*, t_\nu^*), \delta\mathbf{v}_{\nu-1}) \leqq 0. \tag{22}$$

Since $\delta\mathbf{v}_{\nu-1}$ is an arbitrary member of $\Lambda(\mathbf{v}_{\nu-1}^*)$, (22) states that the function $H(\tilde{\mathbf{p}}^*(t_\nu^*), \mathbf{x}^*(t_{\nu-1}^*), \mathbf{v}, t_{\nu-1}^*, t_\nu^*)$ of the variable $\mathbf{v} \in U$ has either a local maximum or a stationary value at the point $\mathbf{v} = \mathbf{v}_{\nu-1}^*$. Since the choice of ν, $0 \leqq \nu \leqq K - 1$, was arbitrary, this must be true for each ν, $0 \leqq \nu \leqq K - 1$. Thus, condition (i) of Theorem 1 has been shown to be necessary.

Consider the element of $\mathcal{K}$ where $\delta\mathbf{v}_i = 0$ for each i, $\delta t_i = 0$ for $i \neq \nu$, and δt_ν is arbitrary.

Then

$$\tilde{\mathbf{y}} = \mathbf{\Phi}(t_K, t_\nu^*)[\tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_{\nu-1}^*) - \tilde{\mathbf{f}}(\mathbf{x}^*(t\nu^*), \mathbf{v}_\nu^*)]\delta t_\nu,$$

so that (16) implies that

$$(\tilde{\mathbf{p}}^*(t_K), \tilde{\mathbf{y}}) = (\tilde{\mathbf{p}}^*(t_\nu^*), [\tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_{\nu-1}^*) - \tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_\nu^*)])\delta t_\nu \leqq 0. \tag{23}$$

But δt_ν can be positive or negative. Hence (22) is satisfied only when

$$(\tilde{\mathbf{p}}^*(t_\nu^*), [\tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu), \mathbf{v}_{\nu-1}^*) - \tilde{\mathbf{f}}(\mathbf{x}^*(t_\nu^*), \mathbf{v}_\nu^*)]) = 0. \tag{24}$$

Again, the choice of ν, $1 \leqq \nu \leqq K - 1$, was arbitrary. Therefore (24) must hold for all ν, $1 \leqq \nu \leqq K - 1$. Thus, condition (ii) of Theorem 1 has been shown to be necessary. Recall that $\tilde{\mathbf{p}}^*(t_K) = \tilde{\mathbf{a}}$, $a_0 \leqq 0$ and $\sum_{i=1}^n a_i\xi_i = 0$ when $(\xi_1, \cdots, \xi_n)$ lies in T^*. Consequently, condition (iii) of Theorem 1 and condition (iv) of Theorem 2 have also been proven.

e. *Case* 2. *Right end constrained to lie at a point.* Next, consider the problem when S is a point. For this case, the following special case of Theorem 2 holds.

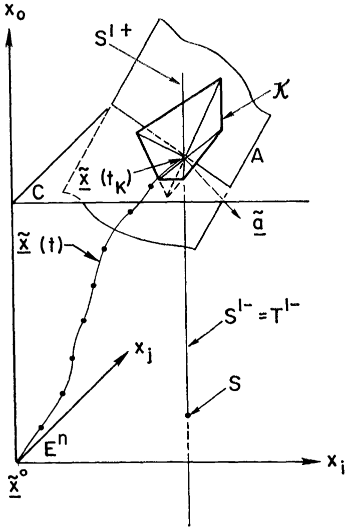


FIG. 2. *Illustration for proof of Theorem 2a*

THEOREM 2a. *Consider the problem* (P) *when the constraint set S is a point in E^n. Then, necessary conditions for* $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$ *to be an optimal control are that conditions* (i), (ii) *and* (iii) *of Theorem* 1 *be satisfied.*

Remark. (See Fig. 2.) Since S is a point, S^1 is a line parallel to the x_0 axis and passing through $\tilde{\mathbf{x}}^*(t_K)$. We define S^{1-} and T^{1-} as before. Clearly, $T^{1-} = S^{1-}$. It follows from Lemma 1 that $\mathcal{K}$ and S^{1-} must be separated by a hyperplane A. Let a normal $\tilde{\mathbf{a}}$ to A be defined as in (14). Then, as was shown in Theorem 2, $a_0 \leqq 0$. However, since $T^{1,0}$ consists of a point, no transversality conditions are imposed on $\tilde{\mathbf{a}}$. By proceeding as in Theorem 2, it is found by letting $\tilde{\mathbf{p}}^*(t_K) = \tilde{\mathbf{a}}$, that conditions (i), (ii) and (iii) of Theorem 1 are necessary for $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$ to be an optimal control. This completes the proof of Theorem 2a.

f. *Case* 3. *Free right end.* Consider the problem where S is the entire space E^n. For this terminal condition, the following special case of Theorem 2 holds.

THEOREM 2b. *Consider the problem* (P) *when $S = E^n$. Then, necessary conditions for* $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$ *to be an optimal control are that conditions* (i), (ii) *and* (iii) *of Theorem* 1 *be satisfied, and in addition that*

$$\text{(iv)} \qquad p_i^*(t_K) = 0, \quad i = 1, 2, \cdots, n.$$

Remark. (See Fig. 3.) We define C and S^{1-} as before. Clearly, S^{1-} is now a closed halfspace. By a simple extension of Lemma 1, it can be shown that for $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$ to be an optimal control, $\mathcal{K}$ must be separated from

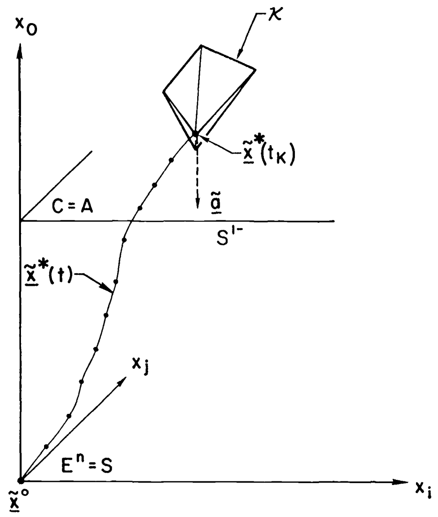


FIG. 3. *Illustration for proof of Theorem 2b*

S^{1-}. Since S^{1-} is a closed halfspace, the only hyperplane which can separate $\mathcal{K}$ from S^{1-} is C. Let a normal $\tilde{\mathbf{c}}$ to C be defined by

$$\tilde{\mathbf{c}} = (1, 0, 0, \cdots, 0). \tag{25}$$

By proceeding as in Theorem 2, and letting $\tilde{\mathbf{p}}^*(t_K) = -\tilde{\mathbf{c}}$, the conditions of Theorem 1 as well as the transversality conditions (iv) of Theorem 2b can be shown to be necessary. This completes the proof of Theorem 2b.

g. *Case* 4. *Right end constrained to lie in an n-dimensional subset of E^n.* The problems where S is a point, a manifold of dimension $(n - l) < n$, and the entire space E^n have been considered. The only problem left is that where S is a closed, convex, n-dimensional proper subset of E^n.

Let S^1, S^{1-}, S^{1+} and C be defined as before. Two possibilities can occur. (See Fig. 4.)

Case 4a. The point $\tilde{\mathbf{x}}^*(t_K)$ is on the boundary of S^1. Clearly, $\tilde{\mathbf{x}}^*(t_K) \in C$.

The following definitions will be used. Let T^1 be the supporting hyperplane to S^1 at $\tilde{\mathbf{x}}^*(t_K)$. Let T^{1+}, T^{1-} and $T^{1,0}$ be defined as in (11), (12) and (13). Let $\tilde{\mathbf{h}}$ be a normal to T^1 directed away from the interior of S^1. Then $\tilde{\mathbf{h}} = (0, h_1, \cdots, h_n)$. Let $\tilde{\mathbf{c}}$ be the normal to C given by $\tilde{\mathbf{c}} = (1, 0, \cdots, 0)$. Let $\tilde{\boldsymbol{\xi}}$ be an arbitrary vector parallel to $T^{1,0}$ and let

$$D = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} - \tilde{\mathbf{x}}^*(t_K) = \alpha\tilde{\mathbf{h}} + \beta\tilde{\mathbf{c}} + \tilde{\boldsymbol{\xi}}, \alpha \leqq 0, \beta \leqq 0\}. \tag{26}$$

Case 4b. Suppose that $\tilde{\mathbf{x}}^*(t_K)$ is in the interior of S^1. The point $\tilde{\mathbf{x}}^*(t_K)$ is, of course, also in C. In this case, a hyperplane analogous to T^1 above cannot be defined, and $\tilde{\mathbf{h}}$ will be assumed to be the zero vector.

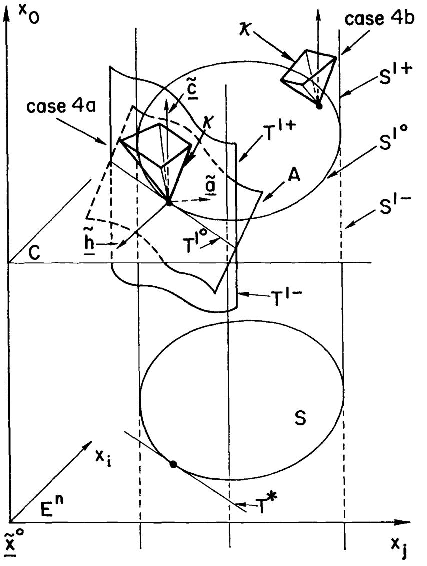


FIG. 4. *Illustration for proof of Theorem 3*

Then, in either case the following theorem holds.

THEOREM 3. *Consider the problem* (P) *when the constraint set S is a closed, convex, n-dimensional proper subset of E^n. Then, necessary conditions for* $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$ *to be an optimal control are that conditions* (i), (ii) *and* (iii) *of Theorem* 1 *be satisfied, and, in addition, that*

(iv) $$\tilde{\mathbf{p}}^*(t_K) = \lambda\tilde{\mathbf{h}} + \mu\tilde{\mathbf{c}},$$

where λ, μ are nonpositive constants.

Proof. (See Fig. 4.) Consider Case 4b. This case is virtually identical to that of the free right end, and Theorem 2b holds. Since $\tilde{\mathbf{h}} = 0$ in this case,

$$\tilde{\mathbf{p}}^*(t_K) = \mu\tilde{\mathbf{c}} = (\mu, 0, 0, \cdots, 0).$$

But from Theorem 2b it is seen that $\mu \leqq 0$. Therefore, Theorem 3 is true for Case 4b.

Consider Case 4a. Again, by a simple extension of Lemma 1, it can be shown that for $\mathbf{u}^*(t) = \mathbf{u}(t; \boldsymbol{\tau}^*, \boldsymbol{\mu}^*)$ to be an optimal control, $\mathcal{K}$ must be separated from D by a hyperplane A. Let a normal $\tilde{\mathbf{a}}$ to A be defined as in (14). Then, by virtue of (26),

$$\tilde{\mathbf{a}} = \lambda\tilde{\mathbf{h}} + \mu\tilde{\mathbf{c}}, \quad \lambda, \mu \leqq 0.$$

By proceeding as in Theorem 2, and letting $\tilde{\mathbf{p}}^*(t_K) = \tilde{\mathbf{a}}$, the conditions of Theorem 1 as well as condition (iv) of Theorem 3 can be shown to be necessary. This completes the proof of Theorem 3.

The conditions of Theorem 1 have been shown to be necessary for each terminal constraint set under consideration. Therefore, the proof of Theorem 1 is completed.

4. Conclusion. Although all the results in this paper were developed for discrete time systems in which the sampling instants (i.e., the components of $\boldsymbol{\tau} \in W$) are not fixed, most of these results also remain valid when the sampling instants are fixed. In this case, condition (ii) of Theorem 1 no longer applies but conditions (i) and (iii) of Theorem 1 and the transversality conditions stated in Theorems 2, 2a, 2b and 3 are still necessary.

For Case 3, the free right end case, it appears that one of the restrictions on the control constraint set U may be relaxed: it does not seem necessary that the sets $\Lambda(\mathbf{v})$ for $\mathbf{v} \in U$ be convex. This is because S^{1-} in this case is a halfspace in E^{n+1}, and the separating hyperplane is uniquely defined. Consequently, the proofs no longer depend on the convexity of $\mathcal{K}$.

The results for all of the cases considered here carry over when the control constraint set is different for each $\mathbf{v}_k$, i.e., when $\mathbf{v}_k$ must belong to U_k, provided that the sets U_k, $k = 0, 1, \cdots, K - 1$, satisfy the restrictions of §2b.

It is hoped that the results of this paper will be useful in developing optimal control systems which are more readily engineered than those operating in a continuous time mode.

Acknowledgment. The authors wish to thank Dr. L. W. Neustadt for pointing out some inconsistencies and inadequacies in the original manuscript.

REFERENCES

[1] L. I. Rozonoer, *The maximum principle of L. S. Pontryagin in optimal system theory*, Avtomat. i. Telemeh., I, II, and III, 20 (1959), pp. 1320–1334, 1411–1458, 1561–1578. [English translation in Automat. Remote Control, I, II, and III, 20 (1959), pp. 1288–1302, 1405–1421, 1517–1532.]

[2] S. Katz, *A discrete version of Pontryagin's maximum principle*, J. Electronics Control, 13, 2 (1962), pp. 179–184.

[3] ———, *Best operating points for staged systems*, Ind. Eng. Chem. Fundamentals, 1 (1962), pp. 226–240.

[4] S. S. L. Chang, *Digitized maximum principle*, Proc. IRE, December 1960, pp. 2030–2031.

[5] ———, *Computer optimization of nonlinear control systems by means of digitized maximum principle*, IRE Intern. Conv. Record, 9 (1961), pp. 48–55.

[6] C. A. Desoer, E. Polak, and J. Wing, *Paper No.* 406, Proc. Second Intern. Congress IFAC on EC, Basel, Switzerland, 1963.

[7] L. S. Pontryagin, et al., *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

SOME APPLICATIONS OF STOCHASTIC DIFFERENTIAL EQUATIONS TO OPTIMAL NONLINEAR FILTERING*

W. M. WONHAM†

1. Introduction. A current problem in control theory is that of estimating the dynamical state of a physical system, on the basis of data perturbed by noise. Solution of the estimation problem is usually immediate if one knows the probability distribution of the system state at each instant of time, conditional on the data available up to that instant. It is therefore of interest to ask how this *posterior* probability distribution evolves with time, and if possible to specify the dynamical structure of a filter (i.e., analog device) which generates the posterior distribution when its input is the time function actually observed.

In the present paper, filters of this type are defined by means of stochastic differential equations[1] for the posterior distribution in which the observed time function appears as a forcing term. Differential equations for this purpose were introduced in 1960 by Stratonovič [1], who also indicated their application to stochastic control problems [2]. When the dynamical system under observation is linear and the noise is white Gaussian it has been shown [3] that Stratonovič's equation can be solved formally to yield the stochastic differential equation of the optimal (linear) filter. When the function to be estimated is a Markov step process and the noise is white Gaussian the optimal (nonlinear) filter equations were stated in [4]. The latter equations are discussed in more detail in §3, below; they differ from those of Stratonovič in a sense to be noted in the sequel. For one example, discussed in §3, performance of the optimal nonlinear filter is evaluated numerically and is found to be somewhat better than that of the simpler Wiener filter, particularly when the noise intensity is low.

In §4, the equations of §3 are generalized heuristically to the case where the state space of the step process is continuous, and in §5 some tentative remarks are made on the form of the solutions.

Some parallel work on noisy observation of a diffusion process has been reported by Kushner in a recent paper [5].

* Received by the editors April 30, 1964, and in revised form November 3, 1964.

† Center for Control Theory, RIAS, Baltimore, Maryland. Now at Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island. This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research, under Contract No. AF 49(638)-1206, and in part by the National Aeronautics and Space Administration under Contract No. NASw-845.

[1] A brief review of stochastic differential equations is given in Appendix 1.

2. Noisy measurement of an unknown constant. The basic idea of a "functional" filter is illustrated by the following simple estimation problem. Let x be a discrete, real-valued random variable with range of values a_1, $\cdots$, a_K and a priori probability distribution $\{p_j(0), j = 1, \cdots, K\}$ at $t = 0$. Suppose that one observes the function

$$y(t) = xt + \int_0^t \beta(s)\, dw(s), \qquad t \geqq 0, \tag{1}$$

where the function β is known to the observer,[2] and $\{w(t), t \geqq 0\}$ is a Wiener process which is independent of x, with[3] $P\{w(0) = 0\} = 1$. The process $y(t)$ defined by (1) can also be written as the solution of the stochastic differential equation

$$dy(t) = x\, dt + \beta(t)\, dw(t), \quad y(0) = 0. \tag{2}$$

Dividing formally by dt one obtains the possibly more familiar version

$$\dot{y}(t) = x + \beta(t)\dot{w}(t), \quad y(0) = 0, \tag{3}$$

where $\dot{w}$ represents Gaussian white noise. Since w is not differentiable in the ordinary sense we shall use instead the differential notation of (2) and interpret (2) to mean the integral representation (1) (see Appendix 1).

Let us now introduce the posterior distribution

$$p_j(t) = P\{x = a_j \mid y(s), 0 \leqq s \leqq t\}, \qquad j = 1, \cdots, K. \tag{4}$$

Evaluation of $p_j(t)$ is straightforward (see Appendix 2); the result is[4]

$$p_j(t) = \frac{p_j(0) \exp\left[a_j \int_0^t \beta(s)^{-2}\, dy(s) - \frac{1}{2} a_j^2 \int_0^t \beta(s)^{-2}\, ds\right]}{\sum_{k=1}^{K} p_k(0) \exp\left[a_k \int_0^t \beta(s)^{-2}\, dy(s) - \frac{1}{2} a_k^2 \int_0^t \beta(s)^{-2}\, ds\right]}. \tag{5}$$

The stochastic integral in (5) is well-defined [6, Chap. IX, §2]. Now write $p(t) = [p_1(t), \cdots, p_K(t)]$ and consider the joint process $\{x, p(t), t \geqq 0\}$, where we regard x as a fixed random variable with distribution $\{p_j(0), j = 1, \cdots, K\}$. Since almost every $w(t)$ sample function is continuous, the same is true of the $p(t)$ sample functions (by [6, Chap. IX, Theorem

[2] We shall assume that β is continuously differentiable, and bounded away from 0 for $t \geqq 0$.

[3] It is assumed that an underlying probability space with elementary events ω is given. Probability measure on this space is denoted by P. To simplify notation, ω-dependence of random variables will not be indicated. In the Appendices "Borel field" means "Borel field of ω-sets".

[4] The qualification "with probability 1" on equalities between conditional probabilities is to be understood.

5.2]). Moreover, it is easily seen (Appendix 3) that the $\{x, p(t)\}$ process is Markov.

Our aim is to describe the evolution in time of the p_j's by means of a system of stochastic differential equations. Having verified the existence of the limits (7) and (8) written below, one can apply to the Markov process $\{x, p(t)\}$ a representation theorem of Doob [6, Chap. VI, Theorem 3.3]. Alternatively, since the $p_j(t)$ are known explicitly, it is more direct to apply a result of Dynkin (Appendix 1 or [7, Theorem 7.2]), and this gives

$$dp_j(t) = m_j[t, x, p(t)]\,dt + \sigma_j[t, x, p(t)]\,dw(t), \qquad j = 1, \cdots, K. \tag{6}$$

The functions m_j and σ_j in (6) have the probabilistic meaning

$$m_j(t, \xi, \pi) = \lim_{h \to 0} E\left\{\frac{p_j(t+h) - p_j(t)}{h} \,\middle|\, x = \xi, p(t) = \pi\right\}, \tag{7}$$

and

$$\begin{aligned} &\sigma_i(t, \xi, \pi)\sigma_j(t, \xi, \pi) \\ &= \lim_{h \to 0} E\left\{\frac{[p_i(t+h) - p_i(t)][p_j(t+h) - p_j(t)]}{h} \,\middle|\, x = \xi, p(t) = \pi\right\}; \\ &\qquad\qquad i, j = 1, \cdots, K. \end{aligned} \tag{8}$$

In Appendix 3 the limits (7) and (8) are computed from (5), using Dynkin's formulas; the results are

$$m_j(t, x, p) = \beta(t)^{-2}(x - \bar{x})(a_j - \bar{x})p_j, \tag{9}$$

$$\sigma_j(t, x, p) = \beta(t)^{-1}(a_j - \bar{x})p_j, \tag{10}$$

$j = 1, \cdots, K$, where

$$\bar{x} = \sum_{k=1}^{K} a_k p_k. \tag{11}$$

Writing out (6) in full and noting (2), we obtain finally

$$\begin{aligned} dp_j(t) = {} & -\beta(t)^{-2}\bar{x}(t)[a_j - \bar{x}(t)]p_j(t)\,dt \\ & + \beta(t)^{-2}[a_j - \bar{x}(t)]p_j(t)\,dy(t), \qquad j = 1, \cdots, K. \end{aligned} \tag{12}$$

The system of stochastic differential equations (12) is the desired result. It can be interpreted as specifying the dynamical structure of a filter (or analog device) which continuously generates the posterior distribution $p(t)$ when the input is the observed function $y(t)$. From a practical viewpoint this interpretation might be useful when the function actually observed is not y, but rather some approximation to the "function" $\dot{y}$ indicated in (3). In that case formal division by dt in (12) yields a system of nonlinear differential equations for the p_j's, in which $\dot{y}$ appears as a forcing term. This system of equations can be simulated by an analog device with input $\dot{y}$ and output p, and thus represents the filter.

The Ito equation (12) is solved (Appendix 1) by replacing (12) by an integral equation and constructing the solution q by successive approximations. It is plausible that the structure of the analog feedback device should be chosen to model that of the successive approximation scheme. However, it is not yet known precisely in what sense a physical process must approximate a $\dot{y}$ "process" in order that the filter output be a useful approximation to the desired posterior distribution. Experimental work along these lines would be of considerable interest.

Example. It is worth emphasizing that the ordinary rules of integration cannot be applied to the stochastic equation (12) in an attempt to regain the explicit solution (5). To illustrate this fact let x have possible values $a_1 = +1$, $a_2 = -1$ and let $p_1(0) = p_2(0) = \frac{1}{2}$, $\beta \equiv 1$. Since $p_1(t) + p_2(t) \equiv 1$, it is sufficient to consider

$$q(t) = p_1(t) - p_2(t). \tag{13}$$

From (11), $\bar{x}(t) = q(t)$; and from (12) the filter is defined by

$$dq = -q(1 - q^2)\,dt + (1 - q^2)\,dy. \tag{14}$$

On the other hand (5) yields the evaluation

$$q(t) = \tanh\,[y(t)], \tag{15}$$

and formal differentiation of (15) gives

$$dq = (1 - q^2)\,dy. \tag{16}$$

We observe that the first term on the right side of (14) is absent from (16). The point is simply that stochastic differential equations of Ito type cannot be manipulated by the usual formal rules (cf. [6, Chap. IX, §5]); also, it is plausible that an analog device for generating q should be set up according to the Ito equation (14), and not according to the "formal" equation (16).

It must be noted finally that Stratonovič's procedure [1] applied to this example leads to (16) and not to (14) (cf. [2, (9)]). In general it appears that the coefficient of dt in Stratonovič's equations (when these are written in differential notation) cannot be given the interpretation indicated in (7).

3. Noisy observations of a Markov step process. Let $\{x(t), t \geqq 0\}$ be a stationary Markov step process[5] with a finite number of states (distinct

[5] For a discussion of the process see [6, Chap. VI, §1], where our ν_i, ν_{ij} are denoted by q_i, q_{ij}.

step levels) $a_1, \cdots, a_K$. Denote the transition probabilities by

$$p_{ij}(h) = P\{x(t + h) = a_j | x(t) = a_i\}.$$

We assume that

$$p_{ij}(h) = \begin{cases} 1 - \nu_i h + o(h), j = i & (h \to 0) \\ \nu_{ij} h + o(h), j \neq i & (h \to 0), \end{cases} \tag{17}$$

where the $\nu_{ij} \geqq 0$ are constants and

$$\nu_i = \sum_{\substack{j=1 \\ j \neq i}}^{K} \nu_{ij}, \; i = 1, \cdots, K. \tag{18}$$

Let the distribution of $x(0)$ be $\{p_j(0), j = 1, \cdots, K\}$. To avoid triviality we assume that $\{p_j(0)\}$ is not concentrated on any absorbing state among the a_j's.

As in §2, suppose that the process observed is $\{y(t), t \geqq 0\}$ defined by

$$dy(t) = x(t)\, dt + \beta(t)\, dw(t), \qquad t \geqq 0, \tag{19}$$

where $P\{y(0) = 0\} = 1$ and the Wiener process $\{w(t), t \geqq 0\}$ is independent of the $x(t)$ process. Introduce the posterior probabilities

$$p_j(t) = P\{x(t) = a_j | y(s), 0 \leqq s \leqq t\}, \qquad j = 1, \cdots, K. \tag{20}$$

We now seek a system of stochastic differential equations for the $p_j(t)$. In contrast to the situation in §2, a representation of the $p(t)$ process as an explicit functional of the $y(t)$ process (cf. (5)) is apparently no longer available. However, the appropriate generalization of (12) can be obtained by an application of Doob's theorem [6, Chap. VI, Theorem 3.3]. The equations, derived in Appendix 4, are

$$\begin{aligned} dp_j(t) = {} & [-\nu_j p_j(t) + \sum_{\substack{i=1 \\ i \neq j}}^{K} \nu_{ij} p_i(t)]\, dt - \beta(t)^{-2} \bar{x}(t)[a_j - \bar{x}(t)] p_j(t)\, dt \\ & + \beta(t)^{-2}[a_j - \bar{x}(t)] p_j(t)\, dy(t), \qquad j = 1, \cdots, K. \end{aligned} \tag{21}$$

Equation (21) is the main result of this paper. The previous remarks on the interpretation of (12) apply also to (21): the equation defines the structure of an ideal analog device for generating the p_j's from the data, y. Comparison of (21) with (12) shows that the only new term in (21) is the first term on the right side. This term is of the form $\mathcal{L}^+[p]\, dt$ where $\mathcal{L}^+$ is the forward operator of the $x(t)$ process, and thus represents a change in p due to the observer's a priori knowledge of how the $x(t)$ process evolves.

We shall now discuss a simple special case of (21) in detail.

Example. Let $\beta \equiv$ const. and suppose that x is a "random telegraph sig-

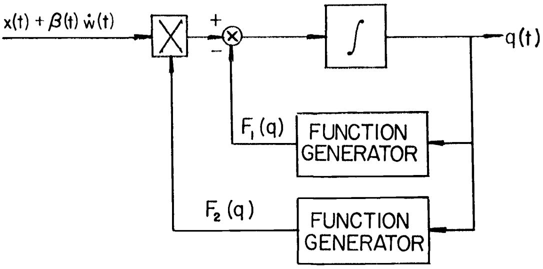


FIG. 1. *Optimal nonlinear filter: Example*, §3.

$$F_1(q) = 2\nu q + \beta^{-2}q(1 - q^2)$$

$$F_2(q) = \beta^{-2}(1 - q^2)$$

nal" [11]; that is,

$$(22) \qquad \begin{aligned} a_1 &= +1, \quad a_2 = -1, \\ \nu_i &= \nu_{ij} = \nu, \qquad i, j = 1, 2. \end{aligned}$$

The parameter ν is the expected number of jumps of $x(\cdot)$ in unit time. Let

$$(23) \qquad q(t) = p_1(t) - p_2(t).$$

From (21) and (22),

$$(24) \qquad dq = -2\nu q\, dt - \beta^{-2}q(1 - q^2)\, dt + \beta^{-2}(1 - q^2)\, dy,$$

or equivalently

$$(25) \qquad dq = [-2\nu q - \beta^{-2}(1 - q^2)(q - x)]\, dt + \beta^{-1}(1 - q^2)\, dw.$$

A block diagram of the optimal filter is shown in Fig. 1.

We shall evaluate the filter performance in terms of mean square estimation error. Thus the optimal estimate of $x(t)$ is

$$(26) \qquad \begin{aligned} \hat{x}(t) &= E\{x(t) \mid y(s), 0 \leqq s \leqq t\} \\ &= a_1 p_1(t) + a_2 p_2(t) \\ &= q(t). \end{aligned}$$

In Appendix 4 it is shown that the joint process $\{x(t), q(t), t \geqq 0\}$ is Markov. It will be assumed that this process has stationary densities $\pi^{\pm}(q)$, $-1 \leqq q \leqq 1$, defined by

$$(27) \qquad \pi^{\pm}(q)\, dq = P\{x(t) = \pm 1, q(t) \in (q, q + dq)\}.$$

Then the mean square estimation error is

$$\sigma^2 = \int_{-1}^{1} (1-q)^2\pi^+(q)\,dq + \int_{-1}^{1} (1+q)^2\pi^-(q)\,dq. \tag{28}$$

It remains to calculate the densities $\pi^\pm$. From (25), the stationary Kolmogorov forward equation [6, Chap. VI] of the $\{x, q\}$ process is

$$\begin{aligned}\tfrac{1}{2}\beta^{-2}[(1-q^2)^2\pi^\pm(q)]'' - [\beta^{-2}(\pm 1 - q)(1-q^2)\pi^\pm(q)\\ - 2\nu q\pi^\pm(q)]' \pm \nu[\pi^-(q) - \pi^+(q)] = 0,\end{aligned} \tag{29}$$

where $'$ denotes d/dq. With the symmetry condition $\pi^-(q) = \pi^+(-q)$, (29) has the unique solution

$$\pi^\pm(q) = c(1 \pm q)(1-q^2)^{-2} \exp[-2\mu(1-q^2)^{-1}], \tag{30}$$

where

$$c = \left[2\int_1^\infty z^{1/2}(z-1)^{-1/2}e^{-2\mu z}\,dz\right]^{-1} \tag{31}$$

and

$$\mu = \beta^2\nu. \tag{32}$$

From (28), (30) the stationary error variance is then

$$\begin{aligned}\sigma^2 &= \frac{\int_0^\infty z^{-1/2}(z+1)^{-1/2}e^{-2\mu z}\,dz}{\int_0^\infty z^{-1/2}(z+1)^{1/2}e^{-2\mu z}\,dz}\\ &= \begin{cases}-2\mu\log 2\mu + o(\mu\log\mu) & \text{as } \mu \to 0,\\ 1 - (4\mu)^{-1} + O(\mu^{-2}) & \text{as } \mu \to \infty.\end{cases}\end{aligned} \tag{33}$$

The result (33) will be compared with the error variance of a Wiener filter which is optimal for the same input. For the Wiener filter a standard computation yields

$$\begin{aligned}\sigma_w^2 &= 2\mu[(1+\mu^{-1})^{1/2} - 1]\\ &= \begin{cases}2\mu^{1/2} + o(\mu^{1/2}) \text{ as } \mu \to 0,\\ 1 - (4\mu)^{-1} + O(\mu^{-2}) \text{ as } \mu \to \infty.\end{cases}\end{aligned} \tag{35}$$

Numerical results are given in Fig. 2. Since the nonlinear filter generates the estimate $\hat{x}(t)$ defined by (26) it is necessarily optimal (with respect to error variance) in the class of all filters which operate on the present and past of the data y. Thus $\sigma^2 \leqq \sigma_w^2$; in fact, σ^2/σ_w^2 is substantially less than unity when the noise intensity is low.

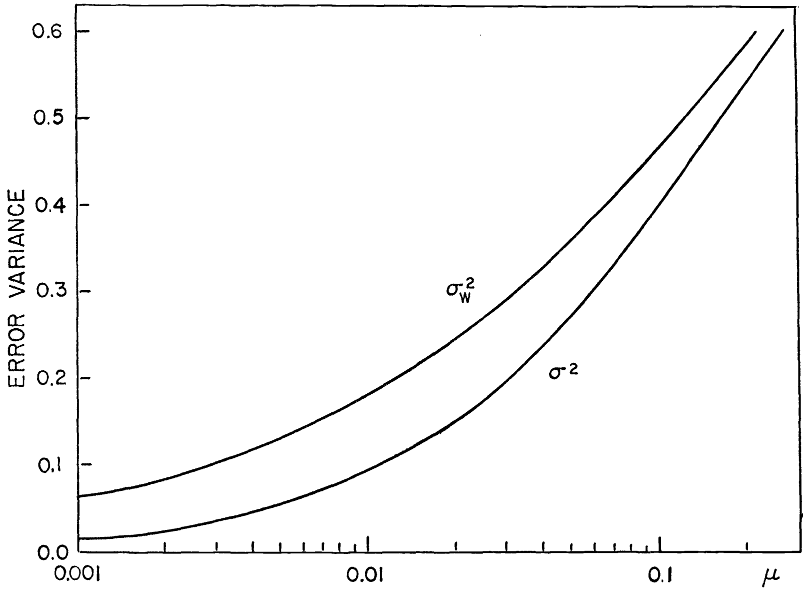


FIG. 2. *Numerical results for example,* §3.

4. Heuristic generalization of (21) **to a continuous state space.** The differential equations (12), (21) were derived on the assumption that the state space of the $x(t)$ process is a finite set. The following is a heuristic generalization to a continuous state space. Let $\{x(t), t \geqq 0\}$ be a real-valued Markov step process with state space X, where X is a closed finite interval. Let $\nu(\xi)$, $\nu(\xi, A)$ be defined for ξ in X and A a Borel subset of X; the functions $\nu(\cdot)$, $\nu(\cdot\ ,\ \cdot)$ are assumed to be a "standard pair" in the sense of Doob [6, Chap. VI, §2][6]. In addition, $\nu(\cdot)$ is assumed to be bounded on X; the $x(t)$ sample functions are then almost all step functions [6]. In analogy to (17) one has

$$p(h, \xi, A) = P\{x(t+h) \in A \mid x(t) = \xi\}$$
$$(36) \qquad = \begin{cases} \nu(\xi, A)h + o(h), \ \xi \notin A, \\ 1 - \nu(\xi)h + o(h), \ A = \{\xi\}. \end{cases}$$

As in §3, suppose next that

$$(37) \qquad dy(t) = x(t)\,dt + \beta(t)\,dw(t), \qquad t \geqq 0,$$

and introduce the posterior probability

$$(38) \qquad p(t, A) = P\{x(t) \in A \mid y(s), 0 \leqq s \leqq t\}.$$

[6] The functions ν are denoted in [6] by q.

Then inspection of (21) suggests the generalization

$$
\begin{aligned}
dp(t, A) = {} & \left[-\int_A \nu(\xi, X - A)p(t, d\xi) + \int_{X-A} \nu(\xi, A)p(t, d\xi)\right] dt \\
& - \left[\beta(t)^{-2}\bar{x}(t) \int_A [\xi - \bar{x}(t)]p(t, d\xi)\right] dt \\
& + \left[\beta(t)^{-2} \int_A [\xi - \bar{x}(t)]p(t, d\xi)\right] dy(t).
\end{aligned} \tag{39}
$$

In (39),

$$
\bar{x}(t) = \int_X \xi p(t, d\xi), \tag{40}
$$

and A is an arbitrary Borel subset of X.

Just as in the case of a (finite-dimensional) Ito equation, (39) might plausibly be interpreted by starting from the corresponding integral equation, obtained by integrating both sides of (39) with respect to t, and seeking the solution as the limit of successive approximations. Unlike (12) and (21), however, the general equation (39) cannot readily be interpreted as specifying the dynamics of a practically realizable filter for generating p from the data y.

5. Sufficient statistics. We have seen in §4 that the stochastic differential equation for the posterior distribution of $x(t)$ cannot be used directly, in general, to design an optimal filter. In practice, construction of the posterior distribution must be reduced to the evaluation of a "small" number of functionals (sufficient statistics) on the observed function y. For example, inspection of (5) shows that the K-dimensional stochastic system (12) can be replaced by the 1-dimensional system

$$
dz(t) = \beta(t)^{-2}\, dy(t), \quad z(0) = 0, \tag{41}
$$

which generates the sufficient statistic

$$
z(t) = \int_0^t \beta(s)^{-2}\, dy(s). \tag{42}
$$

On the other hand the writer knows of no similar reduction of the system (21) or of (39).

Even if a nontrivial sufficient statistic $z(t)$ for the determination of p exists, it may be impossible to write z as an explicit functional of y. It would be enough to know, however, that z satisfies a stochastic differential equation

$$
dz = \zeta_1(t, z)dt + \zeta_2(t, z)\, dy, \tag{43}
$$

where the functions ζ_1, ζ_2 are known; then in principle z could be obtained as the output of an analog device set up according to (43). Thus the "solution" of an equation of type (21), (39) might take the form of a (known) function of a statistic z which satisfies a (known) equation of type (43). The investigation of solutions of this type (if they exist) would be of considerable interest.

Acknowledgment. The author is indebted to H. J. Kushner (RIAS) and to J. Ternan (Cambridge University) for helpful discussions.

Appendix 1. Stochastic differential equations.

1. Since stochastic differential equations are not yet widely used in engineering applications we mention here some definitions and known results. For a detailed account the reader is referred to Doob [6, Chap. VI, §3] and Dynkin [7, Chaps. 7, 11].

Let $\{z(t),\ t \geqq 0\}$ be a stochastic process in K-dimensional Euclidean space R^K; we write $z = (z_1, \cdots, z_K)$, where the z_i are real-valued, and put $\| z \| = (\sum_1^K z_i^2)^{1/2}$. Let $\{w(t),\ t \geqq 0\}$ be a Wiener process (Brownian motion process) in R^J; that is, $w(t) = [w_1(t), \cdots, w_J(t)]$ where the $w_i(t)$ are independent Wiener processes in R^1 and, for $t \geqq s \geqq 0$,

$$E\{[w_i(t) - w_i(s)][w_j(t) - w_j(s)]\} = \begin{cases} t - s, j = i, \\ 0, j \neq i. \end{cases} \tag{44}$$

(See [6, Chap. II, §9] for the definition of the Wiener process in R^1.)

The stochastic differential equation of interest here is written

$$dz(t) = m[t, z(t)]\,dt + \sigma[t, z(t)]\,dw(t), \qquad t \geqq 0, \tag{45}$$

where m is a K-vector and σ is a $K \times J$ matrix. Loosely interpreted, (45) states that in a small time interval $(t,\ t + dt)$ the vector $z(t)$ suffers a "dynamical" displacement $m[t,\ z(t)]\,dt$ plus a random displacement $\sigma[t,\ z(t)]\,dw(t)$, where the latter is a Gaussian random vector with mean 0 and covariance matrix $\sigma[t,\ z(t)]\sigma'[t,\ z(t)]\,dt$ ($'$ denotes transpose).

Dividing both sides of (45) formally by dt one obtains

$$\dot{z} = m(t, z) + \sigma(t, z)\dot{w} \qquad (\cdot = d/dt), \tag{46}$$

where $\dot{w}$ is a J-vector whose components are independent "Gaussian white noise processes". The notation of (46) has been more common in the engineering literature than that of (45) but it is objectionable for two reasons:

(a) almost all $w(t)$ sample functions are nondifferentiable almost everywhere, a fact which is intuitively plausible if we note that $E \| dw(t) \|$ is proportional to $(dt)^{1/2}$;

(b) when applied to (46) the usual formal rules of integration lead in general to results which are definitely incorrect. This statement will be illustrated later.

2. The stochastic differential equation (45) does not specify the value of a derivative. According to Gihman [9], such an equation can be defined by giving an explicit procedure for constructing the solution, but then has meaning only insofar as this construction can be carried out. An alternative interpretation of (45) due to Ito [8] is the following. Replace (45) by the stochastic integral equation

$$z(t) = z(0) + \int_0^t m[s, z(s)]\, ds + \int_0^t \sigma[s, z(s)]\, dw(s). \tag{47}$$

The stochastic integrals on the right side of (47) are defined [6, Chap. IX, §2, §5] under suitable restrictions on the (random) functions $m[\cdot, z(\cdot)]$ and $\sigma[\cdot, z(\cdot)]$. Ito's construction of a $z(t)$ process which satisfies (47) is carried out by successive approximation; one sets $z^{(0)}(t) \equiv 0$ and

$$z^{(n+1)}(t) = z(0) + \int_0^t m[s, z^{(n)}(s)]\, ds + \int_0^t \sigma[s, z^{(n)}(s)]\, dw(s), \qquad n = 0, 1, 2, \cdots . \tag{48}$$

Conditions which guarantee that the sequence $\{z^{(n)}\}$ converges for t in a finite interval $[0, T]$ are given by Doob [6, IX, §3][7] and are, mainly, that the functions $m(t, z)$, $\sigma(t, z)$ satisfy a uniform Lipschitz condition in z, and are bounded in norm by $C(1 + \| z \|^2)^{1/2}$ where C is some constant. Then there exists a process $\{z(t), 0 \leqq t \leqq T\}$ with the following properties.
(i) $\lim_{n\to\infty} z^{(n)}(t) = z(t)$ uniformly in t with probability 1; and the $z(t)$ process is unique, in the sense that the difference of two solutions is zero, with probability 1, for each t.
(ii) The $z(t)$ sample functions are, with probability 1, continuous in $[0, T]$.
(iii) For each $t \in [0, T]$, (47) is true with probability 1.
(iv) If the initial value $z(0)$ is a random variable which is independent of the increments $\{w(t_2) - w(t_1), t_1, t_2 \in [0, T]\}$, then $\{z(t), 0 \leqq t \leqq T\}$ is a Markov process.

In addition, the $z(t)$ process has the following local properties, which make precise the interpretation of (45) given earlier:

$$\text{(v)}\quad \lim_{h\to 0} E\left\{\frac{z(t+h) - z(t)}{h} \,\middle|\, z(t) = \zeta\right\} = m(t, \zeta), \tag{49}$$

$$\begin{aligned}\text{(vi)}\quad \lim_{h\to 0} E&\left\{\frac{[z_i(t+h) - z_i(t)][z_j(t+h) - z_j(t)]}{h} \,\middle|\, z(t) = \zeta\right\} \\ &= \sum_{r=1}^{J} \sigma_{ir}(t, \zeta)\sigma_{jr}(t, \zeta) \\ &= b_{ij}(t, \zeta), \quad \text{say}; \qquad i, j = 1, \cdots, K.\end{aligned} \tag{50}$$

[7] Doob's treatment for $K = J = 1$ can be generalized after replacing $| m |$ by $\| m \|$ and $| \sigma |$ by $\| \sigma \| = (\sum_i\sum_j \sigma_{ij}^2)^{\frac{1}{2}}$ (cf. [7, Chap. 7]).

3. Assume for the moment that the Markov process $\{z(t),\ 0 \leqq t \leqq T\}$ constructed above has a transition probability density $p = p(s, z;\ t, \zeta)$, defined for $0 \leqq s < t \leqq T$ and $z, \zeta \in R^K$. It would be convenient if the hypotheses made on m and σ (strengthened to include the differentiability needed below) guaranteed that the density p exists and satisfies the Kolmogorov equations

$$(51)\qquad -\frac{\partial p}{\partial s} = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} b_{ij}(s, z)\,\frac{\partial^2 p}{\partial z_i\,\partial z_j} + \sum_{i=1}^{K} m_i(s, z)\,\frac{\partial p}{\partial z_i}$$

$$(52)\qquad \frac{\partial p}{\partial t} = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\partial^2}{\partial \zeta_i\,\partial \zeta_j}\,[b_{ij}(t, \zeta)p] - \sum_{i=1}^{K} \frac{\partial}{\partial \zeta_i}\,[m_i(t, \zeta)p].$$

Unfortunately it does not seem possible to establish (51) and (52) without making a priori assumptions on the smoothness of p (see the discussion in [6, Chap. VI, §3]). Nevertheless, if the $z(t)$ process obtained by solving the Ito equation (45) has a transition density which satisfies the Kolmogorov equations, then the coefficients which appear in the latter are related to the functions m and σ of the Ito equation according to (49)–(52) above. Used heuristically, this correspondence between (45) and (51), (52) is convenient in engineering applications (see, e.g., [10]).

4. The following example[8] shows that Ito equations cannot be manipulated by the ordinary rules of integration (cf. also [6, p. 443]). Let $\{w(t),\ t \geqq 0\}$ be a Wiener process in R^1 with $P\{w(0) = 0\} = 1$. Let

$$(53)\qquad z(t) = e^{w(t)}, \qquad t \geqq 0.$$

It can be shown that the $z(t)$ process can be represented as the solution of an Ito equation (45). From (49), (50) we find

$$(54)\qquad m(\zeta) = \tfrac{1}{2}\zeta, \quad \sigma(\zeta) = \zeta,$$

so that

$$(55)\qquad dz(t) = \tfrac{1}{2}z(t)\,dt + z(t)\,dw(t), \qquad t \geqq 0,$$

with initial condition $P\{z(0) = 1\} = 1$. On the other hand if (55) is replaced by

$$(56)\qquad \dot{z} = \tfrac{1}{2}z + z\dot{w}, \qquad z(0) = 1,$$

and the last equation integrated formally, the result is

$$(57)\qquad z(t) = e^{\frac{1}{2}t + w(t)}.$$

5. It is useful to know under what conditions a given process $\{\psi(t),\ t \geqq 0\}$ can be represented as the solution of an Ito equation (45) or equivalently

[8] Suggested to the writer by H. J. Kushner.

of the integral equation (47). One such representation theorem is given by Doob [6, Chap. VI, Theorem 3.3]. We will state here a slightly specialized version of a theorem of Dynkin [7, Theorem 7.2].

THEOREM. *Let the process* $\{z(t),\ t \geqq 0\}$ *satisfy an integral equation of form* (47) (*in particular we can have* $z(t) \equiv w(t)$) *and let* $\phi = \phi(t, \zeta)$ *be a numerical function, twice continuously differentiable in* (t, ζ) *for* $t \geqq 0$ *and* ζ *in* R^K. *Put* $\psi(t) \equiv \phi[t, z(t)]$. *Then the process* $\{\psi(t),\ t \geqq 0\}$ *satisfies the integral equation*

$$\psi(t) = \psi(0) + \int_0^t \tilde{m}[s, z(s)]\, ds + \int_0^t \sum_{r=1}^{J} \tilde{\sigma}_r[s, z(s)]\, dw_r(s). \tag{58}$$

The functions $\tilde{m}$ *and* $\tilde{\sigma}_r$ *are given by*

$$\begin{aligned}\tilde{m}(t, \zeta) = \frac{\partial \phi(t, \zeta)}{\partial t} + \sum_{i=1}^{K} \frac{\partial \phi(t, \zeta)}{\partial \zeta_i}\, m_i(t, \zeta)\\ + \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\partial^2 \phi(t, \zeta)}{\partial \zeta_i \partial \zeta_j}\, b_{ij}(t, \zeta),\end{aligned} \tag{59}$$

where

$$b_{ij}(t, \zeta) = \sum_{r=1}^{J} \sigma_{ir}(t, \zeta) \sigma_{jr}(t, \zeta), \tag{60}$$

and

$$\tilde{\sigma}_r(t, \zeta) = \sum_{i=1}^{K} \frac{\partial \phi(t, \zeta)}{\partial \zeta_i}\, \sigma_{ir}(t, \zeta), \qquad r = 1, \cdots, J. \tag{61}$$

Equation (58) is (by definition) equivalent to the stochastic differential equation

$$d\psi(t) = \tilde{m}[t, z(t)]\, dt + \sum_{i=1}^{J} \tilde{\sigma}_r[t, z(t)]\, dw_r(t). \tag{62}$$

Equation (62) is more general than (45) in the sense that $\tilde{m}$ and the $\tilde{\sigma}_r$ may not be expressible as functions of (t, ψ); however, by adjoining (62) to (45) we obtain a new system of the same type as before.

Finally, it is seen that (62) can also be written

$$\begin{aligned}d\psi(t) = \left\{\frac{\partial \phi[t, z(t)]}{\partial t} + \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\partial^2 \phi[t, z(t)]}{\partial \zeta_i \partial \zeta_j}\, b_{ij}[t, z(t)]\right\} dt\\ + \sum_{i=1}^{K} \frac{\partial \phi[t, z(t)]}{\partial \zeta_i}\, dz_i(t).\end{aligned} \tag{62'}$$

Equation (62′) exhibits the "chain rule" explicitly.

Appendix 2. Derivation of (5). Let $t > 0$ be fixed and put $s_{rn} = rt/n$, $r = 0, 1, \cdots, n$. It will be verified that $\hat{p}_j(t)$, defined by

$$(63) \qquad \hat{p}_j(t) = \underset{n\to\infty}{\text{l.i.m.}}\, P\{x = a_j \mid y(s_{rn}),\ r = 0, 1, \cdots, n\},$$

is given by the expression on the right side of (5); and then that $\hat{p}_j(t)$ is actually the conditional probability $p_j(t)$ defined by (4). Put $\eta_{rn} = y(s_{rn}) - y(s_{r-1,n})$, ($r = 1, \cdots, n$; $n = 1, 2, \cdots$). Then from (1),

$$(64) \qquad \eta_{rn} = \frac{xt}{n} + \int_{(r-1)t/n}^{rt/n} \beta(s)\, dw(s).$$

Thus for each n the random variables $\eta_{rn} - xt/n$, $r = 1, \cdots, n$, are independent and Gaussian with mean 0 and variance

$$(65) \qquad v_{rn} = \int_{(r-1)t/n}^{rt/n} \beta(s)^2\, ds.$$

Defining

$$(66) \qquad \hat{p}_j^{(n)}(t) = P\{x = a_j \mid y(s_{rn}),\ r = 0, 1, \cdots, n\},$$

we have

$$\begin{aligned} \hat{p}_j^{(n)}(t) &= P\{x = a_j \mid \eta_{rn},\ r = 1, \cdots, n\} \\ &= \frac{p_j(0) \exp\left[-\sum_{r=1}^{n} \left(\eta_{rn} - a_j \frac{t}{n}\right)^2 (2v_{rn})^{-1}\right]}{\sum_{k=1}^{K} p_k(0) \exp\left[-\sum_{r=1}^{n} \left(\eta_{rn} - a_k \frac{t}{n}\right)^2 (2v_{rn})^{-1}\right]} \\ &= \frac{p_j(0) \exp\left[a_j \left(\frac{t}{n}\right) \sum_{r=1}^{n} \eta_{rn} v_{rn}^{-1} - a_j^2 \sum_{r=1}^{n} \left(\frac{t}{n}\right)^2 (2v_{rn})^{-1}\right]}{\sum_{k=1}^{K} p_k(0) \exp\left[a_k \left(\frac{t}{n}\right) \sum_{r=1}^{n} \eta_{rn} v_{rn}^{-1} - a_k^2 \sum_{r=1}^{n} \left(\frac{t}{n}\right)^2 (2v_{rn})^{-1}\right]} \end{aligned} \qquad (67)$$

By definition of the stochastic integral ([6, Chap. IX, §2]) and the continuity of $\beta(s)^{-2}$, $0 \leqq s \leqq t$, there follows

$$(68) \qquad \underset{n\to\infty}{\text{l.i.m.}} \left(\frac{t}{n}\right) \sum_{r=1}^{n} \eta_{rn} v_{rn}^{-1} = \int_0^t \beta(s)^{-2}\, dy(s)$$

and

$$(69) \qquad \lim_{n\to\infty} \sum_{r=1}^{n} \left(\frac{t}{n}\right)^2 v_{rn}^{-1} = \int_0^t \beta(s)^{-2}\, ds.$$

From (67)–(69) it follows that $\hat{p}_j(t)$ coincides with the expression given by (5). It is clear that the same result is obtained with any sequence

$\{s_{rn}\}$ such that $0 = s_{0n} < s_{1n} < \cdots < s_{nn} = t$ $(n = 1, 2, \cdots)$, and

$$\max_{1 \leqq r \leqq n} (s_{rn} - s_{r-1,n}) \to 0 \quad \text{as} \quad n \to \infty.$$

Let $\mathfrak{F}_t$ be the smallest Borel field[9] with respect to which the random variables $y(s)$, $0 \leqq s \leqq t$, are measurable. To see that $\hat{p}_j(t) = p_j(t)$, note first that $\hat{p}_j(t)$ is certainly measurable relative to $\mathfrak{F}_t$. Now suppose that $0 \leqq s_\nu \leqq t$ and that A_ν is a Borel subset of R^1, $\nu = 1, \cdots, N$. If Λ is the event $[y(s_\nu) \in A_\nu, \nu = 1, \cdots, N]$ then by adjoining the s_ν's to each of the sets $(s_{1n}, \cdots, s_{nn})$, $n = 1, 2, \cdots$, and including the $y(s_\nu)$ as conditioning variables in (66), we can write

$$\int_\Lambda \hat{p}_j^{(n)}(t)\, dP = P\{\Lambda[x = a_j]\}, \qquad n = 1, 2, \cdots. \tag{70}$$

Since $\hat{p}_j^{(n)} \to \hat{p}_j(t)$ in the mean we obtain, on letting $n \to \infty$,

$$\int_\Lambda \hat{p}_j(t)\, dP = P\{\Lambda[x = a_j]\}. \tag{71}$$

Since (71) holds for every Λ of the form described, it holds for every Λ in the Borel field $\mathfrak{F}_t$ (cf. [6, Chap. I, §7]). That is, we have verified that $\hat{p}_j(t)$ has the defining properties of the conditional probability $p_j(t)$.

Appendix 3. Properties of the $p(t)$ process (5). A simple computation from (5) shows that, for $0 < \tau < t$,

$$p_j(t) = \frac{p_j(\tau) \exp\left[a_j \int_\tau^t \beta(s)^{-2}\, dy(s) - \frac{1}{2} a_j^2 \int_\tau^t \beta(s)^{-2}\, ds\right]}{\sum_{k=1}^{K} p_k(\tau) \exp\left[a_k \int_\tau^t \beta(s)^{-2}\, dy(s) - \frac{1}{2} a_k^2 \int_\tau^t \beta(s)^{-2}\, ds\right]}. \tag{72}$$

Consider the joint process $\{x, p(t), t \geqq 0\}$ where x is regarded as a fixed random variable with distribution $\{p_j(0)\}$. From (72), the vector $p(t)$ depends only on x, $p(\tau)$, and the $w(s)$ increments for $\tau < s < t$. The latter increments are independent of $p(\tau)$, and of x and the $w(s)$ increments for $0 < s < \tau$, on which $p(\tau)$ depends. It follows that the conditional distribution of x, $p(t)$ given x, $p(s)$, $0 \leqq s \leqq \tau$, is a function of x, $p(\tau)$ alone; that is, the process $\{x, p(t), t \geqq 0\}$ is Markov.

The stochastic differential equation (7) for the $p(t)$ process can be established by applying either a representation theorem of Doob [6, Chap. VI, Theorem 3.3] or a related theorem of Dynkin [7, Theorem 7.2]. In either case the theorem mentioned must be extended slightly to take account of the fact that only the $p(t)$-component of the joint $\{x, p(t)\}$

[9] See footnote 3 in §2.

process is of diffusion type (alternatively the constant component x can be regarded as a trivial diffusion process). We shall apply Dynkin's theorem, extended to the present case. From (5) we see that $p_j(t)$ is of the form

$$p_j(t) = \phi_j[t, z(t)], \tag{73}$$

where

$$\begin{aligned} z(t) &= \int_0^t \beta(s)^{-2}\,dy(s) \\ &= \int_0^t x\beta(s)^{-2}\,ds + \int_0^t \beta(s)^{-1}\,dw(s) \end{aligned} \tag{74}$$

and

$$\phi_j(t, z) = \frac{p_j(0)\exp\left[a_j z - \frac{1}{2}a_j^2\int_0^t \beta(s)^{-2}\,ds\right]}{\sum_{k=1}^K p_k(0)\exp\left[a_k z - \frac{1}{2}a_k^2\int_0^t \beta(s)^{-2}\,ds\right]}. \tag{75}$$

Since $\phi_j(t, z)$ is twice continuously differentiable in (t, z), there follows (by Appendix 1 or [7, Theorem 7.2])

$$p_j(t) - p_j(0) = \int_0^t m_j[s, x, p(s)]\,ds + \int_0^t \sigma_j[s, x, p(s)]\,dw(s). \tag{76}$$

The functions m_j, σ_j are given by

$$m_j(t, x, p) = \frac{\partial}{\partial t}\phi_j(t, z) + x\beta(t)^{-2}\frac{\partial}{\partial z}\phi_j(t, z) + \frac{1}{2}\beta(t)^{-2}\frac{\partial^2}{\partial z^2}\phi_j(t, z) \tag{77}$$

and

$$\sigma_j(t, x, p) = \beta(t)^{-1}\frac{\partial}{\partial z}\phi_j(t, z). \tag{78}$$

The probabilistic meaning of m_j, σ_j is expressed by (7) and (8). The expressions (9) and (10) are computed directly from (75), (77) and (78). Finally, the stochastic differential equation (12) is equivalent (by definition) to the integral equation (76).

Appendix 4. Derivation of (21).

1. We first evaluate $p_j(t)$. To simplify the writing of certain conditional expectations it is convenient to adjoin to the probability space of the $\{x(t), w(t)\}$ process, a "dummy" step process $\{\tilde{x}(t), t \geqq 0\}$, defined to have the same range, initial distribution and transition probabilities as the $x(t)$ process, but independent of $\{x(t)\}$ and $\{w(t)\}$.

Now let $s_{rn} = rt/n$ $(r = 0, 1, \cdots, n; n = 1, 2, \cdots)$ and put

$$\begin{aligned} \eta_{rn} &= y(s_{rn}) - y(s_{r-1,n}), \\ \xi_{rn} &= \int_{(r-1)t/n}^{rt/n} x(s)\,ds, \\ \tilde{\xi}_{rn} &= \int_{(r-1)t/n}^{rt/n} \tilde{x}(s)\,ds. \end{aligned} \tag{79}$$

By (19),

$$\eta_{rn} = \xi_{rn} + \int_{(r-1)t/n}^{rt/n} \beta(s)\,dw(s);$$

and for each fixed n the random variables $\eta_{rn} - \xi_{rn}$ $(r = 1, \cdots, n)$ are independent and Gaussian with mean 0 and variance

$$v_{rn} = \int_{(r-1)t/n}^{rt/n} \beta(s)^2\,ds. \tag{80}$$

Using this fact we can write

$$\begin{aligned} p_j^{(n)}(t) &= P\{x(t) = a_j \mid y(s_{rn}), r = 0, 1, \cdots, n\} \\ &= P\{x(t) = a_j \mid \eta_{rn}, r = 1, \cdots, n\} \\ &= \frac{\displaystyle\sum_{i=1}^{K} p_i(0)p_{ij}(t) \cdot E\left\{\exp\left[-\sum_{r=1}^{n}(c_{rn} - \xi_{rn})^2(2v_{rn})^{-1}\right] \middle| x_0 = a_i, x_t = a_j\right\}_{c_{rn}=\eta_{rn}}}{\displaystyle\sum_{k=1}^{K}\sum_{i=1}^{K} p_i(0)p_{ik}(t) \cdot E\left\{\exp\left[-\sum_{r=1}^{n}(c_{rn} - \xi_{rn})^2(2v_{rn})^{-1}\right] \middle| x_0 = a_i, x_t = a_k\right\}_{c_{rn}=\eta_{rn}}}. \end{aligned} \tag{81}$$

In (81) the c_{rn} are arbitrary real numbers and the conditional expectation is regarded as a function of the c_{rn} evaluated at the (random) argument $c_{rn} = \eta_{rn}$ $(r = 1, \cdots, n)$. The last line of (81) follows by Bayes' rule and the fact that the conditional expectations thus evaluated are simply conditional densities of the η_{rn}. From now on a normalizing factor will be denoted by the generic symbol N. Then

$$\begin{aligned} p_j^{(n)}(t) &= N \sum_{i=1}^{K} p_i(0)p_{ij}(t) \\ &\cdot E\left\{\exp\left[\sum_{r=1}^{n} c_{rn}\,\xi_{rn}\,v_{rn}^{-1} - \frac{1}{2}\sum_{r=1}^{n} \xi_{rn}^2\,v_{rn}^{-1}\right] \middle| x_0 = a_i, x_t = a_j\right\}_{c_{rn}=\eta_{rn}}. \end{aligned} \tag{82}$$

By our assumptions on the $\tilde{x}(t)$ process, (82) can also be written

$$p_j^{(n)}(t) = N \sum_{i=1}^{K} p_i(0) p_{ij}(t) \tag{83}$$
$$\cdot E\left\{ \exp\left[\sum_{r=1}^{n} \eta_{rn} \tilde{\xi}_{rn} v_{rn}^{-1} - \frac{1}{2} \sum_{r=1}^{n} \tilde{\xi}_{rn}^{2} v_{rn}^{-1} \right] \middle| \tilde{x}_0 = a_i, \tilde{x}_t = a_j, \eta_{1n}, \cdots, \eta_{nn} \right\}.$$

Since almost every $\tilde{x}(t)$ sample function is a step function, the limit

$$\lim_{n\to\infty} \sum_{r=1}^{n} \tilde{\xi}_{rn}^{2} v_{rn}^{-1} = \int_0^t \beta(s)^{-2} \tilde{x}(s)^2 \, ds \tag{84}$$

exists with probability 1 and hence, by bounded convergence, in mean square. Further

$$\begin{aligned} \sum_{r=1}^{n} \eta_{rn} \tilde{\xi}_{rn} v_{rn}^{-1} &= \sum_{r=1}^{n} \left[\xi_{rn} \tilde{\xi}_{rn} + \tilde{\xi}_{rn} \int_{(r-1)t/n}^{rt/n} \beta(s) \, dw(s) \right] v_{rn}^{-1} \\ &\to \int_0^t \beta(s)^{-2} x(s) \tilde{x}(s) \, ds \\ &\qquad + \int_0^t \beta(s)^{-1} \tilde{x}(s) \, dw(s) \qquad (n \to \infty) \\ &= \int_0^t \beta(s)^{-2} \tilde{x}(s) \, dy(s), \end{aligned} \tag{85}$$

where the integral is again a limit in mean square ([6, Chap. IX, §2]). Let $\theta_n(t)$ be the random variable in the brackets in (83) and put

$$\theta(t) = \int_0^t \beta(s)^{-2} \tilde{x}(s) \, dy(s) - \frac{1}{2} \int_0^t \beta(s)^{-2} \tilde{x}(s)^2 \, ds. \tag{86}$$

From (84) and (85), l.i.m. $\theta_n(t) = \theta(t)$. Furthermore

$$\begin{aligned} \left| e^{\theta_n(t)} - e^{\theta(t)} \right| &\leqq \tfrac{1}{2} \left| \theta_n(t) - \theta(t) \right| [e^{\theta_n(t)} + e^{\theta(t)}] \\ &\leqq \left| \theta_n(t) - \theta(t) \right| e^{\lambda |w(t)| + \mu t} \end{aligned}$$

for some constants λ, $\mu > 0$; and since $E\{[e^{\lambda |w(t)|}]^2\} < \infty$,

$$\underset{n\to\infty}{\text{l.i.m.}} \, e^{\theta_n(t)} = e^{\theta(t)}. \tag{87}$$

Denote by $\mathfrak{F}_t^{\,n}$ (resp. $\mathfrak{F}_t$) the smallest Borel field[10] relative to which the random variables $\tilde{x}(0)$, $\tilde{x}(t)$, $\eta_{1n}, \cdots, \eta_{nn}$, (resp. $\tilde{x}(0)$, $\tilde{x}(t)$, $y(s)$, $0 \leqq s \leqq t$) are measurable. Now

$$\begin{aligned} &E\{e^{\theta_n(t)} \mid \mathfrak{F}_t^{\,n}\} - E\{e^{\theta(t)} \mid \mathfrak{F}_t\} \\ &\quad = E\{e^{\theta_n(t)} \mid \mathfrak{F}_t^{\,n}\} - E\{e^{\theta_n(t)} \mid \mathfrak{F}_t\} + E\{e^{\theta_n(t)} \mid \mathfrak{F}_t\} - E\{e^{\theta(t)} \mid \mathfrak{F}_t\}. \end{aligned} \tag{88}$$

[10] See footnote 3 in §2.

Since $\mathfrak{F}_t^n \subset \mathfrak{F}_t$ and since, by inspection of $\theta_n(t)$, the random variable $E\{e^{\theta_n(t)} \mid \mathfrak{F}_t\}$ is measurable relative to $\mathfrak{F}_t^n$, there follows ([6, Chap. I, Theorem (8.1])

$$(89) \qquad E\{e^{\theta_n(t)} \mid \mathfrak{F}_t^n\} - E\{e^{\theta_n(t)} \mid \mathfrak{F}_t\} = 0$$

with probability 1. Finally, (87) implies

$$(90) \qquad \underset{n\to\infty}{\text{l.i.m.}}\, E\{e^{\theta_n(t)} \mid \mathfrak{F}_t\} - E\{e^{\theta(t)} \mid \mathfrak{F}_t\} = 0.$$

Thus we have shown that

$$(91) \qquad \begin{aligned} \underset{n\to\infty}{\text{l.i.m.}}\, p_j^{(n)}(t) &= N \sum_{i=1}^{K} p_i(0) p_{ij}(t) \\ &\quad \cdot E\{e^{\theta(t)} \mid \tilde{x}(0) = a_i\,, \tilde{x}(t) = a_j\,; y(s), 0 \leqq s \leqq t\}. \end{aligned}$$

By exactly the same argument as in Appendix 2 the left side of (91) can be identified with $p_j(t)$.

2. Next we show that the process $\{x(t), p(t), t \geqq 0\}$ is Markov. For $0 \leqq \tau \leqq t$ write

$$(92) \quad \phi(\tau, t) = \exp\left[\int_\tau^t \beta(s)^{-2}\tilde{x}(s)\, dy(s) - \frac{1}{2}\int_\tau^t \beta(s)^{-2}\tilde{x}(s)^2\, ds\right],$$

and let $\mathcal{H}_\tau^{\,t}$ be the smallest Borel field relative to which the random variables $y(s)$, $\tau \leqq s \leqq t$, are measurable. Thus $\phi(\tau, t)$ is measurable relative to the Borel field generated by $\mathcal{H}_\tau^{\,t}$ and the $\tilde{x}(s)$'s for $\tau \leqq s \leqq t$. With this notation,

$$(93) \quad p_j(t) = N\sum_{i=1}^{K} p_i(0)p_{ij}(t)E\{\phi(0, t) \mid \tilde{x}_0 = a_i\,, \tilde{x}_t = a_j\,, \mathcal{H}_0^{\,t}\}.$$

Now

$$(94) \quad \begin{aligned} &p_j(t + h) \\ &\quad = N\sum_{k=1}^{K} p_k(0)p_{kj}(t + h)E\{\phi(0, t + h) \mid \tilde{x}_0 = a_k\,, \tilde{x}_{t+h} = a_j\,, \mathcal{H}_0^{\,t+h}\} \\ &\quad = N\sum_{k=1}^{K} p_k(0)p_{kj}(t + h)\sum_{i=1}^{K} P\{\tilde{x}_t = a_i \mid \tilde{x}_0 = a_k\,, \tilde{x}_{t+h} = a_j\} \\ &\qquad \cdot E\{\phi(0, t)\phi(t, t + h) \mid \tilde{x}_0 = a_k\,, \tilde{x}_t = a_i\,, \tilde{x}_{t+h} = a_j\,, \mathcal{H}_0^{\,t+h}\} \\ &\quad = N\sum_{i=1}^{K}\sum_{k=1}^{K} p_k(0)p_{ki}(t)p_{ij}(h)E\{\phi(0, t) \mid \tilde{x}_0 = a_k\,, \tilde{x}_t = a_i\,, \mathcal{H}_0^{\,t}\} \\ &\qquad \cdot E\{\phi(t, t + h) \mid \tilde{x}_t = a_i\,, \tilde{x}_{t+h} = a_j\,, \mathcal{H}_t^{\,t+h}\}, \end{aligned}$$

where we have used the fact that the $\tilde{x}(t)$ process is Markov and is independent of the $y(t)$ process. Comparing (93) and (94) we obtain

$$\begin{aligned}(95)\quad p_j(t+h) &= N\sum_{i=1}^{K} p_i(t)p_{ij}(h)\\ &\quad\cdot E\{\phi(t, t+h) \mid \tilde{x}_t = a_i, \tilde{x}_{t+h} = a_j, \mathcal{H}_t^{t+h}\}.\end{aligned}$$

Write $p(t) = [p_1(t), \cdots, p_K(t)]$ and consider the joint process $\{x(t), p(t), t \geqq 0\}$. Equations (92) and (95) show that $x(t+h), p(t+h)$ are fully determined by $x(t)$, $p(t)$ and the $w(s)$ increments for $t \leqq s \leqq t+h$. Reasoning as in Appendix 2 we conclude that the joint process $\{x(t), p(t), t \geqq 0\}$ is Markov.

3. We now evaluate functions m_j and b_{ij} defined by

$$(96)\quad m_j(t, x, p) = \lim_{h\to 0} E\left\{\frac{p_j(t+h) - p_j(t)}{h}\,\middle|\, x(t) = x, p(t) = p\right\}$$

and

$$\begin{aligned}(97)\quad b_{ij}(t, x, p) &= \lim_{h\to 0} E\\ &\cdot\left\{\frac{[p_i(t+h) - p_i(t)][p_j(t+h) - p_j(t)]}{h}\,\middle|\, x(t) = x, p(t) = p\right\}.\end{aligned}$$

To simplify computation note that the conditional expectation in (95) is readily evaluated, if it is modified by including the extra condition that no jump of $\tilde{x}(\cdot)$ occurs in the interval $(t, t+h)$; and the conditional probability that no jump occurs, given $\tilde{x}_t = \tilde{x}_{t+h} = a_i$, is

$$(98)\quad \frac{e^{-\nu_i h}}{p_{ii}(h)} = 1 + O(h^2) \qquad (h \to 0).$$

Also, since $\beta(s)^{-1}$ is assumed bounded, we have

$$(99)\quad \phi(t, t+h) \leqq f(h)e^{\lambda|\Delta y|},$$

where $\Delta y = y(t+h) - y(t)$, $\lambda > 0$ is constant, and $f(h)$ is bounded as $h \to 0$. Put

$$(100)\quad \theta_j = a_j \int_t^{t+h} \beta(s)^{-2}\, dy(s) - \frac{1}{2}a_j^2 \int_t^{t+h} \beta(s)^{-2}\, ds.$$

From (92) and (98)–(100),

$$(101)\quad E\{\varphi(t, t+h) \mid \tilde{x}_t = a_i, \tilde{x}_{t+h} = a_i, \mathcal{H}_t^{t+h}\} = e^{\theta_i} + h^2\omega_i(h),$$

and if $j \neq i$,

$$(102)\quad E\{\phi(t, t+h) \mid \tilde{x}_t = a_\imath, \tilde{x}_{t+h} = a_j, \mathcal{H}_t^{t+h}\} = 1 + h\omega_{ij}(h),$$

where (for a suitable $f(\cdot)$)

$$0 \leqq \omega_i(h), \quad \omega_{ij}(h) \leqq f(h)e^{\lambda|\Delta y|}. \tag{103}$$

Thus

$$\begin{aligned} p_j(t+h) = N[p_j(t)p_{jj}(h)\{e^{\theta_j} + h^2\omega_j(h)\} \\ + \sum_{\substack{i=1 \\ i \neq j}}^{K} p_i(t)p_{ij}(h)\{1 + h\omega_{ij}(h)\}]. \end{aligned} \tag{104}$$

A simple calculation from (100) yields

$$\lim_{h \to 0} h^{-1}E\{\theta_j \mid x(t) = x\} = a_j x\beta(t)^{-2} - \tfrac{1}{2}a_j^2\beta(t)^{-2}, \tag{105}$$

$$\lim_{h \to 0} h^{-1}E\{\theta_i\theta_j \mid x(t) = x\} = a_i a_j \beta(t)^{-2}. \tag{106}$$

We note that in (104), N^{-1} is the sum over j of the terms in brackets. Writing out (104) explicitly, subtracting $p_j(t)$ and then using (105), (106), we can compute the limits (96) and (97). The results are

$$m_j(t, x, p) = -\nu_j p_j + \sum_{\substack{i=1 \\ i \neq j}}^{K} \nu_{ij}p_i + \beta(t)^{-2}(x - \bar{x})(a_j - \bar{x})p_j \tag{107}$$

and

$$b_{ij}(t, x, p) = [\beta(t)^{-1}(a_i - \bar{x})p_i][\beta(t)^{-1}(a_j - \bar{x})p_j], \tag{108}$$

where

$$\bar{x} = \sum_{i=1}^{K} a_i p_i . \tag{109}$$

4. Define

$$\sigma_j(t, x, p) = \beta(t)^{-1}(a_j - \bar{x})p_j, \qquad j = 1, \cdots, K. \tag{110}$$

It will be shown finally that the $p(t)$ process can be represented as the solution of the stochastic differential system

$$\begin{aligned} dp_j(t) = m_j[t, x(t), p(t)]\, dt \\ + \sigma_j[t, x(t), p(t)]\, dw(t), \quad t \geqq 0, j = 1, \cdots, K, \end{aligned} \tag{111}$$

where $\{w(t),\ t \geqq 0\}$ is the Wiener process introduced in (19). We shall apply a representation theorem of Doob [6, Chap. VI, Theorem 3.3], generalized to allow for the fact that only the $p(t)$ component of the $\{x(t), p(t)\}$ process is continuous (that almost every $p(t)$ sample function is continuous follows from (104)).

The $p(t)$ process is obviously bounded; hence by (107), (110) the conditions usually imposed on m and σ [6, Chap. VI, §3] are satisfied. Now let $\mathcal{G}_t$ be the smallest Borel field with respect to which the random variables $x(s)$, $p(s)$, $0 \leqq s \leqq t$ are measurable. Then, since the $\{x, p\}$ process is Markov, the evaluations (107), (108) are unchanged if the conditional expectations in (96), (97) are defined relative to $\mathcal{G}_t$. Reasoning as in the proof[11] of [6, Chap. VI, Theorem 3.3], we conclude that each process

$$\left\{p_j(t) - p_j(0) - \int_0^t m_j[s, x(s), p(s)]\, ds, \mathcal{G}_t\, ; t \geqq 0\right\}, \qquad j = 1, \cdots, K, \tag{112}$$

is a martingale which satisfies the conditions of [6, Chap. IX, Theorem 5.3]. That is, if $\{\tilde{p}_j(t), \mathcal{G}_t\, ; t \geqq 0\}$ is the martingale defined by (112), if $\{\sigma_j\}^{-1} = \sigma_j^{-1}$ where $\sigma_j \neq 0$, and if $\{\sigma_j\}^{-1} = 0$ where[12] $\sigma_j = 0$, then the equation

$$\hat{w}_j(t) = \int_0^t \{\sigma_j[s, x(s), p(s)]\}^{-1}\, d\tilde{p}(s) \tag{113}$$

defines a Wiener process $\hat{w}_j(t)$, $t \geqq 0$.

It remains to identify each $\hat{w}_j(t)$ process with the $w(t)$ process of (19). Let

$$\begin{aligned}\tilde{y}(t) &= y(t) - \int_0^t x(s)\, ds \\ &= \int_0^t \beta(s)\, dw(s).\end{aligned} \tag{114}$$

Using (104)–(106) and (110) we find that

$$\begin{aligned}\lim_{h \to 0} h^{-1} E\{[\tilde{y}(t + h) - \tilde{y}(t)][p_j(t + h) - p_j(t)] \mid \mathcal{G}_t\} \\ = \beta(t)\sigma_j[t, x(t), p(t)].\end{aligned} \tag{115}$$

Reasoning as before we conclude that the process $\{\tilde{y}(t), \mathcal{G}_t\, ; t \geqq 0\}$ is a martingale. Since

$$w(t) = \int_0^t \beta(s)^{-1}\, d\tilde{y}(s),$$

it follows by (113) and (115) that

$$E\{[w(t) - w(s)][\hat{w}_j(t) - \hat{w}_j(s)] \mid \mathcal{G}_s\} = t - s, \qquad 0 < s < t. \tag{116}$$

[11] The Borel field $\mathcal{G}_t$ plays the role of the $\mathcal{F}_t$ of Doob's theorem.

[12] σ_j cannot vanish on a t-interval. Otherwise $\bar{x}(t)$ is constant on this interval, which can be shown to imply that the posterior variance of $x(t)$ vanishes, hence that all but one $p_k(t)$ vanishes, in contradiction to (86) and (91).

Hence for each $t > 0$, $\hat{w}_j(t) = w(t)$ with probability 1; by continuity this implies that the processes $\hat{w}_j(t)$, $w(t)$ are essentially identical. The integrated form of (21) now follows from (113) by inversion.

REFERENCES

[1] R. L. Stratonovič, *Conditional Markov processes*, Theor. Probability Appl., 5 (1960), pp. 156–178.

[2] G. E. Kolosov and R. L. Stratonovič, *A problem of synthesis of an optimal regulator by methods of dynamic programming*, Avtomat. i Telemeh., 24 (1963), pp. 1165–1173.

[3] W. M. Wonham, *Stochastic problems in optimal control*, RIAS TR 63-14, 1963.

[4] R. E. Kalman and W. M. Wonham, *Investigation of filter functions*, Final TR, Contract No. DA-36-034-ORD-3708Z, U.S. Army Ordnance Missile Command, Redstone Arsenal, Alabama, 1964.

[5] H. J. Kushner, *On the differential equations satisfied by conditional probability densities of Markov processes, with applications*, this Journal, 2 (1964), pp. 106–119.

[6] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[7] E. B. Dynkin, *Markovskie Processy*, Fizmatgiz, Moscow, 1963.

[8] K. Ito, *On Stochastic Differential Equations*, Mem. Amer. Math. Soc., 4, 1951.

[9]. I. I. Gihman, *On the theory of differential equations of stochastic processes, I.* Ukrain. Mat. Z., 2 (1950), pp. 37–63.

[10] J. F. Barrett, *Application of Kolmogorov's equations to randomly disturbed automatic control systems*, Proc. IFAC, Automatic and Remote Control, vol. 2, Butterworth, London, 1961, pp. 724–733.

[11] S. O. Rice, *Mathematical analysis of random noise*, Bell System Tech. J., 23, 24 (1944); reprinted in Noise and Stochastic Processes, N. Wax, ed., Dover, New York, 1954.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

THE ASYMPTOTES OF THE TIME LAG ROOT-LOCUS*

ALLAN M. KRALL†

In 1961 [2] and 1963 [3] formal proofs of various properties of the root-locus method appeared. Included in these properties was the existence of asymptotes which some of the zeros of the root-locus approach as the parameter becomes large.

Later Berman and Stanton [1] showed that not only does the root-locus approach these asymptotes, but in addition, the equations of the tangent lines approach the equations of the asymptotes. (The equation of the line $y = mx + b$ approaches the equation of the line $y = m_0 x + b_0$ when $m \to m_0$ and $b \to b_0$.)

Recently [4] the root-locus method has been extended to time lag systems. This paper shows that the time lag root-locus also has the property that the equations of the tangent lines approach the equations of the asymptotes.

Let $z = x + iy$, $g(z) = z^n + az^{n-1} + \cdots$, $h(z) = z^m + bz^{m-1} + \cdots$, where $n \geqq m$. Let κ, θ and $\tau > 0$ be real numbers. The time lag root-locus is the set of all zeros of $g(z) - \kappa e^i e^{-\tau z} h(z)$ for all real values of κ.

All points on the root-locus satisfy the equation

$$\begin{aligned} F(x, y) = \cos(\theta - \tau y) \operatorname{Im}(h(z)\overline{g(z)}) \\ + \sin(\theta - \tau y) \operatorname{Re}(h(z)\overline{g(z)}) = 0, \end{aligned} \tag{1}$$

where the bar over $g(z)$ indicates the complex conjugate. Let the positive root-locus be those zeros for which $\kappa \geqq 0$. The negative root-locus, similarly defined, is the positive root-locus with θ replaced by $\pi + \theta$. It has been shown that as x approaches $+\infty$, the positive root-locus approaches the lines,

$$y = \frac{1}{\tau}(\theta + 2k\pi), \qquad k = 0, 1, 2, \cdots. \tag{2}$$

As x approaches $-\infty$, the positive root-locus approaches the lines

$$y = \frac{1}{\tau}(\theta - [n - m]\pi + 2k\pi), \qquad k = 0, 1, 2, \cdots. \tag{3}$$

It is well known that if (x, y) is a point on the locus of $F(x, y) = 0$, the tangent to the locus at (x, y) is given by

$$F_y(x, y)Y + F_x(x, y)X = F_y(x, y)y + F_x(x, y)x, \tag{4}$$

* Received by the editors May 1, 1964, and in revised form October 14, 1964.

† 333 McAllister Hall, Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania.

where the point (X, Y) is on the tangent line. We now prove that this equation approaches $Y = (1/\tau)(\theta + 2k\pi)$ as x approaches $+\infty$, and approaches $Y = (1/\tau)(\theta - [n - m]\pi + 2k\pi)$ as x approaches $-\infty$. By $\bar{z}$ we mean $x - iy$. If $\theta(x)/\Psi(x)$ remains bounded as x approaches $\pm\infty$, then $\phi(x) = O(\Psi(x))$.

We see that

$$F_x = \cos(\theta - \tau y) \frac{\partial}{\partial x} \operatorname{Im}(h(z)\overline{g(z)}) + \sin(\theta - \tau y) \frac{\partial}{\partial x} \operatorname{Re}(h(z)\overline{g(z)}), \tag{5}$$

$$F_y = \cos(\theta - \tau y)\left[\frac{\partial}{\partial y} \operatorname{Im}(h(z)\overline{g(z)}) - \tau \operatorname{Re}(h(z)\overline{g(z)})\right] + \sin(\theta - \tau y)\left[\frac{\partial}{\partial y} \operatorname{Re}(h(z)g(z)) + \tau \operatorname{Im}(h(z)g(z))\right], \tag{6}$$

$$h(z)\overline{g(z)} = \bar{z}^n z^m + \bar{a}\bar{z}^{n-1} z^m + b\bar{z}^n z^{m-1} + O(|z|^{n+m-2}), \tag{7}$$

$$\operatorname{Re}(\bar{z}^n z^m) = x^{n+m} + O(x^{n+m-2}), \tag{8}$$

$$\operatorname{Im}(\bar{z}^n z^m) = (m - n)x^{n+m-1} y + O(x^{n+m-2}). \tag{9}$$

As x approaches $\pm\infty$, we see

$$F_x = \cos(\theta - \tau y)[O(x^{n+m-2})] + \sin(\theta - \tau y)[(m + n)x^{n+m-1} + O(x^{n+m-2})], \tag{10}$$

and

$$F_y = \cos(\theta - \tau y)[O(x^{n+m-1}) - \tau x^{n+m}] + \sin(\theta - \tau y)[O(x^{n+m-1})]. \tag{11}$$

Since as x approaches $\pm\infty$, y approaches $(1/\tau)(\theta + 2k\pi)$ or $(1/\tau)(\theta - [n - m]\pi + 2k\pi)$, $\cos(\theta - \tau y)$ approaches ± 1 and $\sin(\theta - \tau y)$ approaches 0. Thus

$$F_y = \pm\tau x^{n+m}(1 + O(x^{-1})). \tag{12}$$

Since F_y is ultimately bounded away from zero, we can divide (4) by F_y. Thus we find that if (x, y) is a point on the root-locus, the equation of the tangent line is

$$Y - y = M(X - x), \tag{13}$$

where $M = F_x/F_y$, given in (10) and (11). As x approaches $+\infty$, y approaches $(1/\tau)(\theta + 2k\pi)$, $\cos(\theta - \tau y)$ approaches 1, and $\sin(\theta - \tau y)$

approaches 0. Since both M and Mx are dominated by x^{n+m} in the denominator, both approach 0, and the equation of the tangent line approaches $Y = (1/\tau)(\theta + 2k\pi)$. Similarly as x approaches $-\infty$. the equation of the tangent line approaches $Y = (1/\tau)(\theta - [n - m]\pi + 2k\pi)$. We have proved

THEOREM. *The equations of the tangent lines of the time lag root-locus approach the equations of the asymptotes of the time lag root-locus as x approaches* $\pm \infty$.

REFERENCES

[1] GERALD BERMAN AND R. G. STANTON, *The asymptotes of the root-locus*, SIAM Rev., 5 (1963), pp. 209–218.

[2] ALLAN M. KRALL, *An extension and proof of the root-locus method*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 644–653.

[3] ———, *A closed expression for the root-locus method*, Ibid., 11 (1963), pp. 700–704.

[4] ———, *Stability criteria for feedback systems with a time lag*, this Journal, 2 (1964), pp. 160–170.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

TIME OPTIMAL CONTROL WITH AMPLITUDE AND RATE LIMITED CONTROLS*

W. W. SCHMAEDEKE† AND D. L. RUSSELL‡

Introduction. It has long been recognized that the maximum principle of Pontrjagin would have to be modified to allow for controls whose switching rates were finite, due either to inertial or other factors.

The first insight into the form of the resulting theory was provided by Birch and Jackson in their 1959 paper [2], although they were discussing quite a different problem.

The first discussion of the problem together with a set of necessary conditions characterizing the optimal controllers was provided by Chang. Several of the results in this paper were indicated by him in [1]. The aspect of the problem that is new in the treatment herein is the requirement that solutions of the augmented adjoint equations be differentiable on the whole interval $(0, T)$ instead of merely piecewise differentiable on so called "pang" intervals. It is this requirement which allows the "pang" intervals to be located. To be more specific, it is shown that the optimal control is either at extreme amplitude or extreme velocity. The subintervals of $(0, T)$ over which this behavior occurs can be determined if appropriate initial and final conditions are given.

Preliminaries. We shall consider a dynamical system whose state at any time t is described by an n-dimensional column vector $x(t)$. The law governing the motion of the state (and the law regarding the action of the controls on this motion) are expressed in the form of a vector differential equation,

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

where $A(t)$ is an $n \times n$ matrix, $B(t)$ is an $n \times m$ matrix, and $c(t)$ is an n-vector. The elements of A, B, and c are bounded continuous functions of time on an interval I under consideration. The components of the m-vector $u(t)$ correspond to the control functions whose values may be regulated in order to influence or control the motion of the state vector $x(t)$. The control

* Received by the editors February 21, 1964, and in revised form November 10, 1964. Prepared under Contract NASw-563 for the National Aeronautics and Space Administration.

† Minneapolis-Honeywell Regulator Company, Minneapolis, Minnesota. Now at the School of Mathematics, Institute of Technology, University of Minnesota, Minneapolis, Minnesota.

‡ Minneapolis-Honeywell Regulator Company, Minneapolis, Minnesota. Now at the Mathematics Research Center, United States Army, University of Wisconsin, Madison, Wisconsin.

problem is to select the real functions $u_j(t), j = 1, \cdots, m$, on an interval of time $0 \leqq t \leqq T$ such that the solution $x(t)$ of (1) moves from a prescribed initial point x_0 in R^n (n-dimensional real number space) to a prescribed moving target $\Delta(t)$ in minimum time T. The prescribed target set $\Delta(t)$ is assumed to be a nonempty compact subset of R^n for each fixed t in the given interval I. By considering the collection Σ of all nonempty compact subsets of R^n with the distance $d(C_1, C_2)$ between two such subsets C_1 and C_2 defined to be the infimum of all numbers d such that C_1 lies in the d-neighborhood of C_2 and C_2 lies in the d-neighborhood of C_1, Σ becomes a complete metric space (cf. [3]). It is herein assumed that the target set $\Delta(t)$ varies continuously with t in the sense of the preceding metric (called the Hausdorff metric). For example, if $\Delta(t)$ is a point for each t, then the target is a continuous curve; if $\Delta(t)$ is a constant compact set, then the problem is the familiar regulator problem where the target is fixed.

It is further supposed that there are no constraints on the state variables $x(t)$ other than the given initial point x_0 and the prescribed target set $\Delta(t)$, and that the controls $u(t)$ have components that are bounded in amplitude. Moreover, some (not necessarily all) of the components of the control vector are assumed to be differentiable and to satisfy bounds on these rates. Thus, the class of admissible controls is defined to be all m-vector functions $u(t)$, defined on various subintervals (of the form $[0, \tau]$) in I, whose first k components are absolutely continuous functions and whose remaining $m - k$ components are measurable functions. The following restrictions are imposed on the components of the vectors $u(t)$:

$$(2) \qquad \begin{aligned} a_{1i}(t) &\leqq u_i(t) \leqq a_{2i}(t), \quad i = 1, \cdots, m, \\ b_{1i}(t) &\leqq \dot{u}_i(t) \leqq b_{2i}(t), \quad i = 1, \cdots, k, \end{aligned}$$

where $k \leqq m$. The functions $a_{1i}(t)$, $a_{2i}(t)$, $b_{1i}(t)$, and $b_{2i}(t)$ are bounded continuous functions with the further assumption that $a_{1i}(t)$ and $a_{2i}(t)$ are absolutely continuous and satisfy

$$(3) \qquad \begin{aligned} b_{1i}(t) &< \dot{a}_{2i}(t) < b_{2i}(t), \\ b_{1i}(t) &< \dot{a}_{1i}(t) < b_{2i}(t), \end{aligned}$$

at all times at which the a's are differentiable.

It will be convenient to define new controls $v(t)$ with

$$v_1(t) = \dot{u}_1(t), \cdots, v_k(t) = \dot{u}_k(t), \quad v_{k+1}(t) = u_{k+1}(t), \cdots, v_m(t) = u_m(t),$$

and new state vectors $z(t)$ with

$$z_1(t) = x_1(t), \cdots, z_n(t) = x_n(t), \quad z_{n+1}(t) = u_1(t), \cdots, z_{n+k}(t) = u_k(t).$$

One then obtains the system

$$\dot{z} = F(t)z + G(t)v(t) + h(t), \tag{4}$$

where

$$F = \left[\begin{array}{ccc:ccc} a_{11} & \cdots & a_{1n} & b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} & b_{n1} & \cdots & b_{nk} \\ \hdashline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right] \equiv \begin{bmatrix} A & B_0 \\ 0 & 0 \end{bmatrix},$$

$$G = \left[\begin{array}{ccccc:ccc} 0 & \cdots & & & 0 & b_{1,k+1} & \cdots & b_{1m} \\ \vdots & & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & & & 0 & b_{n,k+1} & \cdots & b_{nm} \\ \hdashline 1 & 0 & \cdots & & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 \cdots & & 0 & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot & & & \\ \cdot & & & \cdot & 0 & \cdot & & \cdot \\ 0 & & & 0 & 1 & 0 & \cdots & 0 \end{array}\right] \equiv \begin{bmatrix} 0 & B_1 \\ I_k & 0 \end{bmatrix}, \tag{5}$$

$$h = \begin{bmatrix} c_1 \\ \vdots \\ c_n \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and where A is the original system's $n \times n$ coefficient matrix, B_0 is an $n \times k$ matrix whose k columns are the first k columns of the original control coefficient matrix B, B_1 is an $n \times (m - k)$ matrix whose columns are the remaining $(m - k)$ columns of B, and I_k is a $k \times k$ identity matrix. The zero matrices are blocks of zeros of the appropriate dimension to make F have dimension $(n + k) \times (n + k)$ and G have dimension $(n + k) \times m$; the number of zeros in h is k so that h is an $(n + k)$-vector.

The system (4) is now in a bounded phase setting, that is,

$$a_{1i} \leqq z_{n+i}(t) \leqq a_{2i}, \quad i = 1, \cdots, k, \tag{6}$$

(this is the bounded phase constraint); furthermore, the bounds on the amplitude of the new control vector $v(t)$ are given by

$$\begin{aligned} b_{1i} &\leqq v_i(t) \leqq b_{2i}, \quad i = 1, \cdots, k, \\ a_{1j} &\leqq v_j(t) \leqq a_{2j}, \quad j = k + 1, \cdots, m. \end{aligned} \tag{7}$$

Necessary conditions for optimal controls. For each time τ with $[0, \tau]$ contained in I, the set of all admissible controls on $[0, \tau]$ together with the set of their corresponding responses is considered. The set of attainability $K(\tau)$ is the set of all points $z(\tau)$ in R^{n+k} which are terminal points of these response trajectories, i.e., if $z(t)$ is the response to the admissible control function $v(t)$ defined on the interval $[0, \tau]$, then the point $z(\tau)$ is to be included in the set $K(\tau)$. By virtue of the conditions (2) imposed upon the controls $u(t)$ and the definition of the control function $v(t)$, it can be shown (see Appendix) that $K(\tau)$ is closed, bounded, and convex. Moreover, $K(\tau)$ is continuous in τ in the sense of the Hausdorff metric previously mentioned. This follows from the easily established fact that given $\epsilon > 0$, there exists a $\delta \neq 0$ such that $|z(\tau + \delta) - z(\tau)| < \epsilon$ for all responses $z(t)$ with $z(\tau)$ in $K(\tau)$.

Next, let $\Delta^*(t)$ be that subset of R^{n+k} obtained by the simple imbedding of $\Delta(t)$ in R^{n+k}, i.e., if z belongs to the set $\Delta^*(t)$, then the first n components of z constitute the components of a point x contained in $\Delta(t)$ and the remaining k components of z are unrestricted. The geometrical significance of this imbedding is simply that $\Delta^*(t)$ is a slab in R^{n+k} for each t. Thus, assuming the existence of an optimal control, if $z(0)$ is not in $\Delta^*(0)$ (i.e., if $x(0)$ is not in $\Delta(0)$), then as τ increases from 0, there is a first time T at which the convex set $K(T)$ comes into contact with the slab $\Delta^*(T)$. This, by virtue of the continuity, can be shown to imply that the optimal response $z(t)$ has its terminal point $z(T)$ in the boundary of $K(T)$.

Properties of controls $v(t)$ on an interval $[0, t_1]$ whose responses hit the boundary set of $K(t_1)$ will be examined. It will be convenient to treat (4) and the control $v(t)$ when investigating these properties because of the simple geometric nature of the problem. The necessary conditions for optimal controls, however, will be phrased in terms of (1) and the controls $u(t)$. To this end, the following definition is made.

Definition 1. The linear control process (4) subject to (6) and (7) is considered. An admissible control $v(t)$ on the interval $[0, T]$ is called an *extremal control* $\hat{v}(t)$ in case there exists a nontrivial solution $\psi(t)$ of the adjoint equations (a prime on a vector or matrix means the transpose of that vector or matrix)

$$\dot{\psi} = -F'\psi,$$

such that

$$\int_0^T \psi'(s) \begin{bmatrix} 0 & B_1(s) \\ I & 0 \end{bmatrix} \hat{v}(s)\, ds = \max_{v(s)} \int_0^T \psi'(s) \begin{bmatrix} 0 & B_1(s) \\ I & 0 \end{bmatrix} v(s)\, ds,$$

where the maximum is taken over all admissible controllers $v(s)$.

Lemma 1. *A control $\hat{v}(t)$ on $[0, T]$ is extremal if and only if the corresponding response $\hat{z}(t)$ has its terminal point $\hat{z}(T)$ in the boundary set of $K(T)$.*

Proof. Assume $\hat{v}(t)$ is such that $\hat{z}(T)$ lies in the boundary of $K(T)$. Then let π be a support plane to $K(T)$ at the point $\hat{z}(T)$, and let η be an outward normal to $K(T)$ at the point $\hat{z}(T)$. Then

$$\eta'[\hat{z}(T) - z(T)] \geqq 0 \tag{8}$$

for any point $z(T)$ belonging to $K(T)$. Now let $\Lambda(t, s)$ be a fundamental solution of the homogeneous equation corresponding to (4) with $\Lambda(s, s) = I$, the $(n + k) \times (n + k)$ identity, and consider the variation of parameters formula for a solution of (4):

$$\begin{aligned} \hat{z}(T) = \Lambda(T, 0)z_0 + \int_0^T \Lambda(T, 0)\Lambda^{-1}(s, 0) \begin{bmatrix} 0 & B_1(s) \\ I & 0 \end{bmatrix} \hat{v}(s)\, ds \\ + \int_0^T \Lambda(T, 0)\Lambda^{-1}(s, 0)h(s)\, ds. \end{aligned} \tag{9}$$

Hence

$$\eta'[\hat{z}(T) - z(T)] = \eta' \int_0^T \Lambda(T, s) \begin{bmatrix} 0 & B_1(s) \\ I & 0 \end{bmatrix} [\hat{v}(s) - v(s)]\, ds \geqq 0. \tag{10}$$

Let $\psi(s)$ be a particular solution of the adjoint equations by defining

$$\psi'(s) = \eta'\Lambda(T, s).$$

Then

$$\int_0^T \psi'(s) \begin{bmatrix} 0 & B_1(s) \\ I & 0 \end{bmatrix} [\hat{v}(s) - v(s)]\, ds \geqq 0, \tag{11}$$

i.e., $\hat{v}$ is an extremal control.

The other case is proven by beginning with (11) and proceeding backwards through the proof of the first case. This completes the proof.

According to Lemma 1, these extremal controls are the only candidates for the optimal controls since previous remarks have established that an optimal control has a response whose terminal point lies on the boundary of $K(T)$.

It will be convenient to decompose the adjoint vector $\psi'(s)$ as follows:

$$\psi'(s) = (\theta'(s), \phi'(s)), \tag{12}$$

where $\theta(s)$ is an n-vector and $\phi(s)$ is a k-vector. Then (11) becomes

$$\int_0^T [\phi'(s), \theta'(s)B_1(s)][\hat{v}(s) - v(s)]\, ds \geqq 0. \tag{13}$$

$v(s)$ is decomposed by defining

$$v(s) = \begin{bmatrix} \tilde{v}(s) \\ \underset{\sim}{u}(s) \end{bmatrix}, \tag{14}$$

where $\tilde{v}(s)$ is a k-vector whose components are

$$\tilde{v}_1(s) \equiv v_1(s), \cdots, \tilde{v}_k(s) \equiv v_k(s),$$

and where $\underset{\sim}{u}(s)$ is an $(m - k)$-vector whose components are

$$\underset{\sim}{u}_1(s) \equiv u_{k+1}(s), \cdots, \underset{\sim}{u}_{(m-k)}(s) \equiv u_m(s).$$

Also, for later use, $\tilde{u}(s)$ is defined as a k-vector whose components are

$$\tilde{u}_1(s) \equiv u_1(s), \cdots, \tilde{u}_k(s) \equiv u_k(s).$$

Then (13) may be written as

$$(15) \quad \int_0^T \phi'(s)[\hat{\tilde{v}}(s) - \tilde{v}(s)]\, ds + \int_0^T \theta'(s)B_1(s)[\hat{\underset{\sim}{u}}(s) - \underset{\sim}{u}(s)]\, ds \geqq 0.$$

It is now possible to refine Lemma 1 as follows:

LEMMA 2. *An extremal control $\hat{v}(t)$ must be such that its first k components (represented by $\hat{\tilde{v}}(t)$) satisfy*

$$(16) \qquad \int_0^T \phi_i'(s)[\hat{\tilde{v}}_i(s) - \tilde{v}_i(s)]\, ds \geqq 0,$$

for $i = 1, \cdots, k$, and for all $\tilde{v}_i(s)$ which are admissible ith *components of admissible controls $v(s)$; furthermore, the remaining $m - k$ components of $\hat{v}(t)$ (represented by $\hat{\underset{\sim}{u}}(t)$) satisfy*

$$(17a) \qquad \int_0^T [\theta'(s)B_1(s)]_i[\hat{\underset{\sim}{u}}_i(s) - u_i(s)]\, ds \geqq 0,$$

or equivalently

$$(17b) \qquad \int_0^T [\theta'(s)B_1(s)]_i[\hat{u}_{k+i}(s) - u_{k+i}(s)]\, ds \geqq 0,$$

for $i = 1, \cdots, m - k$, and for all admissible control components $u_{k+i}(s)$.

Proof. Let a particular choice of $v(s)$ be made as follows: $v_j(s) \equiv \hat{v}_j(s)$ for $j \neq i$ and let $v_i(s)$ be merely admissible. Then $\hat{v}(s) - v(s)$ has at most one nonzero component, namely, $\hat{v}_i(s) - v_i(s)$. With this choice for $v(s)$, the second integral in (15) vanishes and condition (16) of the lemma is established. Condition (17a) and its equivalent condition (17b) are proved in a similar manner.

Returning to (8), η is decomposed as follows:

$$(18) \qquad \eta' = (\lambda', \zeta'),$$

where λ is an n-vector and ζ is a k-vector. According to the definition of $\tilde{u}(t)$ in the remarks following (14), $z(T)$ may be written as

$$z(T) = \begin{bmatrix} x(T) \\ \tilde{u}(T) \end{bmatrix},$$

and (8) becomes

$$(19) \qquad \lambda'[\hat{x}(T) - x(T)] + \zeta'[\hat{\tilde{u}}(T) - \tilde{u}(T)] \geqq 0.$$

By utilizing the variation of parameters representation of a solution of (1) (with $E(t, s)$ as a fundamental solution matrix of the homogeneous equation, where $E(s, s)$ is the $n \times n$ identity), one obtains

$$(20) \quad x(t) = E(t, 0)x_0 + \int_0^t E(t, s)B(s)u(s)\,ds + \int_0^t E(t, s)c(s)\,ds.$$

Now, noting that

$$(21) \qquad \theta'(t) = \lambda'(t)E(t, s)$$

and substituting this and (20) into (19), there results

$$(22) \quad \int_0^T \theta'(s)B(s)[\hat{u}(s) - u(s)]\,ds + \zeta'[\hat{\tilde{u}}(T) - \tilde{u}(T)] \geqq 0,$$

where $u(t)$ is any admissible control vector.

Lemma 3 is established in a manner identical to that used for Lemma 2.

LEMMA 3. *An extremal control $v(s)$ must be such that its first k components (represented by $\dot{\hat{u}}_1(t), \cdots, \dot{\hat{u}}_k(t)$), when integrated, yield control components $\hat{u}_1(t), \cdots, \hat{u}_k(t)$ which satisfy*

$$(23) \quad \int_0^T [\theta'(s)B(s)]_i[\hat{u}_i(s) - u_i(s)]\,ds + \zeta_i[\hat{u}_i(T) - u_i(T)] \geqq 0,$$

for $i = 1, \cdots, k$, and all admissible components $u_i(t)$; furthermore, they must satisfy

$$(24) \qquad \int_0^T [\theta'(s)B(s)]_i[\hat{u}_i(s) - u_i(s)]\,ds \geqq 0,$$

for $i = k + 1, \cdots, m$, and all admissible components $u_i(t)$.

Remark 1. It is observed that the entire matrix B appears in the integrand, whereas in Lemma 2 the matrix was B_1, i.e., the last $m - k$ columns of B. Thus (24) is equivalent to (17) because

$$(25) \qquad [\theta'(s)B(s)]_{k+j} \equiv [\theta'(s)B_1(s)]_j,$$

for $j = 1, \cdots, m - k$.

Some qualitative properties of extremal controls will now be established. These are also necessary conditions for an optimal control. These conditions will be more conveniently phrased in terms of $u(t)$ rather than $v(t)$.

THEOREM 1. *Let $\hat{u}(t)$ be an extremal control for the system* (1). *If $\hat{u}_i(t)$ is at its upper limit during an interval of time, then $[\theta'(s)B(s)]_i \geqq 0$ on that interval. Also, if $\hat{u}_i(t)$ is at its lower limit during an interval of time, then $[\theta'(s)B(s)]_i \leqq 0$ on that interval.*

Proof. Let $\hat{u}_i(t) = a_{2i}(t)$ on an interval $[t_1, t_2]$ and suppose that $[\theta'(\tau)B(\tau)]_i < 0$ at some point τ in $[t_1, t_2]$. By continuity, there is an interval $[\tau_1, \tau_2]$, containing τ in its interior, on which $[\theta'(t)B(t)]_i < 0$. Consider (23) with $u_i(t)$ chosen so that

$$\hat{u}_i(t) - u_i(t) = \begin{cases} 0 & \text{outside of } [\tau_1, \tau_2], \\ \mu(t) > 0 & \text{in } [\tau_1, \tau_2]. \end{cases} \tag{26}$$

Then from (23) (noting $u_i(T) - \hat{u}_i(T) = 0$),

$$\int_{\tau_1}^{\tau_2} [\theta'(s)B(s)]_i \mu(s)\, ds \geqq 0. \tag{27}$$

But the integrand is negative on the entire interval and this is a contradiction. The remainder of the theorem is proved in a similar manner.

Now consider again the adjoint equations for (4),

$$\dot{\psi} = -F'\psi, \tag{28}$$

or, in terms of θ and ϕ,

$$\begin{bmatrix} \dot{\theta} \\ \dot{\phi} \end{bmatrix} = -\begin{bmatrix} A' & 0 \\ B_0' & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \phi \end{bmatrix}. \tag{29}$$

By performing the indicated matrix multiplication,

$$\dot{\theta} = -A'\theta, \tag{30}$$

$$\dot{\phi} = -B_0'\theta \tag{31}$$

are obtained. Notice that θ corresponds to the adjoint vector of the original system (1) whereas ϕ, corresponding to the augmented coordinates of the adjoint vector, is a trivial linear system in that no components of ϕ appear on the right sides.

Given a fundamental solution $E(t, t_0)$ to (30), (with $E(t_0, t_0) = n \times n$ identity), $\theta(t)$ may be represented by

$$\theta(t) = E(t, t_0)\theta_0 . \tag{32}$$

Then (31) yields

$$\phi(t) - \phi(t_0) = \int_{t_0}^{t} - B_0'(s)E(s, t_0)\theta_0\, ds. \tag{33}$$

In the following, a technique for utilizing $\phi(t)$ in the construction of optimal trajectories will be developed.

DEFINITION 2. Let $u_i(t)$ be an admissible component of the control vector for (1) or (4), and define an *interval of type B* as a maximal closed subinterval of the interval $[0, T]$ whereon $u_i(t)$ is extremal, i.e., assumes maximum or minimum amplitude throughout the whole subinterval.

DEFINITION 3. An *interval of type P_1* for $u_i(t)$ is defined to be a maximal closed interval in the interior of which $u_i(t)$ is not extreme valued, i.e., u_i assumes neither its maximum nor its minimum amplitude at any point in the interior of the interval. Note that if $P_1 \not\equiv [0, T]$, then $\hat{u}_i$ is extreme at one end (or both) of P_1.

DEFINITION 4. An *interval of type P_2* for $u_i(t)$ is defined to be a maximal subinterval of $[0, T]$ whereon $u_i(t)$ is not extreme and whereon $\dot{u}_i(t)$ is at one of its extremes, but not both.

Remark 2. The interval $[0, T]$ can be decomposed into nonoverlapping intervals of type B or type P_1 whose union is $[0, T]$.

THEOREM 2. *Let the system* (1) *be normal in the sense of LaSalle* (cf. [5]) *and consider an extremal control vector $\hat{u}(t)$ for* (1). *For each $i = 1, \cdots, m$, on an interval of type P_1 for $\hat{u}_i(t)$, $\dot{\hat{u}}_i(t)$ is either at its maximum value or its minimum value at every t at which $\dot{\hat{u}}_i(t)$ is defined.*

Proof. Let t belong to the interior of P_1 and assume that $\dot{\hat{u}}_1(t)$ is defined and is not extreme. Then, since $[\theta'(s)B(s)]_i$ is not zero on an interval by normality, we may assume further that t is such a point where $[\theta'(s)B(s)]_i$ is not zero. (This would eliminate a set of t in the interior of P_1 whose measure is zero). By continuity, $[\theta'(s)B(s)]_i$ is of one sign on an interval about the point t under consideration. For definiteness, assume $[\theta'(t)B(t)]_i < 0$. Then since $\hat{u}_i(t)$ is not extreme and since $\dot{\hat{u}}_i(t)$ is not extreme, an admissible control $u_i(s)$ is constructed as follows: Let M_1 be a line through the point $(t, \hat{u}_i(t))$ whose slope is $b_{1i}(t)$. (For simplicity, it is assumed that $b_{1i}(s) \leqq 0 \leqq b_{2i}(s)$, where both equalities do not hold simultaneously. Other cases would be treated similarly.) For a given $\delta > 0$, let $m_1(\delta)$ be a line through $(t, \hat{u}_i(t))$ whose slope is equal to the minimum of $b_{2i}(s)$ on the interval $[t, t + \delta]$ and let $m_2(\delta)$ be a line through the same point whose slope is the maximum of $b_{1i}(s)$ on the interval $[t, t + \delta]$. Now let $\delta_{01} > 0$ be chosen so small that the line $m_1(\delta_{01})$ lies entirely between* the curve $\hat{u}_i(s)$ and the line M_1 in the interval $[t, t + \delta_{01}]$, and let $\delta_{02} > 0$ be chosen so small that the line $m_2(\delta_{02})$ lies between the curve $\hat{u}_i(s)$ and the line M_2 in the interval $[t, t + \delta_{02}]$. Let δ_0 denote the smaller of δ_{01} and δ_{02} and, further, be small enough that $[\theta'(s)B(s)]_i < 0$ in $[t, t + \delta_0]$, and let m_1 and m_2 be the lines corresponding to δ_0; see Fig. 1.

According to the previous construction, there is a sgement S of the

* m_1 may coincide with M_1; similarly m_2 may coincide with M_2.

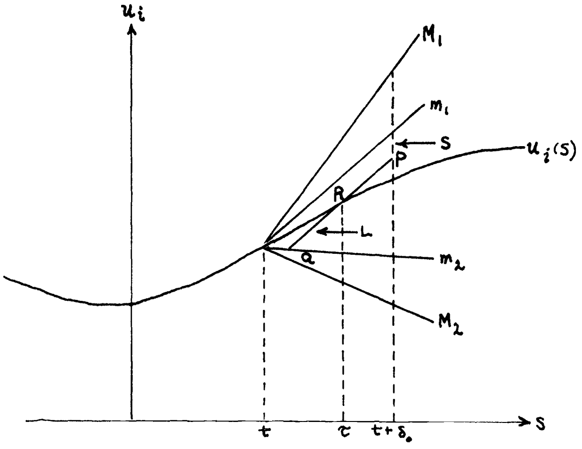


FIG. 1. *Construction of admissible varied control to prove $\hat{u}_i$ extremal*

ordinate at $t + \delta_0$ which is cut out by the line m_1 and the curve $\hat{u}_i(s)$. Since $\hat{u}_i(t)$ is not equal to its minimum value $a_{1i}(t)$, by continuity there is a point P on the segment S such that the line L through P parallel to m_1 will intersect the curve $\hat{u}_i(s)$ at the point R at a time τ in $(t, t + \delta_0)$ and such that it will intersect the line m_2 at a point Q at some time σ for which this intersection is above the height $a_{1i}(\sigma)$. Now define $u_i(s)$ to be equal to $\hat{u}_i(s)$ for $s \leqq t$, and $s \geqq \tau$. On the interval $[t, \tau]$ define $u_i(s)$ to be the segment of m_2 between the point $(t, \hat{u}_i(t))$ and Q, and to be the segment of L between Q and R. Thus $u_i(s)$ is an admissible control satisfying the amplitude and the rate bounds either by construction or because it is equal to $\hat{u}_i(s)$ which is assumed admissible.

Now (23) with the particular $u_i(s)$ just constructed is considered. It is seen that

$$(34) \qquad \int_t^\tau [\theta'(s)B(s)]_i[\hat{u}_i(s) - u_i(s)]\, ds \geqq 0.$$

But $[\theta'(s)B(s)]_i < 0$ in $[t, \tau]$ while $\hat{u}_i(s) - u_i(s) > 0$ in (t, τ). This is a contradiction and hence the velocity of $\hat{u}_i(t)$ must be extreme. The proof goes through in the same way if it is assumed that $[\theta'(t)B(t)]_i > 0$.

Remark 3. It follows from Theorem 2 that the interval $[0, T]$ is decomposable into subintervals of type B or type P_2, which are nonoverlapping and whose union is $[0, T]$. In other words, the optimal control is either at extreme amplitude or extreme velocity, whenever the velocity is defined.

THEOREM 3. *Let the system* (1) *be normal and consider an extremal control vector $\hat{u}(t)$ for* (1). *For each $i = 1, \cdots, m$, if there is an interval of type*

P_1 for $\hat{u}_i$ such that at least one of its endpoints, say t^*, is in the interior of $[0, T]$, then for all t in the P_1 interval for which $u_i(t)$ is defined:
(i) $\phi_i(t) > \phi_i(t^*)$ *implies that* $\dot{\hat{u}}_i(t)$ *is at its maximum value;*
(ii) $\phi_i(t) < \phi_i(t^*)$ *implies that* $\dot{\hat{u}}_i(t)$ *is at its minimum value.*

Proof. From Theorem 2, $\dot{\hat{u}}_i(t)$ is at one extreme or the other in P_1 intervals. By hypothesis, since one of the endpoints of the P_1 interval under consideration is a point t^* interior to $(0, T)$, it may be assumed without loss of generality that the point t^* is the right end of P_1. Furthermore, it is assumed without loss of generality that $\hat{u}_i(t^*)$ is a minimum. Now let t be a point in P_1 at which $\dot{\hat{u}}_i(t)$ is defined and suppose that $\phi_i(t)$ is greater than $\phi_i(t^*)$ but $\dot{\hat{u}}_i(t)$ is minimum (i.e., $b_{1i}(t)$). Fig. 2 supplies the details.

$u_i(s)$ is chosen so that on $[t, t + \delta]$ it has maximum slope and lies above $\hat{u}_i(s)$ while it is parallel to $\hat{u}_i(s)$ from $t + \delta$ to some point $\tau > t^*$ (choose δ so small that $\tau < T$). Then let $\hat{u}_i(s) = u_i(s)$ from τ to T. Now from the construction of $u_i(s)$ and from (23),

$$\int_t^{\tau+\delta} [\theta'(s)B(s)]_i[\hat{u}_i(s) - u_i(s)]\,ds + \int_{t+\delta}^{t^*} [\theta'(s)B(s)]_i[\hat{u}_i(s) - u_i(s)]\,ds + \int_{t^*}^{T} [\theta'(s)B(s)]_i[\hat{u}_i(s) - u_i(s)]\,ds \geqq 0. \tag{35}$$

On $[t + \delta, t^*]$ the function $\hat{u}_i(s) - u_i(s)$ has the constant value, say $-\epsilon$. Thus the middle integral is

$$-\epsilon \int_{t+\delta}^{t^*} [\theta'(s)B(s)]_i\,ds. \tag{36}$$

Note that as δ approaches zero, the integral in (36) (ignoring the mul-

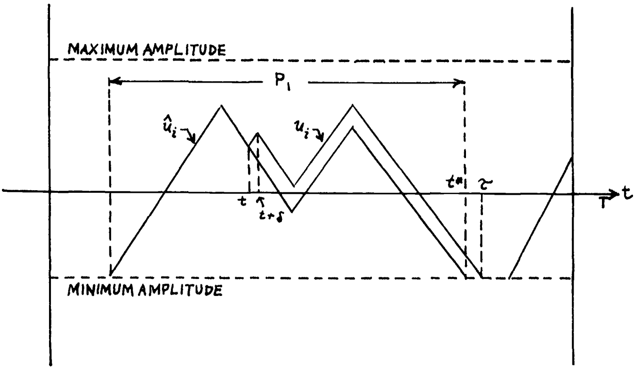


FIG. 2. *Construction of admissible control in proof of necessary conditions for P_1 intervals*

tiplicative factor ϵ) approaches

$$(37) \quad \int_t^{t^*} [\theta'(s)B(s)]_i \, ds = -\int_t^{t^*} \dot{\phi}_i(s) \, ds = \phi_i(t) - \phi_i(t^*) \triangleq r > 0.$$

Choose δ_0 so small that for all $\delta < \delta_0$,

$$(38) \qquad \frac{r}{2} < \int_{t+\delta}^{t^*} [\theta'(s)B(s)]_i \, ds < \frac{3r}{2}.$$

Now the first integral in (35) can be made small on the order of δ^2 as follows: since $| \hat{u}_i(s) - u_i(s) | \leqq K_1 | s - t |$ on $[t, t + \delta]$,

$$(39) \qquad \begin{aligned} &\left| \int_t^{t+\delta} [\theta'(s)B(s)]_i [\hat{u}_i(s) - u_i(s)] \, ds \right| \\ &\qquad \leqq K_1 \int_t^{t+\delta} \left| [\theta'(s)B(s)]_i \right| | s - t | \, ds. \end{aligned}$$

Letting $K_2 = \max_{[t,t+\delta_0]} | [\theta'(s)B(s)]_i |$ yields

$$(40) \qquad \left| \int_t^{t+\delta} [\theta'(s)B(s)]_i [\hat{u}_i(s) - u_i(s)] \, ds \right| \leqq K_1 K_2 \frac{\delta^2}{2}.$$

An easier analysis applies to the last integral, namely,

$$(41) \qquad \left| \int_{t^*}^{\tau} [\theta'(s)B(s)]_i [\hat{u}_i(s) - u_i(s)] \, ds \right| \leqq K_3 \epsilon (\tau - t^*).$$

As a result of (41), (40), (38) and (35),

$$(42) \qquad K_1 K_2 \frac{\delta^2}{2} - \frac{r}{2} \epsilon + K_3 \epsilon (\tau - t^*) \geqq 0.$$

It is next shown that δ is bounded above by a constant times ϵ. It is observed that $\dot{u}_i$ is greater than $\dot{\hat{u}}_i$ on $[t, t + \delta]$; hence their difference is not zero on $[t, t + \delta]$. Let the minimum of this difference be denoted by $c_1 > 0$; then

$$(43) \qquad \int_t^{t+\delta} [\dot{\hat{u}}_i(s) - \dot{u}_i(s)] \, ds \geqq c_1 \delta,$$

or

$$(44) \qquad c_1 \delta \leqq \hat{u}_i(t + \delta) - u_i(t + \delta) - [\hat{u}_i(t) - u_i(t)] = \epsilon.$$

Thus $c_1 \delta \leqq \epsilon$ or $\delta \leqq c_2 \epsilon$ where $c_2 > 0$. Applying this result to (42) yields

$$(45) \qquad K_1 K_2 c_2^2 \epsilon^2 - \frac{r}{2} \epsilon + K_3 \epsilon (\tau - t^*) \geqq 0,$$

or

$$(46) \qquad \epsilon\left(K_1K_2c_2^{\,2}\,\epsilon + K_3(\tau - t^*) - \frac{r}{2}\right) \geqq 0.$$

But (46) is a contradiction because for δ sufficiently small, ϵ and $\tau - t^*$ can be made arbitrarily small which means the quantity in parentheses is negative. Thus it has been shown that when $\phi_i(t) > \phi_i(t^*)$, then $\dot{\hat{u}}_i(t)$ is maximum (where it is defined). A similar proof will show that $\phi_i(t) < \phi_i(t^*)$ implies $\dot{\hat{u}}_i(t)$ is minimum.

Remark 4. Note that intervals of type P_2 coincide with intervals whereon the sign of $\phi_i(t) - \phi_i(t^*)$ is constant for appropriately chosen points t^*. It will be shown later that there are only a finite number of these points t^* for a given $\phi_i(t)$ and that it is possible, in those cases where $u_i(t)$ is given at the final time as well as the initial time, to construct the family of extremal controls.

THEOREM 4. *Let the system* (1) *be normal and consider an extremal control vector* $\hat{u}(t)$ *for* (1). *For each* $i = 1, \cdots, m$, *if the entire interval* $[0, T]$ (*where* T *is the minimal time of response*) *is of type* P_1 *for* $\hat{u}_i(t)$, *then there exists a constant* c_i *such that if* $\phi_i(t) - c_i > 0$ *then* $\dot{\hat{u}}_i(t)$ *is at its maximum value and if* $\phi_i(t) - c_i < 0$ *then* $\dot{\hat{u}}_i(t)$ *is at its minimum value* (*assuming that* $\hat{u}_i(t)$ *is defined at* t). *If there are at least two intervals of type* P_2 *contained in the interval of type* P_1, *then the value of the constant* c_i *is equal to* ϕ_i *evaluated at any of the interior endpoints of the type* P_2 *intervals.*

Proof.

Case I. If the whole interval $[0, T]$ is of type P_2 the theorem is trivially true as the constant c_i in this case may be chosen to be the minimum or the maximum of the function $\phi_i(t)$ on $[0, T]$ depending on whether $\dot{\hat{u}}_i$ is maximum or minimum.

Case II. If there are at least two intervals of type P_2 contained in P_1 then let t^* be an interior endpoint of a P_2 interval. Consider the case where $\dot{\hat{u}}_i(t)$ is minimum to the left of t^* and maximum to the right of t^* (the other case with the maximum and minimum reversed would be treated similarly). Let t' be any interior point of $[0, T]$ which is not the endpoint of an interval of type P_2. Assume that $\phi(t') - \phi(t^*) > 0$ but that $\dot{\hat{u}}_i(t')$ is at its minimum. Let it be supposed that $t' < t^*$. Then construct $u_i(t)$ on $[0, T]$ as follows: let $u_i(t) = \hat{u}_i(t)$ for $t \leqq t$; choose $\delta > 0$ so small that if $u_i(t)$ has maximum velocity on $[t', t + \delta]$, is parallel to $\hat{u}_i(t)$ on $[t' + \delta, t^*]$, and has minimum slope for a suitable time duration to the right of t^*, then the curve $u_i(t)$ will intersect the curve $\hat{u}_i(t)$ at some point τ to the right of t^*. (This choice is possible because the slope of $\hat{u}_i(t)$ is maximum to the right of t^*.) Finally, let $u_i(t) = \hat{u}_i(t)$ on $[\tau, T]$. From here on, one proceeds exactly as in the proof of Theorem 3 beginning with (35).

A method for computation of extremal trajectories for amplitude and rate limited controls. The foregoing theorems will now be given a more useful interpretation. Since the case where a_{1i}, a_{2i}, b_{1i}, b_{2i}, $i = 1, \cdots k$, are constant is of particular interest, as each condition on extremal trajectories is stated its specialization to this case will also be given. Each interval P_1 of type P_1 has a unique decomposition,

$$P_1 = \bar{P}_{2_1} \cup \bar{P}_{2_2} \cup \cdots \cup \bar{P}_{2_r}, \tag{47}$$

into intervals P_{2_ρ} of type P_2 where $\bar{P}_{2_\rho} \cap \bar{P}_{2_{\rho+1}}$ consists of precisely one point, $\rho = 1, 2, \cdots, r - 1$. The bar indicates topological closure. Let $\hat{u}_i$ be the ith component of the optimal control $\hat{u}$ corresponding to the function ϕ. In all that follows let P_1 be an interval of type P_1 for $\hat{u}_i$. Theorems 3 and 4 show that corresponding to the interval P_1 there is a constant c_i such that the subintervals P_{2_ρ} of P_1 coincide with those subintervals of P_1 whereon sgn $(\phi_i(t) - c_i)$ is constant. Let

$$\text{sgn } (P_{2_\rho}) \triangleq \text{sgn } (\phi_i(t) - c_i), \qquad t \in P_{2_\rho}. \tag{48}$$

Then (P_{2_ρ}) is set equal to the length of P_{2_ρ} and several cases are considered. Let τ_1, τ_2 be the endpoints of P_1.

Case I. τ_1, τ_2 both belong to $(0, T)$. Then it is clear that $\hat{u}_i(\tau_1)$ and $\hat{u}_i(\tau_2)$ are both extremal. In fact

$$(49) \quad \begin{aligned} &(\text{a}) \quad \hat{u}_i(\tau_1) = \begin{cases} a_{2i}(\tau_1) & \text{if sgn } (P_{2_1}) = -1, \\ a_{1i}(\tau_1) & \text{if sgn } (P_{2_1}) = +1, \end{cases} \\ &(\text{b}) \quad \hat{u}_i(\tau_2) = \begin{cases} a_{2i}(\tau_2) & \text{if sgn } (P_{2_r}) = +1, \\ a_{1i}(\tau_2) & \text{if sgn } (P_{2_r}) = -1. \end{cases} \end{aligned}$$

Case II. Either τ_1 or τ_2, but not both, belong to $(0, T)$. Then if $\tau_1 \in (0, T)$, 49(a) holds; if $\tau_2 \in (0, T)$, 49(b) holds. In each case the value of $\hat{u}_i$ at the other endpoint, i.e., either $\hat{u}_i(0)$ or $\hat{u}_i(T)$, must be specified in some other manner.

Case III. $\tau_1 = 0$, $\tau_2 = T$. Then both $\hat{u}_i(0)$ and $\hat{u}_i(T)$ must be specified; neither 49(a) nor 49(b) holds.

It is possible to consider problems wherein neither $\hat{u}_i(0)$ nor $\hat{u}_i(T)$ are specified. In this case the procedure to be described below is not immediately applicable. This situation arises in the so-called interception problem. A remark on this will be made at the end of this paper.

The values which $\hat{u}_i(t)$ assumes at τ_1 and τ_2 lead to the following conditions merely by applying the fundamental theorem of calculus for absolutely continuous functions.

Condition 1.

$$\sum_{\rho \ni \text{sgn}(P_{2\rho}) = +1} \int_{P_{2\rho}} b_{2i}(t)\, dt + \sum_{\rho \ni \text{sgn}(P_{2\rho}) = -1} \int_{P_{2\rho}} b_{1i}(t)\, dt = \hat{u}_i(\tau_2) - \hat{u}_i(\tau_1).$$

If the bounds on the control velocity are constant then:

Condition 1a.

$$[\sum_{\rho \ni \operatorname{sgn}(P_{2_\rho})=+1} l(P_{2_\rho})]\, b_{2i} + [\sum_{\rho \ni \operatorname{sgn}(P_{2_\rho})=-1} l(P_{2_\rho})] b_{1i} = \hat{u}_i(\tau_2) - \hat{u}_i(\tau_1).$$

The requirement that $u_i(t)$ shall not achieve an extreme value in the interior of P_1 leads to:

Condition 2. For each σ such that $1 \leqq \sigma < r$ (this set could be void),

$$\sum_{\substack{\rho \ni \operatorname{sgn}(P_{2_\rho})=+1 \\ \rho \leqq \sigma}} \int_{P_{2_\rho}} b_{2i}(t)\, dt$$
$$+ \sum_{\substack{\rho \ni \operatorname{sgn}(P_{2_\rho})=-1 \\ \rho \leqq \sigma}} \int_{P_{2_\rho}} b_{1i}(t)\, dt \begin{cases} < a_{2i}(\tau_{2\sigma}) - \hat{u}_i(\tau_1), \\ > a_{1i}(\tau_{2\sigma}) - \hat{u}_i(\tau_1), \end{cases}$$

where $\tau_{2\sigma}$ is the right endpoint of the interval P_2. The inequalities (3) enable this testing procedure to be restricted to points $\tau_{2\sigma}$. Again if the bounds on the control velocity are constant:

Condition 2a.

$$[\sum_{\substack{\rho \ni \operatorname{sgn}(P_{2_\rho})=+1 \\ \rho \leqq \sigma}} l(P_{2_\rho})] b_{2i} + [\sum_{\substack{\rho \ni \operatorname{sgn}(P_{2_\rho})=-1 \\ \rho \leqq \sigma}} l(P_{2_\rho})] b_{1i} \begin{cases} < a_{2i} - \hat{u}_i(\tau_1), \\ > a_{1i} - \hat{u}_i(\tau_1). \end{cases}$$

DEFINITION 5. Subintervals of [0, T] on which any control $u_i(t)$ may be defined so that Conditions 1 and 2 above are satisfied with $\hat{u}_i(t)$ replaced by $u_i(t)$ (i.e., $\dot{u}_i(t) = b_{2i}(t)$ if sgn $P_{2_\rho} = +1$, and $\dot{u}_i(t) = b_{1i}(t)$ if sgn $P_{2_\rho} = -1$) are called *intervals of type* P_3.

Thus every interval of type P_1 is also of type P_3 by Theorems 3, 4. The converse need not hold since there may be no extension of $u_i(\tau)$ from the given interval of type P_3 into the entire interval [0, T] as an extremal controller.

DEFINITION 6. A decomposition of $I = [0, T]$ into subintervals of types B and P_3 is called *acceptable* if the resulting control $u_i(t)$ is continuous and satisfies the preceding theorems on extremal controllers. The intervals of type P_3 then become intervals of type P_1 for $u_i(t)$.

The following theorem is of primary importance in establishing a procedure for computing extremal controllers.

THEOREM 5. *Assume that* $(\theta' B)_i(t)$ *has at most finitely many zeroes on* [0, T] *and the functions* $a_{1i}(t)$ *and* $a_{2i}(t)$ *are constants. Then there are at most finitely many possible intervals of type* P_3, *provided* $u_i(0)$ *and* $u_i(T)$ *are specified in advance.*

Proof. For a linear differential system the interval [0, T] may be divided

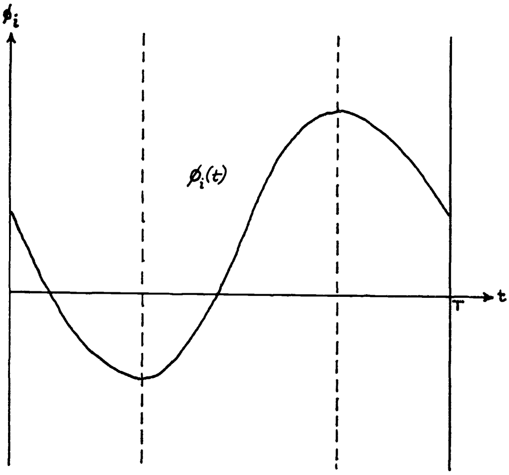


Fig. 3. *The function $\phi_i(t)$, indicating intervals of monotonicity*

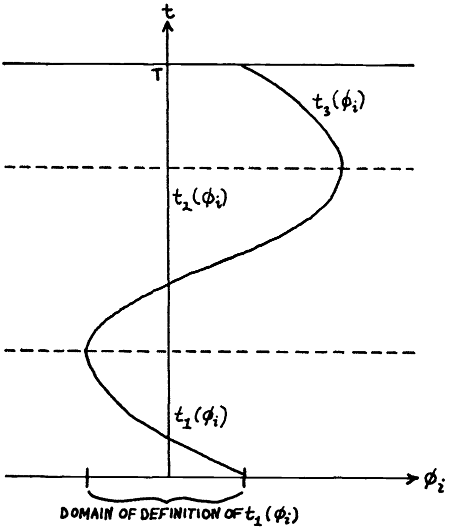


Fig. 4. *The inverse functions $t_\sigma(\phi_i)$ of ϕ_i*

into finitely many subintervals in which $\phi_i(t)$ is monotone. To prove this, note that the negation implies that $(\theta' B)_i(t)$ has infinitely many zeroes in $[0, T]$, contrary to assumption. Thus the inverse function $t(\phi_i)$ of the function $\phi_i(t)$ consists of finitely many functions $t_1(\phi_1), \cdots, t_s(\phi_i)$, each a monotone function defined on some subinterval of

$$\left[\min_{t \in [0,T]} \phi_i(t), \max_{t \in [0,T]} \phi_i(t) \right],$$

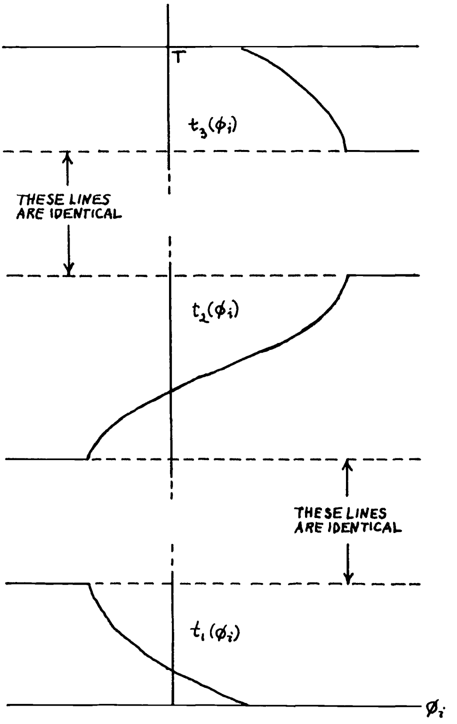


FIG. 5. *Final form of the functions* $t_\sigma(\phi_i)$

and $\sigma_1 < \sigma_2$ implies that $t_{\sigma_1}(\phi_i) < t_{\sigma_2}(\phi_i)$. Without loss of generality, it may be assumed that $t_\sigma(\phi_i)$ is decreasing for odd σ and increasing for even σ (Figs. 3 and 4). The other case is handled similarly. Let $t_0(\phi_i) \equiv 0$, $t_{s+1}(\phi_i) \equiv T$. The domain of definition of each $t_\sigma(\phi_i)$ is extended to all of

$$[\min_{t\in[0,T]} \phi_i(t), \max_{t\in[0,T]} \phi_i(t)]$$

by setting $t_\sigma(\phi_i)$ equal to $t_\sigma(\phi_i^*)$, where ϕ_i^* is the closest point to ϕ_i where $t_\sigma(\phi_i^*)$ is already defined (Fig. 5). For each of the finitely many pairs of indices σ_1, σ_2, $0 \leqq \sigma_1 < \sigma_2 \leqq s + 1$, $\overset{i}{g}_{\sigma_1\sigma_2}(\phi_i)$ is defined by

$$\overset{i}{g}_{\sigma_1\sigma_2}(\phi_i) = \sum_{\sigma=\sigma_1+1}^{\sigma_2} \int_{t_{\sigma-1}(\phi_i)}^{t_\sigma} \beta_{\sigma i}(t)\, dt, \tag{50}$$

where

$$\beta_{\sigma i}(t) = \begin{cases} b_{2i}(t) \text{ if } \sigma \text{ is odd,} \\ b_{1i}(t) \text{ if } \sigma \text{ is even.} \end{cases} \tag{51}$$

Then it is easy to see that each $\overset{i}{g}_{\sigma_1\sigma_2}(\phi_i)$ is a monotone decreasing function of ϕ_i. (If $t_\sigma(\phi_i)$ were increasing for σ odd, decreasing for σ even, then $g_{\sigma_1\sigma_2}(\phi_i)$ would still be a monotone decreasing function.)

It is clear by comparison of (50) with Condition 1 that an interval of type P_3 can occur only when there exist a pair σ_1, σ_2 and a value ϕ_i such that

$$g^i_{\sigma_1\sigma_2}(\phi_i) = u_i(t_{\sigma_2}(\phi_i)) - u_i(t_{\sigma_1}(\phi_i)), \tag{52}$$

where $u_i(t)$ is the control which must be defined on the interval according to the definition of an interval of type P_3. A number of cases are now considered.

If $\sigma_1 = 0$, $\sigma_2 = s + 1$, then we required in the hypotheses of this theorem that $u_i(0)$ and $u_i(T)$ be fixed. Thus $u_i(t_{\sigma_2}(\phi_i)) - u_i(t_{\sigma_1}(\phi_i))$ is a constant known beforehand.

If $\sigma_1 = 0$, σ_2 arbitrary and greater than 0, then $u_i(t_{\sigma_1}(\phi_i))$ is fixed at a constant value known beforehand while $u_i(t_{\sigma_2}(\phi_i)) = a_{2i}(t_{\sigma_2}(\phi_i))$ or $a_{1i}(t_{\sigma_2}(\phi_i))$.

If $\sigma_2 = s + 1$ while σ_1 is arbitrary and less than $s + 1$, then $u_i(t_{\sigma_2}(\phi_i))$ is fixed at a constant value known beforehand while $u_i(t_{\sigma_1}(\phi_i)) = a_{2i}(t_{\sigma_1}(\phi_i))$ or $a_{1i}(t_{\sigma_2}(\phi_i))$.

If $0 < \sigma_1 < \sigma_2 < s + 1$, then $u_i(t_{\sigma_2}(\phi_i)) - u_i(t_{\sigma_1}(\phi_i))$ is one of the four functions $a_{\delta i}(t_{\sigma_2}(\phi_i)) - a_{\gamma i}(t_{\sigma_1}(\phi_i))$, $\delta = 1, 2$; $\gamma = 1, 2$. Thus, since it was assumed that the functions $a_{1i}(t)$, $a_{2i}(t)$ were constants, it has been shown that there are at most finitely many values which $u_i(t_{\sigma_2}(\phi_i)) - u_i(t_{\sigma_1}(\phi_i))$ may assume for each σ_1, σ_2. Since there are finitely many functions $g^i_{\sigma_1\sigma_2}(\phi_i)$ and each of them is monotone, there are but finitely many instances wherein (52) may hold. This completes the proof of the theorem.

Remark 5. In the case where $a_{2i}(t)$ and $a_{1i}(t)$ are not constant but vary with time, the finitely many values, which we have shown in the proof of the theorem may be equal to $u_i(t_{\sigma_2}(\phi_i)) - u_i(t_{\sigma_1}(\phi_i))$, must be replaced by the finitely many functions $u_i(t_{\sigma_2}(\phi_i)) - u(t_{\sigma_1}(\phi_i))$ themselves. Then the conclusion of the theorem remains valid if for each σ_1, σ_2 the function

$$g_{\sigma_1\sigma_2}(\phi_i) = u_i(t_{\sigma_2}(\phi_i)) - u_i(t_{\sigma_1}(\phi_i))$$

has but finitely many zeroes on its domain of definition. It is difficult to give a reasonably general sufficient condition under which this holds; so the restriction to the case where $a_{2i}(t)$ and $a_{1i}(t)$ are constant was made. Clearly the likelihood is very small that any of these functions would have infinitely many zeroes in any given application. Thus it is fairly safe to assume that there are but finitely many P_3 intervals even if $a_{1i}(t)$ and $a_{2i}(t)$ are time-varying but it should be kept in mind that this has not been established and it may be possible to construct pathological functions $a_{2i}(t)$, $a_{1i}(t)$ such that this would not be true.

Remark 6. Note that the theorem also shows a method for finding the intervals of type P_3 since the functions $g_{\sigma_1\sigma_2}(\phi_i)$ and the constants (or functions $u_i(t_{\sigma_2}(\phi_i)) - u_i(t_{\sigma_1}(\phi_i))$ are readily determined. An acceptable decomposition of $[0, T]$ is into intervals of types B and P_3. Thus after having found all possible intervals of type P_3 (and the previous theorem assures us that in many cases this can be done), it remains only to find all acceptable decompositions of $[0, T]$, and hence all possible controls $u_i(t)$ which satisfy the first four theorems. There being only finitely many of these, the control $\hat{u}_i(t)$ which satisfies (11) can easily be found. If no values are given beforehand for $u_i(0)$ and for $u_i(T)$, then these values could be varied and the above results applied to each choice of those values to determine the best (in the sense of (11)) set of values for $u_i(0)$ and for $u_i(T)$.

A short example illustrating the use of the above results is now given.

An example to illustrate the construction of an extremal control. Let the time interval be $[0, 2]$ and

$$(\theta' B)_i(t) = -\frac{5\pi}{2} \cos\left(\frac{5\pi}{2}t\right).$$

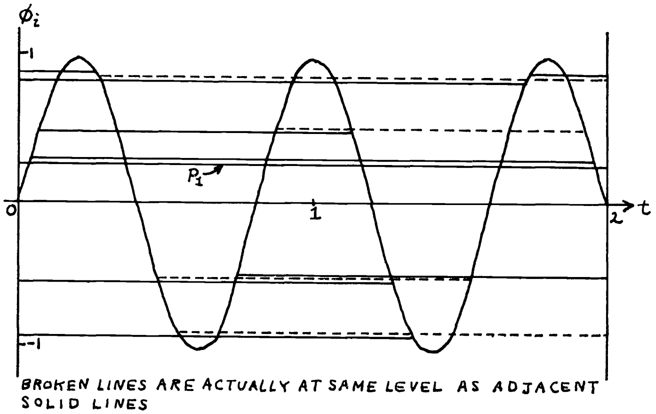
Then

$$\phi_i(t) = \sin\left(\frac{5\pi}{2}t\right).$$

Suppose that $a_{2i} = 1$, $a_{1i} = -1$; require $\hat{u}_i(0) = 0$, $\hat{u}_i(2) = 0$. The extremal control $\hat{u}_i(t)$ on $[0, 2]$ will be constructed. The method used will be graphic and will be special to the constants a_{1i}, a_{1i}, b_{2i}, b_{1i} in this problem. Its relationship to the immediately preceding discussion should be clear, as well as generalizations to different constant bounds. An interval of type P_3 occurs whenever it is possible to draw a level line L through the graph of $\sin(5\pi t/2)$ so that the endpoints of L lie on the graph of $\sin(5\pi t/2)$ or else meet the lines $t = 0$ or $t = 2$ and satisfy the following requirements. (Compare with Conditions 1 and 2 above.)

1. If the endpoints of L are in $(0, 2)$ then the sum S of the lengths of those segments of L lying below $\sin(5\pi t/2)$ minus the sum of the lengths of those segments of L lying above $\sin(5\pi t/2)$ must be 2, -2, or 0. If L' is any segment of L such that the left endpoints of L and L' coincide, then (a) if the first segment of L' lies below $\sin(5\pi t/2)$, the sum S' of the lengths of those segments of L' lying below $\sin(5\pi t/2)$ minus the sum of the lengths of those segments of L' lying above $\sin(5\pi t/2)$ must be less than 2 and greater than 0; (b) if the first segment of L' lies above $\sin(5\pi t/2)$ then the corresponding quantity must be greater than -2 and less than 0.

2. If $t = 0$ is an endpoint of L and the right hand endpoint of L belongs



BROKEN LINES ARE ACTUALLY AT SAME LEVEL AS ADJACENT SOLID LINES

FIG. 6. *All possible P_3 intervals for the function $\phi(t) = \sin(5\pi t/2)$ on the interval* [0, 2]

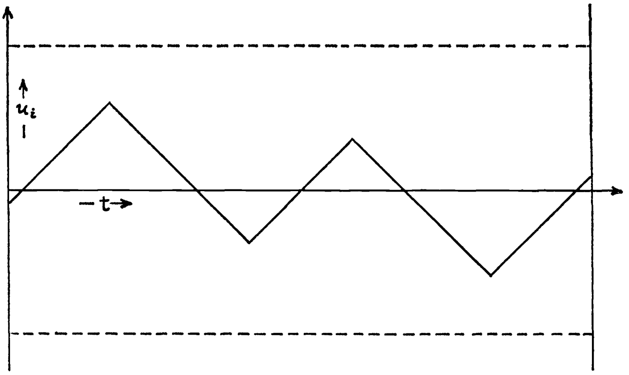


FIG. 7. *Extremal control constructed using results shown on Fig. 6*

to (0, 2) then $S = 1$ or -1 and $-1 < S' < 1$ for any L'. A similar situation occurs if the left hand endpoint of L lies in (0, 2) and the right hand endpoint is at $t = 2$, but here L, L' are taken to have a common right hand endpoint.

3. If L stretches from $t = 0$ to $t = 2$ then $S = 0$ and $-1 < S' < 1$ for any L' having an endpoint in common with L.

The graph in Fig. 6 shows all possible intervals of type P_3 indicated by level lines through the graph. The only acceptable sequence of intervals consists of the single interval P_1 of type P_1 which is indicated in the figure. This is clear by inspection, using the results of the first four theorems. Fig. 7 shows the resulting extremal control $u_i(t)$.

Appendix. The set of attainability $K(\tau)$ for (4) is shown to be closed and bounded as follows: the variation of parameters formula for a solution of (4) (with fundamental solution matrix $\Lambda(t, s)$) yields the representation

$$z(\tau) = \Lambda(\tau, 0)z_0 + \int_0^\tau \Lambda(\tau, s)[G(s)v(s) + h(s)]\, ds.$$

The conditions (2) on the components of $u(t)$ and the definition of $v(t)$ together with the properties of Λ and the boundedness of the components of $h(t)$ clearly imply the boundedness of $z(\tau)$ and hence the boundedness of $K(\tau)$.

To show $K(\tau)$ is closed, let $\{z^{(p)}(\tau)\}$ be any sequence of points in $K(\tau)$ such that $\lim_{p\to\infty} z^{(p)}(\tau) = Q$. Then we must show that Q belongs to $K(\tau)$. To this end, let $\{v^{(p)}(t)\}$ be a sequence of admissible controls whose responses at time τ are $\{z^{(p)}(\tau)\}$. Denote the interval $[0, \tau]$ by I_τ, and consider the functions $a_{1i}(t)$ and $a_{2i}(t)$, for $i = k + 1, \cdots, m$, as elements in $L_2(I_\tau)$. If $\mu_i(t)$ is defined as

$$\mu_i(t) = \max\{| a_{1i}(t) |, | a_{2i}(t) |\},$$

and the nonnegative number M_i is defined as

$$M_i^2 = \int_{I_\tau} | \mu_i(t) |^2\, dt,$$

then each function $v_i^{(p)}(t)$ (for each fixed i and all p) is in the sphere S_{M_i} contained in $L_2(I_\tau)$, and hence there is a subsequence of $\{v_i^{(p)}(t)\}$ (still labeled by $\{v_i^{(p)}(t)\}$ for convenience of notation) such that

$$\lim_{p\to\infty} v_i^{(p)}(t) \overset{\text{wk.}}{=} v_i(t)$$

for each $i = k + 1, \cdots, m$. These functions $v_i(t)$ belong to $L_2(I_\tau)$ and furthermore satisfy the conditions (for $t \in I_\tau$):

$$a_{1i}(t) \leqq v_i(t) \leqq a_{2i}(t), \qquad i = k + 1, \cdots, m.$$

This follows from the fact that the sequence $v_i^{(p)}(t) - a_{1i}(t)$ converges weakly to the limit $v_i(t) - a_{1i}(t)$. If the latter function is assumed to be negative on a set R of positive measure, then by considering the limit,

$$\int_{I_\tau} [v_i^{(p)}(t) - a_{1i}(t)]\delta_R(t)\, dt \to \int_{I_\tau} [v_i(t) - a_{1i}(t)]\delta_R(t)\, dt$$

(where $\delta_R(t)$ is the characteristic function of the set R), it is observed that the sequence of numbers on the left is nonnegative while the number on the right is strictly negative. This is a contradiction, and thus R must have measure zero. By changing $v_i(t)$ on this set of measure zero to conform to the inequality $v_i(t) - a_{1i}(t) \geqq 0$, the fact that $v_i(t)$ is still the weak limit of $v_i^{(p)}(t)$ is not altered, and the inequality then holds for all $t \in I_\tau$.

This same technique can be used to establish the fact that $a_{2i}(t) - v_i(t) \geqq 0$ for all $t \in I_\tau$ and each $i = k + 1, \cdots, m$; hence, the functions $v_i(t)$ are "admissible components".

We shall consider now the sequence $\{v^{(p)}(t)\}$ to consist of the intersection of all of the aforementioned subsequences with the appropriate change in notation. By applying the previous technique to the component sequences $\{v_i^{(p)}(t)\}$ for $i = 1, \cdots, k$, we obtain

$$v_i^{(p)}(t) \overset{\text{wk.}}{\longrightarrow} v_i(t),$$

and moreover

$$b_{1i}(t) \leqq v_i(t) \leqq b_{2i}(t).$$

Then, to each sequence $\{v_i^{(p)}(t)\}$ there corresponds a sequence $\{u_i^{(p)}(t)\}$ which consists of functions which are absolutely continuous on I_τ. In particular, since the derivatives of the $u_i^{(p)}(t)$ (the $v_i^{(p)}(t)$) are bounded, then the functions $u_i^{(p)}(t)$ are of uniform bounded total variation on I_τ and by Helly's theorem (cf. [4]) there exist functions $r_i(t)$ of bounded variation on I_τ for each $i = 1, \cdots, k$ such that $u_i^{(p)}(t) \rightarrow r_i(t)$ for all t in I_τ. Furthermore, since the functions $u_i^{(p)}(t)$ are absolutely continuous on I_τ we have the relation

$$\int_{I_\tau} v_i^{(p)}(s)\delta_t(s)\, ds = u_i^{(p)}(t) - u_i^{(p)}(0).$$

But, by weak convergence, the relation

$$\int_{I_\tau} v_i^{(p)}(s)\delta_t(s)\, ds \rightarrow \int_0^t v_i(s)\, ds$$

holds for all t in I_τ. These last two relations then imply $\dot{r}_i(t) = v_i(t)$ almost everywhere in I_τ.

If the functions $r_i(t)$ can be shown to be absolutely continuous and satisfy

$$a_{1i}(t) \leqq r_i(t) \leqq a_{2i}(t),$$

then the functions $v_i(t)$ will have been shown to be "admissible" components and thus the control $v(t)$ will have been shown to be an admissible control. Now, the fact that $u_i^{(p)}(t) \rightarrow r_i(t)$ pointwise on I_τ implies that $r_i(t)$ does indeed satisfy the pointwise bounds given by $a_{1i}(t)$ and $a_{2i}(t)$. The fact that $r_i(t)$ is absolutely continuous for each $i = 1, \cdots, k$ follows from the inequality

$$| u_i^{(p)}(\beta_j) - u_i^{(p)}(\alpha_j) | = \left| \int_{\alpha_j}^{\beta_j} \dot{u}_i^{(p)}(s)\, ds \right| \leqq \max_{I_\tau} | \dot{u}_i^{(p)}(s) | \; | \beta_j - \alpha_j |.$$

Returning to the sequence $\{z^{(p)}(t)\}$, we know that

$$z^{(p)}(\tau) = \Lambda(\tau, 0)z_0 + \int_0^\tau \Lambda(\tau, s)[G(s)v^{(p)}(s) + h(s)]\,ds.$$

Hence, by weak convergence properties of $\{v^{(p)}(s)\}$,

$$z^{(p)}(\tau) \to \Lambda(\tau, 0)z_0 + \int_0^\tau \Lambda(\tau, s)[G(s)v(s) + h(s)]\,ds.$$

In other words, $z^{(p)}(\tau)$ approaches a point in the set $K(\tau)$ and thus $K(\tau)$ is closed.

The final step required of this Appendix is the proof that $K(\tau)$ is convex. Because of the linearity of (4) in $v(t)$, convexity of $K(\tau)$ would follow from the convexity of the class of admissible controls, $v(t)$. Thus, let $v'(t)$ and $v''(t)$ be any two admissible controls and let α, β be any two nonnegative numbers such that $\alpha + \beta = 1$. Then let $w(t) = \alpha v'(t) + \beta v''(t)$ and consider the following:

$$\begin{array}{c} \alpha a_{1i}(t) \leqq \alpha v_i'(t) \leqq \alpha a_{2i}(t). \\ \underline{\beta a_{1i}(t) \leqq \beta v_i''(t) \leqq \beta a_{2i}(t).} \\ a_{1i}(t) \leqq w_i(t) \leqq a_{2i}(t). \end{array}$$

Hence the components of $w_i(t)$ satisfy the required bounds. Moreover, if $w_i(t)$ is one of the differentiable components of $w(t)$, it arises from the sum $\alpha v_i'(t) + \beta v_i''(t)$, and hence

$$\begin{array}{c} \alpha b_{1i}(t) \leqq \alpha \dot{v}_i'(t) \leqq \alpha b_{2i}(t). \\ \underline{\beta b_{1i}(t) \leqq \beta \dot{v}_i''(t) \leqq \beta b_{2i}(t).} \\ b_{1i}(t) \leqq \dot{w}_i(t) \leqq b_{2i}(t). \end{array}$$

Thus, $w(t)$ is an admissible control for (4), and the Appendix is completed.

REFERENCES

[1] S. S. L. Chang, *Minimal time control with multiple saturation limits*, IEEE Trans. Automatic Control, January 1963.

[2] B. J. Birch and R. Jackson, *The behavior of linear systems with inputs satisfying certain bounding conditions*, J. Electronics Control, 6 (1959), pp. 366–375.

[3] P. Alexandroff and H. Hopf, *Topology*, Springer, Berlin, 1935.

[4] L. M. Graves, *The Theory of Functions of Real Variables*, McGraw-Hill, New York, 1946, Theorem 33 of Chap. XII.

[5] J. P. LaSalle, *The Time Optimal Control Problem*, Ann. of Math. Studies, vol. V, no. 45, Princeton, 1960, pages 1–24.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

ANALYSIS OF LINEAR SYSTEMS BY MEANS OF LAGUERRE FUNCTIONS*

JAMES C. I. DOOGE†

Abstract. The use of Laguerre functions is proposed for the analysis of heavily damped linear systems where the input is not subject to experimental control. An equation is derived which links the corresponding coefficients in the Laguerre function expansions of the input, the impulse response and the output of a linear system. This equation enables the third set of Laguerre coefficients to be calculated when the other two sets of coefficients are known. The connection between the Laguerre function expansion and the representation of the system response by a series of gamma distributions is noted and the latter series identified as defining an analog system composed entirely of branches of linear storage elements.

Introduction. The relationship between the input and the output of a time-invariant linear system can be expressed as

$$y(t) = \int_0^t x(\tau)h(t - \tau)\, d\tau, \tag{1}$$

or

$$y(t) = x(t)*h(t),$$

where $x(t)$ is the input function, $y(t)$ is the output function, and $h(t)$ is the impulse response function of the system.

This impulse response is the output from the system when the input is a delta-function (i.e., a unit impulse). Fourier methods have been used in the analysis of such systems but there is some reason to believe that a method based on Laguerre functions may be more convenient and physically more meaningful in the case of heavily damped systems in which the input is not subject to experimental control. The present paper is an attempt to outline such a method of Laguerre analysis. The author is primarily interested in the input-output system involved in the conversion of rainfall to flood run-off but the approach in this paper should be applicable to any heavily damped system.

Laguerre polynomials and functions. The ordinary Laguerre polynomial of degree n may be defined either by

$$L_n(x) = \sum_{k=0}^{k=n} (-1)^k \binom{n}{k} \frac{x^k}{k!} \tag{2}$$

* Received by the editors May 1, 1964, and in final revised form January 11, 1965.

† Department of Civil Engineering, University College, Cork, Ireland.

or by

$$L_n(x) = \frac{e^x}{n!}\frac{d^n}{dx^n}(e^{-x}x^n), \tag{3}$$

where k and n are nonnegative integers. (It should be noted that some writers use a definition of $L_n(x)$ which is $n!$ times the quantity defined above.) The polynomial is orthonormal with respect to the weighting factor e^{-x} within the range from zero to plus infinity.

$$\int_0^\infty e^{-x}L_m(x)L_n(x)\,dx = \delta_{mn}. \tag{4}$$

The polynomials can be readily computed by means of the following recurrence relationship:

$$(n+1)L_{n+1}(x) = (2n+1-x)L_n(x) - nL_{n-1}(x). \tag{5}$$

Since the Laguerre polynomials exist for every value of n, it is possible to express x^n as a linear function of the first n Laguerre polynomials. It can be shown that

$$\frac{x^n}{n!} = \sum_{k=0}^{k=n}(-1)^k\binom{n}{k}L_k(x), \tag{6}$$

which corresponds exactly to the form of (2):

$$L_n(x) = \sum_{k=0}^{k=n}(-1)^k\binom{n}{k}\frac{x^k}{k!}. \tag{2}$$

Thus a table of Laguerre coefficients, such as those of Lanczos [1], can be used to express x^n as a function of the first n Laguerre polynomials.

The ordinary Laguerre function is defined as

$$f_n(x) = e^{-x/2}L_n(x). \tag{7}$$

In view of the orthonormal relationship of (4), the Laguerre functions are also an orthonormal set, i.e.,

$$\int_0^\infty f_m(x)f_n(x)\,dx = \delta_{mn}. \tag{8}$$

The expression for the ordinary Laguerre function can be written as

$$f_n(x) = \sum_{k=0}^{k=n}(-1)^k\binom{n}{k}\frac{e^{-x/2}x^k}{k!}. \tag{9}$$

The expression $e^{-x/2}x^k/k!$ is of interest in the analysis of heavily damped linear systems since (divided by a scale factor of 2^{k+1}) it represents the result of passing a delta-function through $k+1$ equal storages, each having an average delay time of two units and may indeed be considered as a damped delta-function. It also has the form of a gamma distribution. It can be shown that a damped delta-function of degree n can be represented as a linear function of the first n Laguerre functions:

$$\frac{e^{-x/2}x^n}{n!} = \sum_{k=0}^{k=n} (-1)^k \binom{n}{k} f_k(x). \tag{10}$$

Further properties of the Laguerre polynomials are listed in Erdélyi [2]. Proofs of the properties of Laguerre polynomials and functions can be found in such standard works as Szegö [3] or Sansone [4].

The Laplace transform of the Laguerre polynomial is simple in form (see [5]) and is given by

$$\mathcal{L}[L_n(x)] = \frac{(s-1)^n}{s^{n+1}}. \tag{11}$$

Hence the Laplace transform of the ordinary Laguerre function is given by

$$\mathcal{L}[f_n(x)] = \frac{(s-\frac{1}{2})^n}{(s+\frac{1}{2})^{n+1}}. \tag{12}$$

If a series of Laguerre functions are to be used to represent the arbitrary functions in (1), then we need to know the result of the convolution of two Laguerre functions. In view of (9) and (10), the result of convoluting a Laguerre function of degree m with a Laguerre function of degree n must itself be a linear combination of the first $m + n + 1$ Laguerre functions. It can be shown, however, that all the coefficients in the resulting expression are zero with the exception of the last two. This result was originally derived by applying combinatorial algebra to the form of the Laguerre function based on (2) but the following proof based on the Laplace transform is more compact.

If

$$g(x) = f_m(x) * f_n(x), \tag{13}$$

then

$$\begin{aligned}\mathcal{L}[g(x)] &= \mathcal{L}[f_m(x)]\mathcal{L}[f_n(x)] \\ &= \frac{(s-\frac{1}{2})^m}{(s+\frac{1}{2})^{m+1}} \frac{(s-\frac{1}{2})^n}{(s+\frac{1}{2})^{n+1}} = \frac{(s-\frac{1}{2})^{m+n}}{(s+\frac{1}{2})^{m+n+2}} \\ &= \frac{(s-\frac{1}{2})^{m+n}}{(s+\frac{1}{2})^{m+n+1}} - \frac{(s-\frac{1}{2})^{m+n+1}}{(s+\frac{1}{2})^{m+n+2}};\end{aligned}$$

therefore

$$g(x) = f_{m+n}(x) - f_{m+n+1}(x). \tag{14}$$

The generalized Laguerre polynomial of degree n is defined as follows:

$$L_n^a(x) = \sum_{k=0}^{k=n} \binom{n+a}{n-k} \frac{(-x)^k}{k!}, \tag{15}$$

where k and n are nonnegative integers and a is greater than minus 1. Since they will suffice for the vast majority of applications of the method suggested in the present paper, only the ordinary Laguerre polynomials and functions will be discussed in the body of the paper. The corresponding forms of the equations for the generalized Laguerre polynomials and functions are, however, listed in Appendix A.

Linkage between Laguerre coefficients. O'Donnell [6] has shown that the harmonic coefficients, α_n and β_n, of the impulse response function $h(t)$ can be found from the harmonic coefficients of the input function $x(t)$, namely a_n and b_n, and the harmonic coefficients of the output function $y(t)$, namely A_n and B_n, by means of the linkage equations

$$\alpha_0 = \frac{1}{T}\frac{A_0}{a_0}, \tag{16}$$

$$\alpha_n = \frac{2}{T}\frac{a_n A_n + b_n B_n}{a_n^2 + b_n^2}, \tag{17}$$

$$\beta_n = \frac{2}{T}\frac{a_n B_n - b_n A_n}{a_n^2 + b_n^2}. \tag{18}$$

In these equations T is the common base period (not less than the range of the output function) over which all three functions are harmonically analyzed. Equation (14) enables us to derive an analogous relationship for the case where the functions involved are expressed in terms of ordinary Laguerre functions. The input function, the impulse response function and the output can each be represented by an infinite series of Laguerre functions.

$$x(t) = \sum_{n=0}^{\infty} a_n f_n(t), \tag{19}$$

$$h(t) = \sum_{n=0}^{\infty} \alpha_n f_n(t), \tag{20}$$

$$y(t) = \sum_{n=0}^{\infty} A_n f_n(t), \tag{21}$$

where $f_n(t)$ is the nth Laguerre function. Since the functions are orthonormal, the coefficients in (19), (20) and (21) are given by

$$a_n = \int_0^\infty x(t) f_n(t)\, dt, \tag{22}$$

$$\alpha_n = \int_0^\infty h(t) f_n(t)\, dt, \tag{23}$$

$$A_n = \int_0^\infty y(t) f_n(t)\, dt. \tag{24}$$

Now if we return to the basic equation linking input and output,

$$y(t) = h(t) * x(t), \tag{1}$$

this becomes

$$\begin{aligned} y(t) &= \sum_{m=0}^{\infty} \alpha_m f_m(t) * \sum_{n=0}^{\infty} a_n f_n(t) \\ &= \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \alpha_m f_m(t) * a_n f_n(t), \end{aligned}$$

which we can write as

$$\sum_{p=0}^{\infty} A_p f_p(t) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \alpha_m a_n [f_{m+n}(t) - f_{m+n+1}(t)]. \tag{25}$$

In order to find a linkage equation, it is necessary to identify the coefficient of an individual Laguerre function $f_p(t)$ on both sides of (25). But f_p can only appear on the right hand side of the equation when either $m + n = p$ or $m + n + 1 = p$. Hence

$$A_p f_p(t) = \sum_{k=0}^{k=p} \alpha_k a_{p-k} f_p(t) - \sum_{k=0}^{k=p-1} \alpha_k a_{p-1-k} f_p(t). \tag{26}$$

The required linkage equation can therefore be written as

$$A_p = \sum_{k=0}^{k=p} \alpha_k a_{p-k} - \sum_{k=0}^{k=p-1} \alpha_k a_{p-1-k} . \tag{27}$$

If two of the sets of coefficients are known, the third set can be determined by means of the linkage equation (27). It is not necessary to solve simultaneously for all values of α (or other unknown set of coefficients) since we can start with $p = 0$ and substitute each value as found. Thus

$$\sum_{k=0}^{p} \alpha_k a_{p-k} = A_p + \sum_{k=0}^{k=p-1} \alpha_k a_{p-1-k} = \sum_{k=0}^{k=p} A_k,$$

so that

$$\alpha_p a_0 = \sum_{k=0}^{k=p} A_k - \sum_{k=0}^{k=p-1} \alpha_k a_{p-k} . \tag{28}$$

Thus, by means of (28), the successive values of α_p can be readily computed one by one.

Equations (19) to (28) are based on Laguerre expansions with an infinite number of terms and Laguerre coefficients defined by integration over an infinite range. In practical applications, the data will be finite in extent, some form of numerical integration will be involved and the Laguerre representation of the functions will be a finite series. These limitations will affect the accuracy of the solution to a varying degree which will depend on the nature of the input and of the system response. The examples given later in the paper will illustrate this effect.

Response characteristics of heavily damped systems. For heavily damped systems, the response to a unit impulse will show no oscillatory features but will be positive throughout the range from zero to infinity. In fact, the response approaches zero asymptotically and will become negligible after a certain time.

In a system consisting of a series of r equal elements of linear storage each having a delay time k, the response is

$$h(t) = \frac{e^{-t/k}(t/k)^{r-1}}{k(r-1)!}. \tag{29}$$

For a series of r such storages in parallel with a second series of s storages of equal size the response is given by

$$h(t) = C_r \frac{e^{-t/k}(t/k)^{r-1}}{k(r-1)!} + C_s \frac{e^{-t/k}(t/k)^{s-1}}{k(s-1)!}, \tag{30}$$

in which $C_r + C_s = 1$. The determination of an unknown response function by Laguerre analysis and the application of (9) and (10) derived earlier enables us to express the response function as a series of terms of the type given in (30).

If the system under examination actually consists of equal storage elements arranged both in parallel and series, then analysis by means of Laguerre functions could be used to determine the size of the storage elements and the number of elements in each branch. In the more general case of any heavily damped system the same procedure would give the parameters of a system of linear storage elements equivalent in its effect to the system under examination. The gamma distribution series expansion representing the impulse response function of a heavily damped system to a given degree of accuracy could be expected to contain less terms than the corresponding series obtained by harmonic analysis since the individual terms would be positive throughout the whole range.

Since Laguerre functions are defined in terms of $e^{-t/2}$ and the gamma-distribution (or damped delta-function) in terms of $e^{-t/k}$, it is necessary to multiply the data times by a time factor of $2/k$ before starting the analysis. If the double argument system of Lanczos [1] is used, a time scale of $1/k$ should be used. For a completely unknown system the appropriate scale must be found by trial. If the system can be represented by a finite model system composed of equal storage elements, then translation to the scale appropriate to the size of these storage elements will lead to a response function having a finite series of terms each with positive coefficients and of the damped delta-function type. Where the model system approximating to the response system under examination contains storages of unequal size, no single time scale will reduce the series to a finite number of terms.

The solution based on Laguerre analysis has the added advantage that it automatically provides a model of the system in terms of linear storage elements, thus allowing the impulse response of the system to be written directly. Such models are often useful in the comparison of different systems of unknown constitution and in the formation of theoretical hypotheses concerning the phenomena being studied.

The results of Laguerre analysis can be represented in the form of a "storage spectrum" in which the coefficients C_r, C_s, etc., of the gamma distributions are plotted against the corresponding values of rk, sk, etc. The latter values represent the mean delay of the cascade of storage elements equivalent to each term. Such a storage spectrum would indicate immediately whether the system could be readily separated into distinct parts with appreciably different delay times.

Examples of computation. The use of Laguerre functions to determine an impulse response function is illustrated below by two simple examples. In the first of these, all of the functions involved can be represented exactly by a short series of Laguerre functions. Hence the coefficients can be readily computed by hand and the whole procedure closely followed. In the second example, the functions involved are zero outside limited ranges and hence cannot be represented exactly by finite Laguerre expansions.

If we have a system whose impulse response function is given by

$$h(t) = \frac{e^{-t/2}t^3}{3!}, \tag{31}$$

it can be readily shown by direct convolution that for an input

$$x(t) = \frac{1}{2}\left(t + \frac{t^2}{2}\right)e^{-t/2}, \tag{32}$$

the output will be given by

$$y(t) = \frac{1}{2}\left(\frac{t^5}{5!} + \frac{t^6}{6!}\right)e^{-t/2}. \tag{33}$$

Since all of the terms in these functions consist of gamma distributions, they can, by virtue of (9) and (10), be written in terms of Laguerre functions as follows:

$$h(t) = f_0(t) - 3f_1(t) + 3f_2(t) - f_3(t), \tag{34}$$

$$x(t) = f_0(t) - 1.5f_1(t) + 0.5f_2(t), \tag{35}$$

$$\begin{aligned} y(t) = f_0(t) - 5.5f_1(t) + 12.5f_2(t) - 15.0f_3(t) \\ + 10.0f_4(t) - 3.5f_5(t) + 0.5f_6(t). \end{aligned} \tag{36}$$

If we consider the input and output as known and the impulse response function as unknown, the Laguerre coefficients of the response can be readily determined by the inversion of the linkage equation.

The given coefficients and calculated coefficients for the impulse response are given in Table 1. The successive values of α_p can be readily calculated by means of (28). For example, the value of α_2 is determined as follows:

$$\alpha_2 a_0 = \sum_{k=0}^{k=2} A_k - \sum_{k=0}^{k=1} \alpha_k a_{2-k}\,, \tag{28}$$

$$\begin{aligned} \alpha_2 = [(1) - (5.5) + (12.5)] \\ - [(1)(0.5) + (-3)(-1.5)] = (8) - (5) = 3. \end{aligned}$$

Empirical input and output data are usually in discrete form and in such cases the Laguerre coefficients will be estimated by some form of numerical integration. The procedure for discrete data was examined by programming the problem on an IBM 1620 computer.

A program was developed to do the following: (a) to generate synthetic values of the input and output at fixed intervals of time; (b) to compute the Laguerre coefficients of these inputs and outputs by numerical integration; (c) to use these computed values of the Laguerre coefficients of input and output to determine the Laguerre coefficients of the impulse response function; (d) to compare the response function generated by these com-

Table 1

Term n	Input a_n (given)	Output A_n (given)	Response α_n (calculated)
0	1	1	1
1	−1.5	−5.5	−3
2	+0.5	+12.5	+3
3	—	−15.0	−1
4	—	+10.0	—
5	—	−3.5	—
6	—	+0.5	—

Table 2

Term n	Input a_n (by integration)	Output A_n (by integration)	Response α_n (by inversion)
0	9.999×10^{-1}	9.999×10^{-1}	1.000
1	-1.500	-5.499	-3.000
2	4.999×10^{-1}	1.249×10	2.999
3	-8.047×10^{-5}	-1.499×10	-1.000
4	-1.332×10^{-4}	9.999	8.600×10^{-6}
5	1.983×10^{-4}	-3.499	1.226×10^{-5}
6	-2.752×10^{-4}	4.999×10^{-1}	-2.484×10^{-5}
7	-3.633×10^{-4}	1.560×10^{-4}	1.336×10^{-4}
8	-4.620×10^{-4}	-4.801×10^{-4}	-3.427×10^{-4}
9	-5.702×10^{-4}	1.174×10^{-3}	8.314×10^{-4}
10	-6.876×10^{-4}	-2.240×10^{-3}	-1.396×10^{-3}

puted Laguerre coefficients with the original function used to synthesize the output.

The program was first applied to the functions used in the above example. For a time interval of $t = 0.2$ and integration up to the time $t = 40$, the computed values of the Laguerre coefficients were as shown in Table 2. As can be seen by comparing Tables 1 and 2, the Laguerre coefficients obtained by numerical methods closely approximate the exact values. When the impulse response function generated by the 10 coefficients found from the linkage equation was compared to the original impulse response, the root mean square absolute error over the range $t = 0$ to $t = 40$ was found to be 2.5×10^{-4}, which is small compared to the maximum ordinate of 1.8.

As a second example, let us consider a system whose response function cannot be represented exactly by a finite number of Laguerre functions. The functions chosen are those used by O'Donnell in some of his more recent work on response analysis via Fourier series [7]. All of the functions are zero outside the limited ranges indicated below.

The response function is

$$h(t) = \frac{t}{10}(e^{8-t} - 1), \qquad 0 \leqq t \leqq 8.$$

For an input given by

$$x(t) = 10t(1 - t)e^{1-t}, \qquad 0 \leqq t \leqq 1,$$

the output is in three segments

$$\text{(i)} \quad 0 \leqq t \leqq 1, \quad y(t) = e^{9-t}\frac{t^3}{12}(2 - t) + e^{1-t}(t^2 + 3t + 4) - e(4 - t);$$

$$\text{(ii)} \quad 1 \leqq t \leqq 8, \qquad y(t) = e^{9-t}\,\frac{(2t-1)}{12} + (11 - 3t) - e(4 - t);$$

$$\text{(iii)} \quad 8 \leqq t \leqq 9, \qquad y(t) = \frac{e^{9-t}}{12}\,[12t^2 - 130t + 335 - (t-8)^2(152 - 14t - t^2)] + (11 - 3t).$$

Values of the input and output were generated in the computer and used to calculate the Laguerre coefficients of the response function. These coefficients were then used to generate the values of the response functions at fixed values of t. Because of the form of the exponentials in the functions used, the analysis was carried out using a time scaling factor of 2.

Table 3 shows the comparison of the ordinates of the impulse response function originally used to derive the synthetic output with the ordinates found by Laguerre analysis using various numbers of coefficients and a time interval of 0.05. The root mean square absolute error over the range $t = 0$ to $t = 9.25$ is also shown. The very close representation by a series of only two terms is due to the fact that the response function peaks at $t = 1$ and then decays rapidly.

Neither of the examples reveals any serious difficulty in the numerical application of Laguerre analysis to noiseless synthetic data. Whether it can be as easily applied to empirical data is a matter for separate investigation in the different fields concerned with heavily damped systems. In the two

TABLE 3

Time t	Original function	Impulse response function			
		$N = 2$	$N = 10$	$N = 20$	$N = 30$
0	0.00	0.40	0.07	−0.10	−0.37
.5	90.4	90.3	90.4	90.3	90.4
1.0	109.5	109.3	109.5	109.5	109.5
1.5	99.6	99.5	99.6	99.6	99.6
2.0	80.5	80.4	80.5	80.5	80.5
2.5	60.9	61.0	60.9	60.9	60.9
3.0	44.2	44.4	44.2	44.2	44.2
4.0	21.4	21.8	21.5	21.4	21.4
5.0	9.5	10.0	9.6	9.5	9.5
6.0	3.8	4.4	3.8	3.9	3.9
7.0	1.2	1.9	1.2	1.2	1.2
8.0	0.00	0.80	0.19	0.12	0.11
9.0	0.00	0.33	−0.07	−0.04	−0.04
RMS Error	—	4.4×10^{-1}	4.7×10^{-2}	3.0×10^{-2}	4.2×10^{-2}

examples the time scale was chosen to suit the exponential term of the response function involved. The determination of the appropriate scale would be a necessary part of the Laguerre analysis of a system whose natural time scale was unknown.

Appendix A. *Generalized Laguerre polynomials and functions.* A generalized Laguerre polynomial can be defined in either of the following forms:

$$(2') \qquad L_n^a(x) = \sum_{k=0}^{k=n} \binom{n+a}{n-k} \frac{(-x)^k}{k!},$$

or

$$(3') \qquad L_n^a(x) = \frac{e^x x^{-a}}{n!} \frac{d^n}{dx^n} (e^{-x} x^{n+a}).$$

These polynomials have the following orthogonal property:

$$(4') \quad \int_0^\infty e^{-x} x^a L_m^a(x) L_n^a(x)\, dx = \left[\frac{\Gamma(m+a+1)\Gamma(n+a+1)}{\Gamma(m+1)\Gamma(n+1)}\right]^{1/2} \delta_{mn},$$

and are connected by the recurrence relationship

$$(5') \quad (n+1)L_{n+1}^a(x) = (2n+a+1-x)L_n^a(x) - (n+a)L_{n-1}^a(x)$$

As in the case of the ordinary polynomials, the inverse relationship between x^n and $L_n^a(x)$ is the same as the direct relationship, i.e.,

$$(6') \qquad \frac{x^n}{n!} = \sum_{k=0}^{k=n} (-1)^k \binom{n+a}{n-k} L_k^a(x).$$

The generalized Laguerre function is defined by

$$(7') \qquad f_n^a(x) = e^{-x/2} x^{a/2} L_n^a(x) \left\{\frac{\Gamma(n+1)}{\Gamma(n+a+1)}\right\}^{1/2},$$

and has the following orthonormal property

$$(8') \qquad \int_0^\infty f_m^a(x) f_n^a(x)\, dx = \delta_{mn}.$$

The Laplace transforms of the generalized polynomials and functions are given by

$$(11') \qquad \mathcal{L}[x^a L_n^a(x)] = \frac{\Gamma(n+a+1)}{\Gamma(n+1)} \cdot \frac{(s-1)^n}{(s)^{n+a+1}},$$

and

$$(12') \qquad \mathcal{L}[x^{a/2} f_n^a(x)] = \left\{\frac{\Gamma(n+a+1)}{\Gamma(n+1)}\right\}^{1/2} \frac{(s-\frac{1}{2})^n}{(s+\frac{1}{2})^{n+a+1}}.$$

If

$$g(x) = x^{a/2} f_m{}^a(x) * x^{a/2} f_n{}^a(x) \tag{13'}$$

then

$$\begin{aligned} g(x) = x^a & \left\{ \frac{\Gamma(m + a + 1)\Gamma(n + a + 1)}{\Gamma(m + 1)\Gamma(n + 1)} \right\}^{1/2} \\ & \cdot \left[\left\{ \frac{\Gamma(m + n + 1)}{\Gamma(m + n + 2a + 1)} \right\}^{1/2} f_{m+n}^{2a}(x) \right. \\ & \left. - \left\{ \frac{\Gamma(m + n + 2)}{\Gamma(m + n + 2a + 2)} \right\}^{1/2} f_{m+n+1}^{2a}(x) \right]. \end{aligned} \tag{14'}$$

The above equation can be used to derive a general linkage equation for the case of the generalized Laguerre functions. Thus, if we have $y(t) = x(t)*h(t)$, we can write

$$x(t) = \sum_{n=0}^{n=\infty} a_n t^{a/2} f_n{}^a(t), \tag{19'}$$

where

$$a_n = \int_0^\infty t^{-a/2} x(t) f_n{}^a(t)\, dt; \tag{22'}$$

and

$$h(t) = \sum_{n=0}^{n=\infty} \alpha_n t^{-a/2} f_n{}^a(t), \tag{20'}$$

where

$$\alpha_n = \int_0^\infty t^{-a/2} h(t) f_n{}^a(t)\, dt; \tag{23'}$$

and

$$y(t) = \sum_{n=0}^{n=\infty} A_n t^a f_n^{2a}(t), \tag{21'}$$

where

$$A_n = \int_0^\infty t^{-a} y(t) f_n{}^{2a}(t)\, dt. \tag{24'}$$

By proceeding as shown in the body of the paper for the ordinary Laguerre functions, we can obtain the following expression for the calculation of the Laguerre coefficients of the response functions:

$$\left\{\frac{\Gamma(p+\alpha+1)\Gamma(\alpha+1)}{\Gamma(p+1)}\right\}^{1/2}\alpha_p a_0=\left\{\frac{\Gamma(p+2\alpha+1)}{\Gamma(p+1)}\right\}^{1/2}\sum_{k=0}^{k=p}A_k \tag{28'}$$
$$-\sum_{k=0}^{k=p-1}\left\{\frac{\Gamma(k+\alpha+1)\Gamma(p-k+\alpha+1)}{\Gamma(k+1)\Gamma(p-k+1)}\right\}^{1/2}\alpha_k a_{p-k}\,.$$

Appendix B. *Notation.*

a =	parameter of generalized Laguerre polynomial
a_n =	coefficient in Laguerre expansion of input
A_n =	coefficient in Laguerre expansion of output
C_n =	coefficient in series of gamma terms
$f_n(x)$ =	ordinary Laguerre function of degree n
$f_n^{\,a}(x)$ =	generalized Laguerre function of degree n
$g(x)$ =	result of convolution
$h(t)$ =	impulse response function
$L_n(x)$ =	ordinary Laguerre polynomial
$L_n^{\,a}(x)$ =	generalized Laguerre polynomial
s =	variable in the Laplace transform
$x(t)$ =	input to linear system
$y(t)$ =	output from linear system
α_n =	coefficient in Laguerre expansion of impulse response
δ_{mn} =	Kronecker delta ($\delta_{mn} = 1$ if $m = n$; $\delta_{mn} = 0$ if $m \neq n$)
$\delta(t)$ =	Dirac delta-function or impulse
$\mathcal{L}[\ ]$ =	Laplace transform
$\Gamma(n+1)$ =	$n! =$ factorial n
$\Gamma(n+a+1)$ =	gamma function
$*$ =	operation of convolution
$\binom{n}{k}$ =	the binomial coefficient

Acknowledgments. The author is indebted to C. Lanczos of the Dublin Institute for Advanced Studies and T. O'Donnell of Imperial College, London, for their helpful comments on an early version of this paper and to P. O'Regan of University College, Cork, who devised and ran the necessary computer programs.

REFERENCES

[1] C. Lanczos, *Applied Analysis*, Pitman, London, 1957, Table X.

[2] A. Erdélyi, *Higher Transcendental Functions*, Bateman Manuscript Project, vol. 2, McGraw-Hill, New York, 1953, Chap. X.

[3] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, New York, 1939.

[4] G. Sansone, *Orthogonal Functions*, Interscience, New York, 1959, Chaps. I and IV.

[5] G. Doetsch, *Theorie und Anwendung der Laplace-Transformation*, Dover, New York, 1943, p. 403, transforms 41 and 42.

[6] T. O'Donnell, *Instantaneous unit hydrograph derivation by harmonic analysis*, International Association of Scientific Hydrology, 51 (1960), pp. 546–557.

[7] ———, Private communication, 1964.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

PENALTY FUNCTIONS AND BOUNDED PHASE COORDINATE CONTROL*

D. L. RUSSELL†

Abstract. This paper studies the use of two different kinds of penalty functions to obtain approximate and, in the limit, exact solutions to a class of bounded phase coordinate optimal control problems. The first type of penalty function assumes small values within the phase constraint and large values outside, while the second type is defined only within the phase constraints, assuming small values away from the constraint boundary but increasing to infinity as that boundary is approached.

0. Introduction. Much attention has recently been given to the problem of bounded, or, more generally, constrained phase optimal control. Several papers, in particular [1] and [2], have dealt with this problem.

One method of attack on this problem is the following: instead of attempting a direct solution of the constrained phase optimal control problem, an unconstrained problem is considered wherein the original cost functional is augmented by a nonnegative penalty function which sharply increases the cost associated with trajectories which violate the phase constraints. By using sequences of cost functionals involving more and more severe penalty functions it is to be expected in many cases that the desired constrained phase solution of the original optimization problem may be approximated to any desired degree of accuracy by solutions to these unconstrained problems.

The purpose of this paper is to study this method of approximation for rather general sequences of penalty functions. In §1 the optimization problem is rigorously posed and two different kinds of sequences of penalty functions are defined. The question of the existence of optimal solutions for unconstrained problems involving a given penalty function is considered in §2. In §3 we study the convergence of such solutions to a solution of the original constrained problem. Finally in §4 we relate our work to papers which have been published by other authors.

1. Statement of the problem. Consider the system of n ordinary differential equations

$$\dot{x}^i = g^i(t, x) + \sum_{j=1}^{m} h_j^i(t, x)u^j(t), \qquad i = 1, \cdots, n. \tag{1.1}$$

* Received by the editors February 21, 1964, and in revised form December 15, 1964.

† Mathematics Research Center, U. S. Army, The University of Wisconsin, Madison, Wisconsin. This paper was prepared under Contract NASw-563 for the National Aeronautics and Space Administration, while the author was employed with Honeywell, Incorporated, St. Paul, Minnesota.

In vector notation this becomes

$$(1.2) \qquad \dot{x} = g(t, x) + H(t, x)u(t).$$

The dot denotes differentiation with respect to t. In (1.2) the vectors x and g are n-dimensional and the vector u is m-dimensional while the matrix H has dimensions n by m.

Throughout this paper I will denote a compact interval of the line E^1 with nonzero length while G will be a closed subset of E^n which possesses a nonempty interior. The symbol Ω will represent a compact, convex subset of E^m having a nonempty interior. We will use O to indicate an open subset of E^n which contains G. $T(t)$ denotes a compact subset of $\operatorname{int}(G)$ which varies continuously with $t \in I$ and x_0 is a point of $\operatorname{int}(G)$ such that x_0 does not lie in $T(t)$ for any $t \in I$.

We assume that $g(t, x)$ and $H(t, x)$ are continuous and have continuous first order partial derivatives with respect to x in $I \times G$. Whenever a symbol u is used it will denote a measurable vector function whose range is a subset of Ω. Such a function u will be assumed to be defined on a subinterval $I_u = [a_u, b_u]$ of I. Corresponding to such a function u is the unique solution x of (1.2) which has the property $x(a_u) = x_0$. Whenever a function u is distinguished by some marking (e.g., $\hat{u}$, u^*) the corresponding x with $x(a_u) = x_0$ will be similarly distinguished (e.g., $\hat{x}$, x^*, respectively). We will use the symbol (u, x) to denote a control and trajectory pair.

Let Δ be the set of all pairs (u, x) such that in each case (i) $x(t) \in G$ for $a_u \leqq t \leqq b_u$; (ii) $x(b_u) \in T(b_u)$ and for all $t < b_u$, $x(t) \notin T(t)$; (iii) $x(a_u) = x_0$.

Let $C(u)$ be defined for each pair (u, x) in Δ by

$$(1.3) \qquad C(u) = \int_{a_u}^{b_u} \{g^0(t, x(t)) + \sum_{j=1}^{m} h_j^0(t, x(t))u^j(t)\}\, dt,$$

where g^0 and h_j^0, $j = 1, \cdots, m$, are continuous in $I \times G$ and bounded on bounded subsets of $I \times G$.

The optimization problem is this: determine a pair $(\bar{u}, \bar{x}) \in \Delta$ such that

$$(1.4) \qquad C(\bar{u}) = \min_{(u,x) \in \Delta} C(u).$$

This is the constrained optimization problem. If in the definition of Δ we replace G by the open set $O \supset G$ and assume the functions $g(t, x)$, $H(t, x)$, $g^0(t, x)$, $h_j^0(t, x)$, $j = 1, \cdots, m$, to be defined and bounded in bounded subsets of O and otherwise possessing the properties already specified, then the resulting optimization problem in the domain O will be called *unconstrained*. Moreover, any solution of an unconstrained problem or any solution of a constrained problem which is such that x does not meet ∂G will be called an *unconstrained solution.*

DEFINITION 1.1. *The sequence of functions* $\{p_i(x)\}$ *is said to be a sequence of penalty functions of the first kind for* G *in case there is an open set* O *containing* G *such that* (i) $p_i(x)$ *is defined and continuous on* O *and assumes only nonnegative values there for* $i = 1, 2, \cdots$; (ii) *given a compact set* $D \subset \operatorname{int}(G)$ *we have*

$$\lim_{i\to\infty} (\max_{x\in D} p_i(x)) = 0; \tag{1.5}$$

(iii) *given any compact set* $D \subset O - G$,

$$\lim_{i\to\infty} (\min_{x\in D} p_i(x)) = +\infty. \tag{1.6}$$

DEFINITION 1.2. *The sequence of functions* $\{p_i(x)\}$ *is said to be a sequence of penalty functions of the second kind for* G *in case*: (i) $p_i(x)$ *is defined and continuous and assumes only nonnegative values in* $\operatorname{int}(G)$, $i = 1, 2, \cdots$; (ii) *given a compact set* $D \subset \operatorname{int}(G)$,

$$\lim_{i\to\infty} (\max_{x\in D} p_i(x)) = 0; \tag{1.7}$$

(iii) *letting* $B(\delta)$ *denote the set of all points* $x \in \operatorname{int}(G)$ *such that the Euclidean distance from* x *to* ∂G *is less than* δ *we have*

$$\lim_{\delta\to 0} (\text{g.l.b.}_{x\in B(\delta)}\, p_i(x)) = +\infty, \qquad i = 1, 2, \cdots ; \tag{1.8}$$

(iv) *for each vector valued function* $x(s)$ *defined, absolutely continuous, and possessing a uniformly bounded derivative* (*where defined*) *on* $[0, 1]$ *and such that for* $s \in [0, 1)$, $x(s) \in \operatorname{int}(G)$, *and* $x(1) \in \partial G$, *we have, for* $i = 1, 2, \cdots$,

$$\int_0^1 p_i(x(s))\, ds = +\infty. \tag{1.9}$$

DEFINITION 1.3. *The sequence of pairs* $\{(u_k, x_k)\} \subseteq \Delta$ *will be said to approximate the pair* $(u, x) \in \Delta$ *in case*: (i) *we have*

$$\lim_{k\to\infty} a_{u_k} = a_u \quad \textit{and} \quad \lim_{k\to\infty} b_{u_k} = b_u ; \tag{1.10}$$

(ii) *given any compact subinterval* J *of* I_u ,

$$\lim_{k\to\infty} (\max_{t\in J} \| x_k(t) - x(t) \|) = 0; \tag{1.11}$$

(iii) *extending the definition of the* u_k *and* u *to* I *by setting them equal to zero where previously undefined, the sequence* $\{u_k\}$ *converges to* u *in the weak topology of* $L_2(I)$.

2. Existence theorems. The work of E. B. Lee and L. Markus in [4] provides a proof for the following theorem.

THEOREM 2.1. *Let the optimization problem, either constrained or uncon-*

strained, be defined as above. Let $\{(u_k, x_k)\}$ *be a sequence of pairs in* Δ *such that for some positive number* B,

$$\| x_k(t) \| \leqq B, t \in I_{u_k}, \qquad k = 1, 2, \cdots . \tag{2.1}$$

Then there are a function u *and a function* x *defined on an interval* I_u *and a subsequence of* $\{(u_k, x_k)\}$, *which we shall still call* $\{(u_k, x_k)\}$, *such that*

$$\lim_{k\to\infty} a_{u_k} = a_u, \qquad \lim_{k\to\infty} b_{u_k} = b_u ; \tag{2.2}$$

$$\lim_{k\to\infty} u_k = u, \tag{2.3}$$

in the weak topology of $L_2(I)$ *(defining the functions to be zero where previously undefined)*;

$$\lim_{k\to\infty} (\max_{t\in J} \| x_k(t) - x(t) \|) = 0 \tag{2.4}$$

on each compact subinterval $J \subseteq I_u$. *Moreover, if the optimization problem is a constrained problem, or if* $O = E^n$, *or if* x *lies wholly in* O, *then*

$$(u, x) \in \Delta, \tag{2.5}$$

and

$$\lim_{k\to\infty} C(u_k) = C(u). \tag{2.6}$$

Lee and Markus do not state the *uniform* convergence of the x_k to x on J but this is an immediate consequence of the boundedness of $g(t, x)$, $H(t, x)$ and the u_k in the domains in question.

Corollary 2.1. *If there is a number* $B > 0$ *such that for all* $(u, x) \in \Delta$,

$$\| x(t) \| < B, \qquad t \in I_u, \tag{2.7}$$

and if Δ *is nonempty, the constrained optimization problem has a solution* $(\bar{u}, \bar{x})$.

Proof. Since g^0 and h_j^0, $j = 1, \cdots, m$, are bounded in $G \cap \{x \mid \| x \| \leqq B\}$, there is a real number M such that

$$C(u) > M, \qquad u \in \Delta. \tag{2.8}$$

By the same reasoning and since Δ is nonempty, there is some (u, x) in Δ with finite cost $C(u)$. Hence we may assume that M is the largest such number. Then there is a sequence of (not necessarily distinct) pairs $(u_k, x_k) \in \Delta$ such that

$$\lim_{k\to\infty} C(u_k) = M. \tag{2.9}$$

From Theorem 2.1 there is a pair $(\bar{u}, \bar{x}) \in \Delta$ and a subsequence of $\{(u_k, x_k)\}$, which we still call $\{(u_k, x_k)\}$, such that $\{(u_k, x_k)\}$ approximates $(\bar{u}, \bar{x})$ and

$$C(\bar{u}) = \lim_{k\to\infty} C(u_k) = M. \tag{2.10}$$

Then $(\bar{u}, \bar{x})$ is a solution to the constrained optimization problem.

COROLLARY 2.2. *Let $\{p_i(x)\}$ be a sequence of penalty functions of the first kind for G defined in an open set $O \supset G$. If i is sufficiently large then the unconstrained optimization problem with $C(u)$ replaced by*

$$C_i(u) = C(u) + \int_{a_u}^{b_u} p_i(x(t))\, dt \tag{2.11}$$

has a solution (u^i, x^i) such that x^i lies wholly in O, provided Δ satisfies (2.7) *and there is a pair $(\hat{u}, \hat{x}) \in \Delta$ such that $\hat{x}$ lies in* $\operatorname{int}(G)$.

Proof. Let D denote the set

$$D = O \cap \{x \mid \| x \| \leqq B\}. \tag{2.12}$$

Since $g^0(t, x)$ and $h_j^0(t, x), j = 1, \cdots, m$, are bounded in D and each $p_i(x)$ is nonnegative in D, the nonemptiness of Δ implies that for each i there is a largest real number M_i such that

$$C_i(u) \geqq M_i, \qquad (u, x) \in \Delta. \tag{2.13}$$

Let the pair $(\hat{u}, \hat{x}) \in \Delta$ be such that $\hat{x}$ lies wholly in the interior of G. Let D_0 be a compact subset of $\operatorname{int}(G)$ such that

$$\hat{x}(t) \in D_0, \qquad t \in I_{\hat{u}}. \tag{2.14}$$

Let D_2 be a relatively open subset of D containing $G \cap \{x \mid \| x \| \leqq B\}$ whose closure $\bar{D}_2$ is compact and is contained in D. Let D_3 be a compact subset of D whose interior, relative to D, contains $\bar{D}_2$. Let

$$D_1 = D_3 - D_2. \tag{2.15}$$

Then D_1 is a compact subset of $O - G$.

From Definition 1.1 it is clear that there is a number M such that

$$C_i(\hat{u}) < M, \qquad i = 1, 2, \cdots. \tag{2.16}$$

For each i, let $\{(u_k^i, x_k^i)\}$ be a sequence in Δ such that

$$\lim_{k\to\infty} C_i(u_k^i) = M_i. \tag{2.17}$$

Since $M_i < M$, $i = 1, 2, \cdots$, we may assume without loss of generality that

$$C_i(u_k^i) < M, \qquad i = 1, 2, \cdots; \quad k = 1, 2, \cdots. \tag{2.18}$$

Let (u^*, x^*) be any pair in Δ such that x^* meets $O - D_3$. Since $g(t, x)$, $H(t, x)$ and u^0 are bounded there are a number $\delta > 0$ and at least two intervals $[t_1, t_2]$ and $[t_3, t_4]$, each of length $\geqq \delta$, such that

$$x^*(t) \in D_1, \qquad t \in [t_1, t_2] \cup [t_3, t_4]. \tag{2.19}$$

Then

$$\begin{aligned} C_i(u^*) &= \left(\int_{a_u^*}^{t_1} + \int_{t_1}^{t_2} + \int_{t_2}^{t_3} + \int_{t_3}^{t_4} + \int_{t_4}^{b_u^*}\right) \\ &\cdot\left(g^0(t, x^*(t)) + \sum_{j=1}^{m} h_j^0(t, x^*(t))u^*(t) + p_i(x^*(t))\right) dt. \end{aligned} \tag{2.20}$$

Since $g^0(t, x)$ and $h_j^0(t, x)$, $j = 1, \cdots, m$, are bounded in D and $p_i(x)$ is nonnegative, there is a real number $M_1 > 0$ such that for all i,

$$C_i(u^*) \geqq \int_{[t_1,t_2] \cup [t_3,t_4]} p_i(x^*(t))\, dt - M_1. \tag{2.21}$$

But then, from (iii) of Definition 1.1,

$$\lim_{i \to \infty} C_i(u^*) = +\infty, \tag{2.22}$$

and hence, for i sufficiently large,

$$M < C_i(u^*), \tag{2.23}$$

and (u^*, x^*) cannot be a member of the sequence $\{(u_k^i, x_k^i)\}$. Thus all of the members of sequences $\{(u_k^i, x_k^i)\}$, for i sufficiently large, are such that x_k^i always lies in D_3. Then Theorem 2.1 immediately gives the existence of the optimal (u^i, x^i) for such i with $x^i(t) \in O$, $t \in I_{u^i}$.

COROLLARY 2.3. *Let $\{p_i(x)\}$ be a sequence of penalty functions of the second kind for G. Assume that the constrained optimization problem is such that (2.7) is satisfied for $(u, x) \in \Delta$ and there is a $(\hat{u}, \hat{x}) \in \Delta$ such that $\hat{x}$ lies wholly in* int(G). *Consider a new constrained optimization problem with $C(u)$ replaced by*

$$C_i(u) = C(u) + \int_{a_u}^{b_u} p_i(x(t))\, dt. \tag{2.24}$$

Then for each i the new constrained optimization problem possesses an unconstrained optimal solution (u^i, x^i) such that x^i lies wholly in int(G).

Proof. Again there is a largest real number M_i such that

$$C_i(u) \geqq M_i \quad \text{for} \quad (u, x) \in \Delta, \qquad i = 1, 2, \cdots. \tag{2.25}$$

Fix some integer i. Let $\{(u_k^i, x_k^i)\}$ be a sequence in Δ such that

$$\lim_{k \to \infty} C_i(u_k^i) = M_i. \tag{2.26}$$

By Theorem 2.1 there is a pair $(u^i, x^i) \in \Delta$ such that for some subsequence, which we still call $\{(u_k^i, x_k^i)\}$, $\{u_k^i\}$ converges weakly to u^i, $\{x_k^i\}$ converges uniformly to x^i on compact subsets of I_{u^i}, and

$$\lim_{k\to\infty} C(u_k^i) = C(u^i). \tag{2.27}$$

Since (2.26) holds we conclude that

$$\lim_{k\to\infty} \left(\int_{a_{u_k}}^{b_{u_k}} p_i(x_k^i(t))\, dt \right) = M_i - C(u^i). \tag{2.28}$$

Suppose there were a time $t^* \in I_{u^i}$ such that $x^i(t^*) \in \partial G$. We may assume that t^* is the smallest such time. Then from condition (iv) in the definition of a penalty function of the second kind we conclude that, given a small number $\delta > 0$,

$$\int_{a_{u^i}+\delta}^{t^*} p_i(x^i(t))\, dt = +\infty, \tag{2.29}$$

since, as is easily verified, $x^i(t)$ satisfies (1.2) on $[a_{u^i} + \delta, t^*)$ with $u(t)$ replaced by $u^i(t)$ and hence has a uniformly bounded derivative there. Now since $\{x_k^i\}$ converges uniformly to x^i on $[a_{u^i} + \delta, t^*]$ we certainly have

$$\lim_{k\to\infty} p_i(x_k^i(t)) = p_i(x^i(t)) \tag{2.30}$$

for $t \in [a_{u^i} + \delta, t^*)$. By Fatou's Lemma, therefore,

$$\begin{aligned} \int_{a_{u^i}+\delta}^{t^*} p_i(x^i(t))\, dt &= \int_{a_{u^i}+\delta}^{t^*} (\lim_{k\to\infty} p_i(x_k^i(t)))\, dt \\ &\leqq \lim_{k\to\infty} \int_{a_{u^i}+\delta}^{t^*} p_i(x_k^i(t))\, dt = M_i - C(u^i). \end{aligned} \tag{2.31}$$

Hence (2.29) is impossible and x^i cannot meet ∂G on the interval I_{u^i}. Then it is clear that

$$C_i(u^i) = M_i, \tag{2.32}$$

and hence the pair (u^i, x^i) is the desired solution to the constrained optimization problem with cost functional $C_i(u)$ and this solution is clearly unconstrained.

3. Convergence results. We have demonstrated in the previous section the existence of an optimal solution for certain problems involving penalty functions and also the existence of a solution to the constrained optimization problem. Our next task will be to study the question whether, as i approaches infinity, the unconstrained optimal solution (u^i, x^i) to the problem with cost functional

$$C_i(u) = C(u) + \int_{a_u}^{b_u} p_i(x(t))\,dt$$

converges in any sense to a solution of the constrained problem with cost functional $C(u)$. Preparatory to this study we introduce the following definition.

DEFINITION 3.1. *Consider the constrained optimization problem. A pair* $(u, x) \in \Delta$ *is said to be approximable from the interior of* G *if there is a sequence* $\{(u_k, x_k)\}$ *of pairs belonging to* Δ *such that in each case* $x_k(t) \in \operatorname{int}(G)$ *for* $t \in I_{u_k}$ *and the sequence* $\{(u_k, x_k)\}$ *approximates* (u, x).

Then we have the following result.

THEOREM 3.1. *Let* $\{p_i(x)\}$ *be a sequence of penalty functions of the first kind for* G. *Let* (u^i, x^i) *be a solution to the unconstrained optimization problem with cost functional* $C_i(u)$ *given by* (2.11) *for each natural number* i. *Assume for each such* i *that*

$$\| x^i(t) \| \leqq B, \qquad t \in I_{u^i}, \tag{3.1}$$

where B *is a fixed positive number independent of* i. *If there is any solution* $(\hat{u}, \hat{x})$ *of the constrained optimization problem with cost functional* $C(u)$ *such that* $(\hat{u}, \hat{x})$ *is approximable from the interior of* G, *then there is a subsequence of* $\{(u^i, x^i)\}$ *which approximates a solution* $(\bar{u}, \bar{x})$ *to the constrained optimization problem with cost functional* $C(u)$.

Proof. Let D be a compact set containing $G \cap \{x \mid \| x \| \leqq B\}$ in its interior and such that $O \supset D$. The proof of Corollary 2.2 allows us to assume without loss of generality that $x_i(t) \in D$, $t \in I_{u^i}$, for all i. Then using Theorem 2.1 it is easy to see that there is a subsequence of $\{(u^i, x^i)\}$, which we shall continue to call $\{(u^i, x^i)\}$, and a pair $(\bar{u}, \bar{x}) \in \Delta$ (for the unconstrained problem) such that the sequence $\{(u^i, x^i)\}$ approximates $(\bar{u}, \bar{x})$.

We shall show that $(\bar{u}, \bar{x}) \in \Delta$ for the constrained problem, i.e., $\bar{x}(t) \in G$ for $t \in I_{\hat{u}}$. Suppose for contradiction this were not so. Let $t^* \in I_u$ be such that $\bar{x}(t^*) \in O - G$. Let D_0 be a compact subset of $O - G$ which contains a neighborhood of $\bar{x}(t^*)$. Then there is an interval $[t^* - \delta, t^* + \delta]$, for some $\delta > 0$, such that $\bar{x}(t) \in D_0$ for $t \in [t^* - \delta, t^* + \delta]$. Let D_1 be a compact subset of $O - G$ which contains D_0 in its interior. Since, if δ is sufficiently small, the sequence $\{x^i\}$ converges uniformly to $\bar{x}$ on $[t^* - \delta, t^* + \delta]$, for sufficiently large i, $x^i(t) \in D_1$ for $t \in [t^* - \delta, t^* + \delta]$. But then using (iii) of Definition 1.1 and the boundedness of $g^0(t, x)$ and $h_j^0(t, x)$, $j = 1, \cdots, m$, on bounded subsets of O, we see that

$$\lim_{i \to \infty} C_i(u^i) = +\infty. \tag{3.2}$$

We have assumed that $(\hat{u}, \hat{x})$ is a solution to the constrained optimization problem which is approximable from the interior. Hence there is a sequence of pairs $(\hat{u}_k, \hat{x}_k)$ approximating $(\hat{u}, \hat{x})$ such that $\hat{x}_k(t) \in \operatorname{int}(G)$, $t \in I_{\hat{u}_k}$. Consider first just the pair $(\hat{u}_1, \hat{x}_1)$. From the boundedness of $g^0(t, x)$ and $h_j^0(t, x)$, $j = 1, \cdots, m$, and (ii) of Definition 1.1 there is a fixed real number M such that

$$(3.3) \qquad C_i(\hat{u}_1) < M, \qquad i = 1, 2, \cdots.$$

But (3.2) together with (3.3) contradicts the optimality of the pair (u^i, x^i) for sufficiently large i. Hence $\bar{x}(t) \in G$ for $t \in I_{\bar{u}}$ and the pair $(\bar{u}, \bar{x}) \in \Delta$ for the constrained problem.

Finally, we must show that $(\bar{u}, \bar{x})$ is a solution of the constrained optimization problem. We have assumed that $(\hat{u}, \hat{x})$ is such a solution. Hence if we can show that

$$(3.4) \qquad C(\bar{u}) = C(\hat{u}),$$

the proof will be complete. Again proof is by contradiction; we assume

$$(3.5) \qquad C(\bar{u}) > C(\hat{u}),$$

and show that this leads to an absurdity.

We know that

$$(3.6) \qquad \lim_{i\to\infty} C(u^i) = C(\bar{u}).$$

Therefore

$$(3.7) \qquad \liminf_{i\to\infty} C_i(u^i) \geqq C(\bar{u})$$

by positivity of $p_i(x^i(t))$, $t \in I_{u^i}$. Thus if we let $d = C(\bar{u}) - C(\hat{u})$, there is a natural number i_0 such that

$$(3.8) \qquad C_i(u^i) > C(\hat{u}) + \frac{d}{2}, \qquad i \geqq i_0.$$

On the other hand we know that

$$(3.9) \qquad \lim_{k\to\infty} C(\hat{u}_k) = C(\hat{u}).$$

Let k_0 be chosen so large that for $k \geqq k_0$,

$$(3.10) \qquad | C(\hat{u}_k) - C(\hat{u}) | < \frac{d}{4}.$$

Now the trajectory $\hat{x}_{k_0}$ of the pair $(\hat{u}_{k_0}, \hat{x}_{k_0})$ lies entirely within some compact set $\hat{D}$ which lies entirely in $\operatorname{int}(G)$. Using (ii) of Definition 1.1, if i is sufficiently large,

$$(3.11) \qquad | C_i(\hat{u}_{k_0}) - C(\hat{u})| < \frac{d}{2}.$$

Together with (3.8) this implies

$$(3.12) \qquad C_i(u^i) > C_i(\hat{u}_{k_0}),$$

which contradicts the optimality of the pair (u^i, x^i) for the unconstrained optimization problem, which we have assumed. Hence we must conclude (3.5) false and (3.4) true and the proof of the theorem is complete.

Next we shall see that we can obtain similar results using a sequence of penalty functions of the second kind.

THEOREM 3.2. *Let the sequence $\{p_i(x)\}$ of penalty functions of the first kind in Theorem* 3.1 *be replaced by a sequence $\{p_i(x)\}$ of penalty functions of the second kind. Then the theorem remains true.*

Proof. Let us use the notation already introduced in the proof of Theorem 3.1. Theorem 2.1 again establishes the existence of a subsequence, which we shall still call $\{(u^i, x^i)\}$ which approximates a pair $(\bar{u}, \bar{x}) \in \Delta$. Corollary 2.3 assures us that for each i, $x^i(t) \in \text{int}(G)$ for $t \in I_{u^i}$, so we know that $\bar{x}(t) \in G$ for $t \in I_{\bar{u}}$; there is no need to prove it. The rest of the proof is word for word as in Theorem 3.1 except that we refer to (ii) of Definition 1.2 instead of (ii) of Definition 1.1. Thus we may regard the proof of this theorem as complete.

In order to show that the condition on approximability from the interior is necessary, let us consider the familiar system

$$(3.13) \qquad \ddot{x} = u, \qquad -1 \leqq u \leqq 1.$$

The time optimal trajectories of the two-dimensional first order system

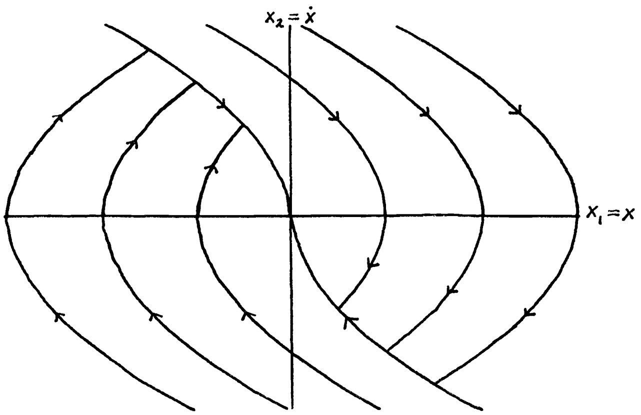


FIG. 1. *Time optimal trajectories for $\ddot{x} = u$*

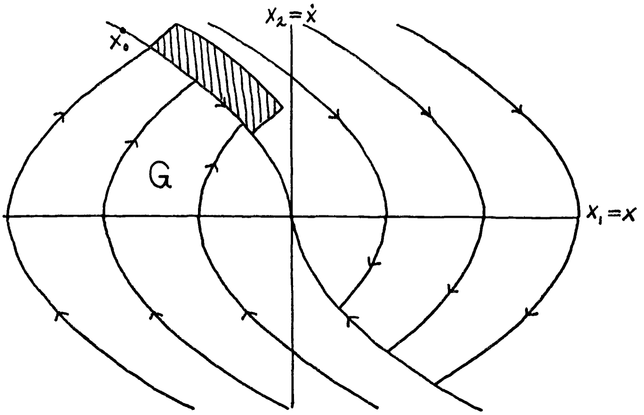


FIG. 2. *An optimal trajectory not approximable from the interior of G*

corresponding to (3.13) are shown in Fig. 1. In Fig. 2 we insist that trajectories be confined to the closed region G represented by the unshaded portion of the figure. While there is a time optimal trajectory $\bar{x}(t)$ from x_0 to 0 there are no nearby trajectories whatsoever lying entirely in $\operatorname{int}(G)$ with which to approximate $\bar{x}(t)$.

In order to make the results of this section complete one should give conditions on the system (1.2) and the domain G which are sufficient for a pair $(\hat{u}, \hat{x})$, wherein $\hat{x}$ lies in part on the boundary of G and otherwise in the interior of G, to be approximable by pairs $(\hat{u}_k, \hat{x}_k)$, wherein $\hat{x}_k$ lies entirely in the interior of G. At present this appears to be a difficult problem. We shall present here a simple result for linear autonomous systems.

THEOREM 3.3 *Consider the n-dimensional linear autonomous system*

$$\dot{x} = Ax + Bu, \tag{3.14}$$

where A and B are $n \times n$ and $n \times m$ matrices, respectively. Assume that u is restricted to a compact convex subset $\Omega \subset E^m$. We shall suppose the system (3.14) is proper, i.e., there is a vector $v \in \Omega$ such that the vectors Bv, ABv, $\cdots$, $A^{n-1}Bv$ are linearly independent.

Let G be a closed convex subset of E^n and let the target set $T(t)$ be identically the origin $x = 0$. Let $(\hat{u}, \hat{x})$ be a pair belonging to Δ such that $\hat{x}$ (possibly) lies on ∂G for certain subintervals of $I_{\hat{u}}$ and otherwise lies in $\operatorname{int}(G)$. *If there is any pair of functions (u^*, x^*) satisfying (1.2) and such that $u^*(t) \in \Omega$ for $t \in I_{\hat{u}}$ with $x^*(a_{\hat{u}}) = x_0$ and $x^*(t) \in \operatorname{int}(G)$ for $t \in I_{\hat{u}}$, then the pair $(\hat{u}, \hat{x})$ is approximable from the interior of G.*

Proof. For each natural number k, let u_k be defined on $I_{\hat{u}}$ by

$$u_k(t) = \left(1 - \frac{1}{k}\right)\hat{u}(t) + \frac{1}{k}u^*(t). \tag{3.15}$$

From the variation of parameters formula we have

$$x_k(t) = e^{A(t-a\hat{u})} x_0 + e^{A(t-a\hat{u})} \int_{a\hat{u}}^{t} e^{-A(s-a\hat{u})} \cdot B\left[\left(1-\frac{1}{k}\right)\hat{u}(s) + \frac{1}{k}u^*(s)\right] ds, \tag{3.16}$$

whence it is easily seen that

$$x_k(t) = \left(1-\frac{1}{k}\right)\hat{x}(t) + \frac{1}{k}x^*(t) \tag{3.17}$$

for $t \in I_{\hat{u}}$. Since $x^*(t) \in \operatorname{int}(G)$ for $t \in I_{\hat{u}}$, it is clear by the convexity of G that $x_k(t) \in \operatorname{int}(G)$ for $t \in I_{\hat{u}}$. Also

$$\lim_{k\to\infty} \left(\max_{t\in I\hat{u}} \| \hat{x}(t) - x_k(t) \|\right) = 0, \tag{3.18}$$

and

$$\lim_{k\to\infty} \left(\max_{t\in I\hat{u}} \| \hat{u}(t) - u_k(t) \|\right) = 0. \tag{3.19}$$

In particular, then,

$$\lim_{k\to\infty} x_k(b_{\hat{u}}) = 0. \tag{3.20}$$

The work of [4] shows that for k sufficiently large there is a pair of functions $(\tilde{u}_k, \tilde{x}_k)$ defined on an interval $[b_{\hat{u}}, b_{\hat{u}} + \delta_k]$ satisfying (1.2) and such that $\tilde{u}_k(t) \in \Omega$ for $t \in [b_{\hat{u}}, b_{\hat{u}} + \delta_k]$, with the properties

$$\tilde{x}_k(b_{\hat{u}}) = x_k(b_{\hat{u}}), \qquad \tilde{x}_k(b_{\hat{u}} + \delta_k) = 0. \tag{3.21}$$

Moreover

$$\lim_{k\to\infty} \delta_k = 0. \tag{3.22}$$

Since $0 \in \operatorname{int}(G)$ and $\dot{x}_k$ is uniformly bounded, if k is sufficiently large,

$$\tilde{x}_k(t) \in \operatorname{int}(G), \qquad t \in [b_{\hat{u}}, b_{\hat{u}} + \delta_k]. \tag{3.23}$$

Setting $I_{u_k} = [a_{\hat{u}}, b_{\hat{u}} + \delta_k]$ and

$$x_k(t) = \tilde{x}_k(t), \qquad u_k(t) = \tilde{u}_k(t), \qquad t \in [b_{\hat{u}}, b_{\hat{u}} + \delta_k], \tag{3.24}$$

for k sufficiently large, it is clear that the sequence of pairs (u_k, x_k) approximates $(\hat{u}, \hat{x})$ from the interior of G and the proof of our theorem is complete.

4. Concluding remarks. The use of penalty functions in the calculus of variations was introduced by R. Courant. Some of this material may be found in [5]. In the supplementary notes to Courant's *Calculus of Variations*, [6], the following theorem is proved by Martin Kruskal and Hanan Rubin.

THEOREM. *If*

(i) $\Phi(p)$ *and* $\Psi(p)$ *are lower semi-continuous real valued functions on a convergence space* S (*i.e., any space in which a notion of convergence is defined*);

(ii) $\Psi(p) \geqq 0$ *for all* p *in* S *and there exist points in* S *for which* $\Psi(p) = 0$;

(iii) A *denotes the problem: Find a point satisfying the side condition* $\Psi(p) = 0$, *at which* $\Phi(p)$ *takes on its least value for all* p *in* S *satisfying the side condition*;

(iv) A_t *denotes the problem: Find a point for which* $\Phi(p) + t\Psi(p)$ *takes on its least value for all* p *in* S; *and*

(v) *there exist a sequence* $\{t_n\}$ *of positive real numbers, a sequence* $\{p_n\}$ *of points in* S, *and a point* p_∞ *in* S *such that* $t_n \longrightarrow \infty$ *as* $n \longrightarrow \infty$, p_n *solves* A_{t_n} *and* $p_n \longrightarrow \infty$ *as* $n \longrightarrow \infty$;

then: p_∞ *solves* A.

A somewhat more general version of this theorem is presented by T. Butler and A. V. Martin in [7]. The counterpart to their theory is obtained in this paper by selecting a function $g(x)$ which vanishes on G and is positive outside G and then defining a series of penalty functions of the first kind by

$$p_i(x) = ig(x), \qquad i = 1, 2, \cdots. \tag{4.1}$$

This method has been used by S. S. L. Chang in [2] and [3] to obtain necessary conditions which must be obeyed by a solution $(\bar{u}, \bar{x})$ of the constrained optimization problem. Chang's theory is not confined to systems (1.2) and cost functionals (1.3) which are linear in u. We have confined our work to systems and cost functionals linear in u so that we may use the work of Lee and Markus [4]. It is interesting to note that when a sequence of penalty functions of the type (4.1) is used, the requirement concerning approximability from the interior is not necessary. One would like to know for what general class of sequences of penalty functions of the first kind this is true.

Penalty functions of the second kind do not appear to have received much attention in the literature. For certain finite dimensional programming problems a technique similar to this has been used with some success by A. V. Fiacco and G. P. McCormick in [8]. Apparently such methods offer certain advantages in numerical computation. We would like to point out two advantages enjoyed by a procedure which uses penalty functions of the second kind. First of all, each of the approximating pairs (u^i, x^i) is such that x^i does not violate the phase constraints even though it is an unconstrained solution to the augmented problem. Thus if the boundary is truly "hard", i.e., must not be violated at all, and an approximate solution to the optimization problem is desired, penalty functions of the second kind provide a suitable method of approximation. Second, it could conceivably happen that either the system (1.2) or the cost functional (1.3) is undefined outside of G in which case again interior approximation is mandatory.

An important question is the rate of convergence of the x^i to $\bar{x}$ in terms of the given sequence of penalty functions. Such error estimates are what is needed to provide a rigorous demonstration of the results obtained in a somewhat formal manner by Chang in [2] and [3]. This is important because Chang's results appear to be the most useful yet obtained for the constrained phase optimization problem.

REFERENCES

[1] R. V. Gamkrelidze, *Optimal processes with restricted phase coordinates*, The Mathematical Theory of Optimal Processes, L. S. Pontryagin et al, Interscience, New York, 1962, Chap. VI.

[2] S. S. L. Chang, *Optimal control in bounded phase space*, Automatica, vol. 1, Pergamon Press, New York, 1962, pp. 55–67.

[3] ———, *An extension of Ascoli's theorem and its applications to the theory of optimal control*, AFOSR Report 1973, 1962.

[4] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[5] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.

[6] R. Courant, *Calculus of variations and supplementary notes and exercises, 1945–1946*, revised and amended by J. Moser, New York University Institute of Mathematical Sciences, New York, 1956–1957 (mimeographed lecture notes.)

[7] T. Butler and A. V. Martin, *On a method of Courant for minimizing functionals*, J. Math. and Phys., 41 (1962), pp. 291-299.

[8] A. V. Fiacco and G. P. McCormick, *The sequential unconstrained minimization technique for nonlinear programing, a primal-dual method*, Management Sci., 10 (1964), pp. 360–366.

J.SIAM CONTROL
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

OPTIMAL-THRUST TRAJECTORIES IN AN ARBITRARY GRAVITATIONAL FIELD*

JOSEPH G. GURLEY†

Abstract. The problem of optimal-thrust trajectories is studied using a slight variation of the usual calculus of variations technique. The results include the usual first-order criteria for optimality, which are that the direction of thrust must be everywhere parallel to a solution $\boldsymbol{\psi}$ of the adjoint differential equation, and that the magnitude of the thrust must be zero in regions where the magnitude of $\boldsymbol{\psi}$ is less than a critical value, and equal to the maximum permissible value in regions where the magnitude of $\boldsymbol{\psi}$ is greater than the critical value. Singular arcs, on which the magnitude of $\boldsymbol{\psi}$ is continuously equal to the critical value, are shown to exist in the case of all except the simplest gravitational fields, and in some cases may form part of an optimal trajectory. A means of calculating the unique value of thrust required to sustain a singular arc is described, and a test for the optimality of such arcs is given. The test shows that a family of singular arcs discovered by D. F. Lawden is nonoptimal.

Introduction. Optimal trajectories of thrusting vehicles in a gravitational field have been extensively studied in recent years [1]–[5]. This particular problem, in which the trajectories are characterized by a very high degree of predictability, is particularly appropriate for the newly developed theories of optimal control [6], [7]. In this paper, a related theory is developed which includes second-order variations as well as the usual linear variations. This makes it possible to exhibit the characteristics of the singular solutions and to test for their optimality. The test for optimality is basically similar to one proposed by Kelly [8], modified so as to apply to an acceleration-controlled system rather than a velocity-controlled system.

Background. The motion of a vehicle acted on by a gravitational acceleration $\mathbf{g}(\mathbf{r}, t)$ and by a propulsive force or thrust $\mathbf{F}$ whose magnitude is proportional to the rate of fuel consumption $-\dot{m}$, is described by the equations

$$\begin{aligned} \ddot{\mathbf{r}} - \mathbf{g} - m^{-1}\mathbf{F} &= 0, \\ c\dot{m} + |\,\mathbf{F}\,| &= 0, \end{aligned} \tag{1}$$

where m is the mass of vehicle plus remaining fuel and c is the *characteristic velocity* of the rocket-type thrusting engine. There generally exists a constraint on the maximum thrust and on the minimum weight of vehicle plus fuel (the latter can not be less than the weight m_v of the vehicle alone).

* Received by the editors July 6, 1964, and in revised form October 30, 1964.

† Hughes Aircraft Company, Post Office Box 90919, Airport Station, Los Angeles, California.

$$(2) \qquad \begin{aligned} | \mathbf{F} | &\leqq F_{\max}, \\ m &\geqq m_v . \end{aligned}$$

We define admissible trajectories as those trajectories satisfying (1) and (2) which originate at a fixed time t_1, with fixed position $\mathbf{r}_1$, velocity $\dot{\mathbf{r}}_1$, and mass m_1, which terminate at a fixed time $t_2 > t_1$, and which lie entirely within a closed region S where $\mathbf{g}$ is nonsingular. The boundaries of S, which may be time-dependent, serve to exclude trajectories which encounter celestial bodies. Admissible trajectories exist except in cases where the propulsion limitations (2) make it impossible to remain within S until the final time t_2.

The values of the final mass m_2 are bounded; the maximum is less than m_1 by the minimum amount of fuel, if any, which must be expended in order to remain in S until time t_2, and the minimum is the larger of the two quantities m_v (corresponding to total exhaustion of the fuel supply) and $m_1 - (t_2 - t_1)c^{-1}F_{\max}$ (corresponding to continuous consumption of fuel at the maximum rate $c^{-1}F_{\max}$, from time t_1 to time t_2). Each value of m_2 between these limits determines a region $A(m_2)$ of final position-velocity space $(\mathbf{r}_2, \dot{\mathbf{r}}_2)$ which is accessible with that particular expenditure of fuel. The accessible region $A(m_2)$ expands as m_2 decreases, attaining its maximum extent A^* when m_2 is equal to its minimum value. Within the region A^* there is associated with every point $(\mathbf{r}_2, \dot{\mathbf{r}}_2)$ a maximum value of m_2 and a value

$$(3) \qquad J(\mathbf{r}_2, \dot{\mathbf{r}}_2) = \max \left[c \log \frac{m_2}{m_1} \right].$$

A trajectory which terminates at a point $(\mathbf{r}_2, \dot{\mathbf{r}}_2)$ within A^*, and for which $c \log (m_2/m_1)$ assumes its maximum value $J(\mathbf{r}_2, \dot{\mathbf{r}}_2)$, is said to be *optimal*

Conditions for optimality. On subtracting linear integral functionals of (1) from

$$c \log \frac{m_2}{m_1} \leqq J(\mathbf{r}_2, \dot{\mathbf{r}}_2),$$

one obtains

$$c \log \frac{m_2}{m_1} - \int_{t_1}^{t_2} [\boldsymbol{\psi} \cdot (\ddot{\mathbf{r}} - \mathbf{g} - m^{-1}\mathbf{F}) + m^{-1}\phi(c\dot{m} + | \mathbf{F} |)] \, dt \leqq J(\mathbf{r}_2, \dot{\mathbf{r}}_2);$$

or, after partial integration, assuming that $\boldsymbol{\psi}$ can be differentiated at least twice and ϕ at least once,

$$
\begin{aligned}
&(1 - \phi_2)c \log \frac{m_2}{m_1} - \boldsymbol{\psi}_2 \cdot \dot{\mathbf{r}}_2 + \dot{\boldsymbol{\psi}}_2 \cdot \mathbf{r}_2 + \boldsymbol{\psi}_1 \cdot \dot{\mathbf{r}}_1 - \dot{\boldsymbol{\psi}}_1 \cdot \mathbf{r}_1 \\
&\qquad + \int_{t_1}^{t_2} [m^{-1}(\boldsymbol{\psi} \cdot \mathbf{F} - \phi \mid \mathbf{F} \mid) \\
&\qquad + c\dot{\phi} \log \frac{m}{m_2} + \boldsymbol{\psi} \cdot \mathbf{g} - \ddot{\boldsymbol{\psi}} \cdot \mathbf{r}]\, dt \leqq J(\mathbf{r}_2 , \dot{\mathbf{r}}_2),
\end{aligned} \tag{4}
$$

where subscripts 1 and 2 refer to initial and final conditions, respectively.

Let $\mathbf{r}(t)$ be an optimal trajectory generated by the thrust $\mathbf{F}(t)$, with $m(t)$ the resulting mass; and let $\mathbf{R}(\epsilon, t)$ be the parameterized family of trajectories generated by the thrust $\mathbf{F}(t) + \epsilon\delta\mathbf{F}(t)$, with $M(\epsilon, t)$ the resulting mass. We consider the case where the trajectory lies entirely in the interior of the region S, and where $\delta\mathbf{F}(t)$ is such that, for some positive ϵ_0 and $0 \leqq \epsilon \leqq \epsilon_0$, each member of the family $\mathbf{R}(\epsilon, t)$ is admissible.

In the case of the optimal trajectory $\mathbf{r}(t)$, the equality sign applies in (4), but in the case of the family $\mathbf{R}(\epsilon, t)$ the weak inequality applies as shown. On taking the difference, one obtains

$$
\begin{aligned}
&(1 - \phi_2)c \log \frac{M_2}{m_2} - \boldsymbol{\psi}_2 \cdot (\dot{\mathbf{R}}_2 - \dot{\mathbf{r}}_2) + \dot{\boldsymbol{\psi}}_2 \cdot (\mathbf{R}_2 - \mathbf{r}_2) \\
&\quad + \int_{t_1}^{t_2} \Big\{ M^{-1}[\boldsymbol{\psi} \cdot (\mathbf{F} + \epsilon\delta\mathbf{F}) - \phi \mid \mathbf{F} + \epsilon\delta\mathbf{F} \mid] - m^{-1}(\boldsymbol{\psi} \cdot \mathbf{F} - \phi \mid \mathbf{F} \mid) \\
&+ c\dot{\phi} \log \frac{M}{m} + \boldsymbol{\psi} \cdot [\mathbf{g}(\mathbf{R}) - \mathbf{g}(\mathbf{r})] - \ddot{\boldsymbol{\psi}} \cdot (\mathbf{R} - \mathbf{r}) \Big\} dt \leqq J(\mathbf{R}_2 , \dot{\mathbf{R}}_2) - J(\mathbf{r}_2 , \dot{\mathbf{r}}_2).
\end{aligned}
$$

For a sufficiently small value of ϵ_0, the left side of this equation can be approximated arbitrarily closely by a power series in ϵ. For points in A^* which can be reached by optimal trajectories in the neighborhood of $r(t)$, the right side can also be expanded in powers of ϵ. These expansions give

$$
\begin{aligned}
&\epsilon \Big\{ (1 - \phi_2)cm_2^{-1} \frac{\partial M_2}{\partial \epsilon} - (\boldsymbol{\psi}_2 + \boldsymbol{\nabla}_2{}' J_2) \cdot \frac{\partial \dot{\mathbf{R}}_2}{\partial \epsilon} + (\dot{\boldsymbol{\psi}}_2 - \boldsymbol{\nabla}_2 J) \cdot \frac{\partial \mathbf{R}_2}{\partial \epsilon} \\
&+ \epsilon \int_{t_1}^{t_2} \Big\{ -\frac{\partial \mathbf{R}}{\partial \epsilon} \cdot [\ddot{\boldsymbol{\psi}} - \boldsymbol{\nabla}(\boldsymbol{\psi} \cdot \mathbf{g})] + m^{-2} \frac{\partial M}{\partial \epsilon} [mc\dot{\phi} - \boldsymbol{\psi} \cdot \mathbf{F} + \phi \mid \mathbf{F} \mid] \\
&+ m^{-1}[\boldsymbol{\psi} \cdot \delta\mathbf{F} - \phi\delta \mid \mathbf{F} \mid] \Big\} dt + \frac{\epsilon^2}{2} \Big\{ (1 - \phi_2)cm_2^{-1} \Big[\frac{\partial^2 M_2}{\partial \epsilon^2} \\
&- m_2^{-1} \Big(\frac{\partial M_2}{\partial \epsilon} \Big)^2 \Big] - (\boldsymbol{\psi}_2 + \boldsymbol{\nabla}_2{}' J_2) \cdot \frac{\partial^2 \dot{\mathbf{R}}_2}{\partial \epsilon^2} + (\dot{\boldsymbol{\psi}}_2 - \boldsymbol{\nabla}_2 J) \cdot \frac{\partial^2 \mathbf{R}_2}{\partial \epsilon^2} \\
&- \Big(\frac{\partial \mathbf{R}_2}{\partial \epsilon} \cdot \boldsymbol{\nabla}_2 + \frac{\partial \dot{\mathbf{R}}_2}{\partial \epsilon} \cdot \boldsymbol{\Delta}_2{}' \Big)^2 J_2 \Big\} + \frac{\epsilon^2}{2} \int_{t_1}^{t_2} \Big\{ -\frac{\partial^2 \mathbf{R}}{\partial \epsilon^2} \cdot [\ddot{\boldsymbol{\psi}} - \boldsymbol{\nabla}(\boldsymbol{\psi} \cdot \mathbf{g})]
\end{aligned} \tag{5}
$$

$$+ m^{-2} \frac{\partial^2 M}{\partial \epsilon^2} [mc\dot{\phi} - \boldsymbol{\psi}\cdot\mathbf{F} + \phi \mid \mathbf{F} \mid] - m^{-3} \left(\frac{\partial M}{\partial \epsilon}\right)^2 [mc\dot{\phi} - 2\boldsymbol{\psi}\cdot\mathbf{F}$$

$$+ 2\phi \mid \mathbf{F} \mid] - 2m^{-2} \frac{\partial M}{\partial \epsilon} [\boldsymbol{\psi}\cdot\delta\mathbf{F} - \phi\delta \mid \mathbf{F} \mid] - m^{-1}\phi\delta^2 \mid \mathbf{F} \mid$$

$$+ \left(\frac{\partial \mathbf{R}}{\partial \epsilon}\cdot\nabla\right)^2 (\boldsymbol{\psi}\cdot\mathbf{g})\Bigg\} \, dt + \text{terms of order } \epsilon^3 \text{ and higher} \leqq 0,$$

where it is assumed that $\mid \mathbf{F} + \epsilon\delta\mathbf{F} \mid$ can be approximated by an expression of the form $\mid \mathbf{F} \mid + \epsilon\delta \mid \mathbf{F} \mid + (\epsilon^2/2)\delta^2 \mid \mathbf{F} \mid +$ terms of order ϵ^3 and higher. The symbols ∇, ∇_2, and ∇_2' represent the gradient operators $\partial/\partial\mathbf{R}$, $\partial/\partial\mathbf{R}_2$, and $\partial/\partial\dot{\mathbf{R}}_2$ respectively. The derivatives of $\mathbf{R}$, M, $\mathbf{R}_2$, $\dot{\mathbf{R}}_2$, and M_2 with respect to ϵ are evaluated for $\epsilon = 0$.

The terms in (5) which are linear in ϵ and proportional to $\delta\mathbf{r}(t)$, $\delta m(t)$, $\delta\mathbf{r}_2$, $\delta\dot{\mathbf{r}}_2$, and δm_2 can be eliminated by letting $\boldsymbol{\psi}$ and ϕ be the solution of the adjoint differential equations

$$\begin{aligned} \ddot{\boldsymbol{\psi}} - \nabla(\boldsymbol{\psi}\cdot\mathbf{g}) &= 0, \\ mc\dot{\phi} - \boldsymbol{\psi}\cdot\mathbf{F} + \phi \mid \mathbf{F} \mid &= 0, \end{aligned} \tag{6}$$

satisfying the boundary conditions

$$\begin{aligned} \boldsymbol{\psi}_2 &= -\nabla_2' J, \\ \dot{\boldsymbol{\psi}}_2 &= \nabla_2 J, \\ \phi_2 &= 1. \end{aligned} \tag{7}$$

In this case, (5) becomes

$$\begin{aligned} &\epsilon \int_{t_1}^{t_2} \{m^{-1}[\boldsymbol{\psi}\cdot\delta\mathbf{F} - \mid \boldsymbol{\psi} \mid \delta \mid \mathbf{F} \mid] + m^{-1}[\mid \boldsymbol{\psi} \mid - \phi]\delta \mid \mathbf{F} \mid\} \, dt \\ &- \frac{\epsilon^2}{2} \left(\frac{\partial \mathbf{R}_2}{\partial \epsilon}\cdot\nabla_2 + \frac{\partial \dot{\mathbf{R}}_2}{\partial \epsilon}\cdot\nabla_2'\right)^2 J_2 + \frac{\epsilon^2}{2} \int_{t_1}^{t_2} \Bigg\{ m^{-2} \left(\frac{\partial M}{\partial \epsilon}\right)^2 c\dot{\phi} \\ &- 2m^{-2} \frac{\partial M}{\partial \epsilon} [\boldsymbol{\psi}\cdot\delta\mathbf{F} - \phi\delta \mid \mathbf{F} \mid] - m^{-1}\phi\delta^2 \mid \mathbf{F} \mid \\ &\qquad + \left(\frac{\partial \mathbf{R}}{\partial \epsilon}\cdot\nabla\right)^2 (\boldsymbol{\psi}\cdot\mathbf{g})\Bigg\} \, dt + \text{terms of order } \epsilon^3 \text{ and higher} \leqq 0. \end{aligned} \tag{8}$$

Since (8) is valid for any variations in thrust which do not violate the conditions imposed by (2), it is valid for a variation which rotates the thrust towards $\boldsymbol{\psi}$, without changing the magnitude of the thrust, in some region where the thrust is not parallel to $\boldsymbol{\psi}$. But such a variation makes the integrand of the first term of (8) positive, and therefore (8) is violated for ϵ sufficiently small. Consequently, there can be no region in which the thrust is not parallel to $\boldsymbol{\psi}$, in the interval $t_1 \leqq t \leqq t_2$.

Equation (8) is also violated if, assuming $\mathbf{F}$ and $\boldsymbol{\psi}$ parallel, the magnitude of $\mathbf{F}$ can be increased in a region where $|\boldsymbol{\psi}| - \phi$ is positive or decreased in a region where $|\boldsymbol{\psi}| - \phi$ is negative. Consequently, $|\mathbf{F}|$ must be equal to $F_{\max}$ throughout every region where $|\boldsymbol{\psi}|$ exceeds ϕ, and be zero throughout every region where $|\boldsymbol{\psi}|$ is less than ϕ.

The preceding rules can be summarized as follows:

$$\mathbf{F} = \begin{cases} F_{\max}\, |\boldsymbol{\psi}|^{-1}\boldsymbol{\psi}, & \text{where } |\boldsymbol{\psi}| > \phi; \\ |\mathbf{F}|\, |\boldsymbol{\psi}|^{-1}\boldsymbol{\psi}, \quad 0 \leqq |\mathbf{F}| \leqq F_{\max}, & \text{where } |\boldsymbol{\psi}| = \phi; \\ 0, & \text{where } |\boldsymbol{\psi}| < \phi. \end{cases} \tag{9}$$

These are the usual necessary conditions for the optimality of the trajectory $\mathbf{r}(t)$.

Singular arcs. The first-order optimality conditions (9) are defective in that they fail to specify completely the value or values of thrust $\mathbf{F}$ required to continue an optimal trajectory, once one has reached a *singular arc* of finite length where $|\boldsymbol{\psi}|$ is equal to ϕ throughout. To attack this problem, we will repeatedly differentiate $|\boldsymbol{\psi}|^2$ with respect to t, noting that on a singular arc, the second line of (9) and the second line of (6) imply that $\dot{\phi}$ vanishes. Therefore $|\boldsymbol{\psi}|$ is constant and successive differentiation of $\frac{1}{2}|\boldsymbol{\psi}|^2$ gives

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}|\boldsymbol{\psi}|^2 &= \boldsymbol{\psi}\cdot\dot{\boldsymbol{\psi}} = 0, \\ \frac{1}{2}\frac{d^2}{dt^2}|\boldsymbol{\psi}|^2 &= \boldsymbol{\psi}\cdot\ddot{\boldsymbol{\psi}} + |\dot{\boldsymbol{\psi}}|^2 = (\boldsymbol{\psi}\cdot\nabla)(\boldsymbol{\psi}\cdot\mathbf{g}) + |\dot{\boldsymbol{\psi}}|^2 = 0, \\ \frac{1}{2}\frac{d^3}{dt^3}|\boldsymbol{\psi}|^2 &= (\dot{\mathbf{r}}\cdot\nabla)(\boldsymbol{\psi}\cdot\nabla)(\boldsymbol{\psi}\cdot\mathbf{g}) + (\boldsymbol{\psi}\cdot\nabla)\left(\boldsymbol{\psi}\cdot\frac{\partial\mathbf{g}}{\partial t}\right) \\ &\qquad + 4(\dot{\boldsymbol{\psi}}\cdot\nabla)(\boldsymbol{\psi}\cdot\mathbf{g}) = 0, \\ \frac{1}{2}\frac{d^4}{dt^4}|\boldsymbol{\psi}|^2 &= [(\mathbf{g} + m^{-1}|\boldsymbol{\psi}|^{-1}|\mathbf{F}|\boldsymbol{\psi})\cdot\nabla](\boldsymbol{\psi}\cdot\nabla)(\boldsymbol{\psi}\cdot\mathbf{g}) \\ &\qquad + (\dot{\mathbf{r}}\cdot\nabla)\frac{d}{dt}[(\boldsymbol{\psi}\cdot\nabla)(\boldsymbol{\psi}\cdot\mathbf{g})] + \frac{d}{dt}\left[(\boldsymbol{\psi}\cdot\nabla)\left(\boldsymbol{\psi}\cdot\frac{\partial\mathbf{g}}{\partial t}\right)\right. \\ &\qquad \left. + 4(\dot{\boldsymbol{\psi}}\cdot\nabla)(\boldsymbol{\psi}\cdot\mathbf{g})\right] = 0. \end{aligned} \tag{10}$$

The second form of the second line follows from (5), and the last line makes use of the expression for $\ddot{\mathbf{r}}$ from (1). The forms of the third and fourth lines of this equation have been simplified by taking advantage of the conservative property of the gravitational field, but the conclusions drawn are applicable to a broader class of force fields.

Necessary conditions for a singular arc are, that the last equation of (10) be satisfied everywhere on the arc, and that the first three conditions all be satisfied at some one point. In this case the first three conditions, being integrals of the last condition, are necessarily satisfied everywhere on the arc.

The conditions for a singular arc can be rephrased as follows: the first three conditions must be satisfied by $\mathbf{r}$, $\dot{\mathbf{r}}$, $\boldsymbol{\psi}$, and $\dot{\boldsymbol{\psi}}$ at the point where the arc is initiated, and the fourth condition must have a root $|\mathbf{F}| = F_0$ lying in the range of permissible thrust magnitude $0 \leqq F_0 \leqq F_{\max}$. If the coefficient of $|\mathbf{F}|$, which is $(\boldsymbol{\psi}\cdot\nabla)^2(\boldsymbol{\psi}\cdot\mathbf{g})$, does not vanish, then there is a unique root and therefore a unique continuation of the singular arc.

The actual existence of singular arcs can be demonstrated constructively for the inverse-square-law central force field, for any positive value of $F_{\max}$. It appears that singular arcs exist in most problems of this class, provided $\mathbf{g}$ is a nonlinear function of position.

An optimality test for singular arcs. If the last equation of (10) prescribes a value of $|\mathbf{F}|$ which is intermediate between 0 and $F_{\max}$, then the optimality of the resulting singular arc can be tested by means of a thrust variation

$$\epsilon\delta\mathbf{F} = \epsilon\,|\boldsymbol{\psi}|^{-1}\boldsymbol{\psi}A(t)m(t),$$

$$A(t) = \begin{cases} -a & \text{if } \; 0 \leqq |t - t_0| < \tau, \\ +a & \text{if } \; \tau \leqq |t - t_0| < 2\tau, \\ 0 & \text{if } \; 2\tau \leqq |t - t_0|. \end{cases} \tag{11}$$

For small enough values of ϵ and τ, this variation results in a thrust consistent with (2), provided the reference trajectory does not touch the boundary of S.

The associated variation in $\mathbf{r}(t)$, according to (1), is

$$\epsilon\delta\mathbf{r}(\epsilon, t) = \epsilon\,|\boldsymbol{\psi}|^{-1}\boldsymbol{\psi}\int_{t_0}^{t}(t - t')A(t')\,dt' \tag{12}$$

$$+ \text{ terms of order } \epsilon a\tau^4 \text{ and higher.}$$

The leading term, of order $\epsilon a\tau^2$, is linear in ϵ and vanishes outside the interval $t_0 - 2\tau < t < t_0 + 2\tau$ (Fig. 1b). Therefore, on substituting (12) into (8), one obtains

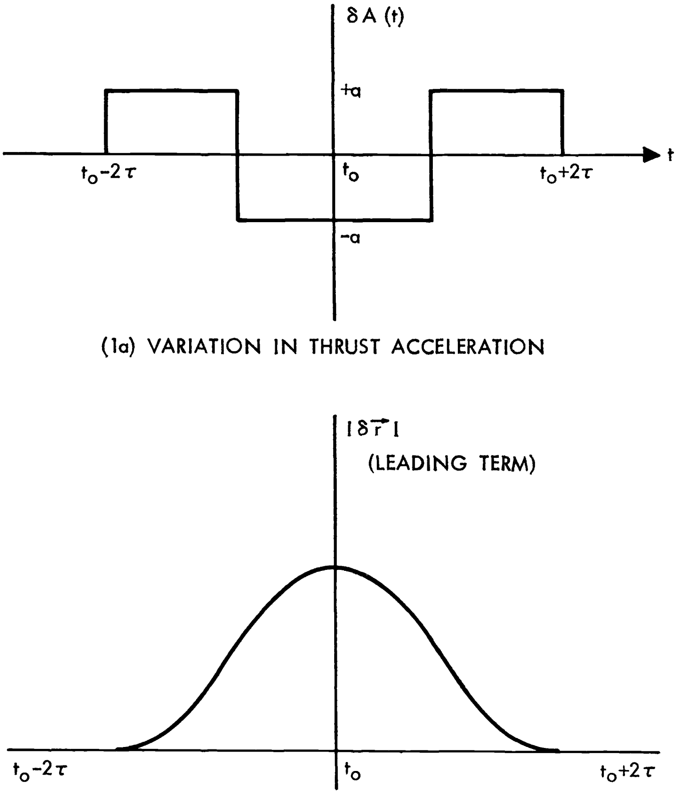
$$\frac{\epsilon^2}{2} \,|\, \boldsymbol{\psi} \,|^{-2} \int_{t_1}^{t_2} \left[\int_{t_1}^{t} (t - t')A(t')\, dt' \right]^2 (\boldsymbol{\psi}\cdot\nabla)^2(\boldsymbol{\psi}\cdot\mathbf{g})\, dt \tag{13}$$
$$+ \text{ terms of order } \epsilon^2 a^2 \tau^7 \text{ and higher } \leqq 0.$$

Since the leading term is of order $\epsilon^2 a^2 \tau^5$, this condition is violated for small enough values of τ, unless

$$(\boldsymbol{\psi}\cdot\nabla)^2(\boldsymbol{\psi}\cdot\mathbf{g}) \leqq 0. \tag{14}$$

This inequality is a necessary condition for the optimality of a singular arc.

Assume that the first three constraints of (10) and $|\,\boldsymbol{\psi}\,| = \phi$ are satisfied at a particular point on an optimal trajectory. Then the final constraint of



(1a) VARIATION IN THRUST ACCELERATION

(1b) VARIATION IN POSITION (THE DIRECTION OF $\delta\vec{r}$ IS PARALLEL TO $\vec{\psi}$)

Fig. 1. *Variation used to test the optimality of a singular arc*

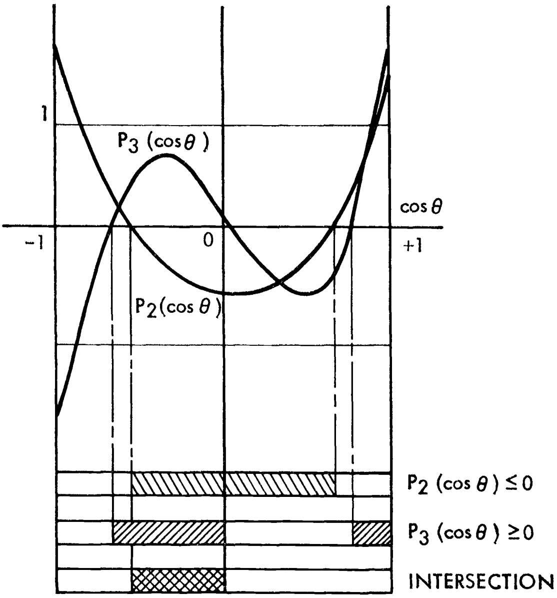


FIG. 2. *Intersection of the inequalities* P_2 *(cos θ)* $\leqq 0$ *and* P_3 *(cos θ)* $\geqq 0$

(10) can be written, if $(\boldsymbol{\psi}\cdot\nabla)^2(\boldsymbol{\psi}\cdot\mathbf{g})$ is not zero, in the form

$$m \mid \boldsymbol{\psi} \mid \frac{d^4}{dt^4} \mid \boldsymbol{\psi} \mid = [(\boldsymbol{\psi}\cdot\nabla)^2(\boldsymbol{\psi}\cdot\mathbf{g})](\mid \mathbf{F} \mid - F_0). \tag{15}$$

If the strong inequality is satisfied in (14), then the fourth derivative of $\mid \boldsymbol{\psi} \mid$, the lowest derivative not zero by hypothesis, is opposite in sign to $\mid \mathbf{F} \mid - F_0$. Therefore the optimal trajectory can continue with $\mid \mathbf{F} \mid$ equal to $F_{\max}$ only if $\mid \mathbf{F} \mid - F_0$ is nonpositive, so that $\mid \boldsymbol{\psi} \mid$ will not immediately begin to decrease. This can occur only if F_0 equals or exceeds $F_{\max}$. Similarly, the optimal trajectory can continue with $\mid \mathbf{F} \mid$ equal to zero only if $\mid \mathbf{F} \mid - F_0$ is nonnegative, so that $\mid \boldsymbol{\psi} \mid$ will not immediately begin to increase. This can occur only if F_0 is less than or equal to zero. To summarize, the strong inequality sign in (14) requires one to continue an optimal trajectory along the singular arc, provided the thrust required to sustain that arc lies in the permissible range $0 \leqq \mid \mathbf{F} \mid \leqq F_{\max}$.

An optimal singular arc can terminate in the interior of the interval $t_1 \leqq t \leqq t_2$, if F_0, which is a continuous function of time when the strong inequality (14) applies, exceeds $F_{\max}$ or becomes negative. In the former case, the singular arc terminates in a maximum-thrust arc, in the latter case in a zero-thrust arc. The general rule is, that $\mid \mathbf{F} \mid$ is continuous at all interior and boundary points of a singular arc which forms part of an optimal trajectory, except at points where $(\boldsymbol{\psi}\cdot\nabla)^2(\boldsymbol{\psi}\cdot\mathbf{g})$ is zero.

If (14) is violated, then the fourth derivative of $| \boldsymbol{\psi} |$ has the same sign as $| \mathbf{F} | - F_0$. In this case, the optimal trajectory can continue with $| \mathbf{F} |$ equal to $F_{\max}$ if $| \mathbf{F} | - F_0$ is positive, or if F_0 has any value less than $F_{\max}$. Similarly, the optimal trajectory can continue with $| \mathbf{F} |$ equal to zero if $| \mathbf{F} | - F_0$ is negative, or if F_0 has any positive value. If F_0 lies in the range $0 < F_0 < F_{\max}$, there are two optimal trajectories proceeding from the same set of initial conditions, and going off in quite different directions, one using maximum thrust and one using zero thrust.

Application to a central force. For an inverse-square-law central force

$$\mathbf{g} = \nabla \left(\frac{\mu}{r} \right), \tag{16}$$

where μ is a constant and r is the magnitude of the position vector $\mathbf{r}$ measured relative to the center of the field, then the second line of (10) shows $(\boldsymbol{\psi} \cdot \nabla)(\boldsymbol{\psi} \cdot \mathbf{g})$ to be negative and (14) shows $(\boldsymbol{\psi} \cdot \nabla)^2 (\boldsymbol{\psi} \cdot \mathbf{g})$ to be negative. These two conditions lead to

$$\begin{aligned} P_2(\cos \theta) &\leqq 0, \\ P_3(\cos \theta) &\geqq 0, \end{aligned} \tag{17}$$

where P_n is the Legendre polynomial of order n, and θ is the angle between the thrust vector $\mathbf{F}$ and the radius vector $\mathbf{r}$. The intersection of these two conditions (Fig. 2) gives, as a necessary condition for the existence of an optimal singular arc,

$$-(\tfrac{1}{3})^{1/2} \leqq \cos \theta \leqq 0. \tag{18}$$

Lawden [3] gives equations for all planar singular trajectories for the inverse-square-law central force field, and discusses in some detail a family of spiral planetary escape or planetary approach trajectories. The latter involve thrust in a direction such that $\cos \theta$ is positive, and so fail to satisfy the necessary condition for optimality of a singular arc. Hence no segment of these spirals can be part of an optimal trajectory.

REFERENCES

[1] J. V. Breakwell, *The optimization of trajectories*, J. Soc. Indust. Appl. Math., 7 (1959), pp. 215–247.

[2] G. Leitmann, *On a class of variational problems in rocket flight*, J. Aerospace Sci., 26 (1959), pp. 586–591.

[3] D. F. Lawden, *Optimal intermediate-thrust arcs in a gravitational field*, Astronaut. Acta, 8 (1962), pp. 106–123.

[4] P. Contensou, *Etude théoretique des trajectoires optimale dans un champ de gravitation*, Astronaut. Acta, 8 (1962), pp. 134–149.

[5] S. Pines, *Constants of the motion for optimal thrust trajectories in a central force*

field, Joint AIAA-IMS-SIAM-ONR Symposium on Control and System Optimization, Monterey, California, January 27–29, 1964.

[6] L. S. Pontryagin et al, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[7] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, 1957.

[8] H. J. Kelley, *A second variation test for singular extremals*, Joint AIAA-IMS-SIAM-ONR Symposium on Control and System Optimization, Monterey, California, January 27–29, 1964.

J.SIAM Control
Ser. A, Vol. 2, No. 3
Printed in U.S.A., 1965

MINIMAX CONTROL OF DISCRETE TIME STOCHASTIC SYSTEMS*

D. D. SWORDER†

1. Introduction. In this paper the synthesis of a control policy for an object with stochastic elements will be investigated. The randomness associated with the object to be controlled can come about in several different ways. For example, the control rule to be used may depend explicitly on output measurements from the system which are contaminated with additive random noise. On the other hand, it might be that some of the parameters which are contained in the equations describing the process are random variables. If the criterion of performance is a nonnegative functional of the system state and of the control policy, one might hope to choose the control in such a way that the expected value of this functional is minimized. In the case where certain parameters of the process are incompletely specified, this leads to the conceptual problem that an optimal control rule may be a function of these undetermined parameters. Since the performance index now provides only a partial ordering of control policies, an auxiliary criterion must be chosen to provide the designer with a "best" control.

In what follows we will make extensive use of the definitions and results from the theory of games as presented by Blackwell and Girshick [1]. It will be shown that the basic structural properties of the control problem can be formulated within the framework provided by this theory.

2. Mathematical formulation. To make the ideas of §1 more precise, consider the discrete time system described by the equation

$$\begin{aligned} x_{j+1} &= f_j(x_j, v_j, \xi_j), \qquad 0 \leqq j \leqq N - 1, \\ x_0 &= x(0), \end{aligned} \tag{1}$$

where:

x_j = the n-dimensional state vector at time $t = j\Delta$. Δ is the unit increment of time.

v_j = the k-dimensional control vector at time $t = j\Delta$.

ξ_j = the r-dimensional disturbance vector at time $t = j\Delta$.

f_j is continuous for all j in the interval $0 \leqq j \leqq N - 1$.

* Received by the editors April 20, 1964, and in final revised form December 23, 1964.

† Department of Electrical Engineering, University of Southern California, Los Angeles, California. This research was supported under Grant AFOSR 62-68 while the author was at the University of California, Los Angeles.

We will assume that the object of the control action is to cause the plant state vector, x_j, to follow a random command input vector, $\tilde{x}_j$, generated by the equation

$$\begin{aligned} \tilde{x}_{j+1} &= g_j(\tilde{x}_j, \xi_j), \qquad 0 \leqq j \leqq N-1, \\ \tilde{x}_0 &= \tilde{x}(0), \end{aligned} \tag{2}$$

where:
$\tilde{x}_j$ = the m-dimensional command input vector of time $t = j\Delta$.
g_j is also assumed to be continuous for all j in the interval $0 \leqq j \leqq N-1$.

The performance of this system will be measured by a continuous nonnegative functional of the system tracking accuracy and the control action,

$$h(x, v, \tilde{x}) = \sum_{i=0}^{N} W_i(x_i, v_i, \tilde{x}_i), \tag{3}$$

where W_i is, itself, a continuous nonnegative functional.

The basic problem is to choose a control action in such a way that the cost of the control process is small. It is well to consider in detail what the choice entails. For each value of j it will be assumed that $v_j \in V_j$, where V_j is a closed convex set in E_k which represents the set of all allowable control actions at time $j\Delta$.

Definition 1. Let the Cartesian product at V_j sets $(V_0 \times V_1 \times \cdots \times V_N)$ be denoted by V. Any element of V will be called an allowable control action and will be denoted by v.

When the compensation element chooses the control action v_j, it will have available to it a certain quantity of information on the loop performance. This data will take the form of a vector composed of sequences of the observed plant variables. For example, it might contain the output sequence $\{x_0, \cdots, x_j\}$, and past control actions $\{v_1, \cdots, v_{j-1}\}$. We will say that at time $t = j\Delta$, the control element can observe the vector

$$\begin{aligned} z_j &= r_j(x_j, \tilde{x}_j, v_{j-1}, z_{j-1}, \xi_j), \\ z_0 &= z(0). \end{aligned} \tag{4}$$

On the basis of the observed data the control element chooses an action $v_j \in V_j$. That is,

$$v_j = \bar{u}_j(z_j). \tag{5}$$

With this in mind, we make the following definitions.

Definition 2. Let the information available to the control element at time $t = j\Delta$ be denoted by z_j. Let the range of z_j be Z_j, and let $(Z_0 \times Z_1 \times \cdots \times Z_N)$ be given by Z. An element of Z will be indicated by z.

Definition 3. A control policy is a function, $\bar{u}$, from Z to V such that if

a and b are elements of Z and if $a_i = b_i$ for some integer i in $[0, N]$, then $\bar{u}(a)_i = \bar{u}(b)_i$. The set of all such $\bar{u}$ will be denoted by Γ.

In (1), (2) and (4) we have used the random vector ξ_j to account for the random or undetermined elements of the system and/or the environment in which the system operates. For example one component of ξ_j, $0 \leqq j \leqq N$, could be a constant which represents an unknown parameter in the plant description. Another could be measurement noise in v_j. In any case, the sequence $\{\xi_j\}$ will be assumed to have a joint probability distribution function of the form $P(\xi_0, \xi_1, \cdots, \xi_N \mid \theta)$, where θ is in general a vector. The parameter vector θ represents the basic uncertainty about the process, and it is constrained to be an element of a known parameter set Θ.

Before we complete the description of the control problem, it would be apropos to examine the fundamental characteristics of the disturbance. Basically, the process $\{\xi_j\}$ is a sequence of random vectors chosen according to one of a class of known probability distributions. This class is indexed by the vector θ and it is known that $\theta \in \Theta$. In this formulation θ is the unspecified portion of the system model and for convenience θ will be referred to as the "unknown parameter" in the process description. It may be the case that the engineer initially has some information about the relative probabilities of the elements of Θ. In this event the uncertainty about the true value of θ may perhaps be re-expressed as an a priori probability statement. The vector θ would be considered to be a random variable in this description of the system. But because θ is time invariant, it will be labeled "unknown" even in this circumstance.

With the above definitions in mind we can write the cost of the control process as

$$(6) \qquad h(x, v, \tilde{x}) = \sum_{i=0}^{N} W_i(x_i, \bar{u}(z_i), \tilde{x}_i).$$

Since all of the arguments of W_i are implicitly functions of the random variable ξ_k, $0 \leqq k \leqq i$, $h(x, v, \tilde{x})$ is a random number. In many situations it is appropriate to use the expected value of this number as a performance index. For a given initial state the expected cost can depend only upon the control policy, $\bar{u}$, and the value of the unknown parameter vector θ. Thus, we can write

$$(7) \qquad H(\bar{u}, \theta) = E\left\{\sum_{i=0}^{N} W_i(x_i, v_i, \tilde{x}_i)\right\}.$$

3. Game theoretic aspects of the control problem. In the usual optimization problem in which Θ contains only one element, θ_0, the optimal control policy is chosen to minimize $H(\bar{u}, \theta_0)$. Unfortunately, in the general case the criterion of performance induces only a partial ordering on Γ. This follows from the fact that the ordering is a function of the unknown vector θ, and it may change for different values of $\theta \in \Theta$. To obtain a "best" policy it is necessary to utilize an additional performance measure. For this purpose we will introduce a few definitions and results from the theory of games (see [1]).

Definition 4. Denote the set of all probability distribution functions over elements of Θ by Θ^*. Elements of Θ are represented in Θ^* by degenerate distributions. Elements of Θ^* are denoted by θ^*.

Definition 5. If $\theta_0^* \in \Theta^*$ and $\bar{u}^{(0)} \in \Gamma$, then define the Bayes cost of $\bar{u}^{(0)}$ with respect to θ_0^* as

$$H(\bar{u}^{(0)}, \theta_0^*) = \int H(\bar{u}^{(0)}, \theta)\, d\theta_0^*.$$

Definition 6. If $\theta_0^* \in \Theta^*$ and $\bar{u}^{(0)} \in \Gamma$, and if

$$H(\bar{u}^{(0)}, \theta_0^*) = \inf_{\bar{u} \in \Gamma} H(\bar{u}, \theta_0^*),$$

then $\bar{u}^{(0)}$ is called a Bayes control policy with respect to θ_0^*.

Definition 7. If for every $\epsilon > 0$, there exists $\theta_\epsilon^* \in \Theta^*$ such that

$$H(\bar{u}^{(0)}, \theta_\epsilon^*) \leqq \inf_{\bar{u} \in \Gamma} H(\bar{u}, \theta_\epsilon^*) + \epsilon,$$

then $\bar{u}^{(0)}$ is called an extended Bayes control policy.

Definition 8. If there exists $\bar{u}^{(0)} \in \Gamma$ such that

$$\sup_{\theta^* \in \Theta^*} H(\bar{u}^{(0)}, \theta^*) = \inf_{\bar{u} \in \Gamma} \sup_{\theta^* \in \Theta^*} H(\bar{u}, \theta^*),$$

then $\bar{u}^{(0)}$ is called a minimax policy.

Definition 9. If there exists an element $\bar{u}^{(0)} \in \Gamma$ such that $H(\bar{u}^{(0)}, \theta) = C$ for all $\theta \in \Theta$, then $\bar{u}^{(0)}$ is an equalizer policy.

Result 1.[1] *If $\bar{u}^{(0)} \in \Gamma$ is an equalizer and if it is also extended Bayes, then it is a minimax policy.*

In the above definitions we have used the same symbol for the Bayes cost and the expected cost for a specific value of $\theta \in \Theta$. This is done for notational convenience since Θ will be treated as if it were a subset of Θ^*. If we suppose that the choice of the true value of θ is made by an entity called nature, it is clear that allowing randomized strategies for nature

[1] This result is not stated explicitly in [1] but follows in an obvious way from the above definitions (see [5]).

$$\sup_{\theta^*} H(\bar{u}^{(0)}, \theta^*) = C = H(\bar{u}^{(0)}, \theta_\epsilon^*) \leqq \inf_{\bar{u}} H(\bar{u}, \theta_\epsilon^*) + \epsilon$$

$$\leqq \sup_{\theta^*} \inf_{\bar{u}} H(\bar{u}, \theta^*) + \epsilon \leqq \inf_{\bar{u}} \sup_{\theta^*} H(\bar{u}, \theta^*) + \epsilon.$$

as we have done makes the control problem more difficult in the sense that

$$\sup_{\theta^* \in \Theta^*} H(\bar{u}, \theta^*) \geqq \sup_{\theta \in \Theta} H(\bar{u}, \theta). \tag{8}$$

The reader might then suggest that the control system designer might do better to consider probability distributions over elements of Γ. Such a suggestion was made by Feldbaum [2]. It can be shown, however, that under rather mild restrictions, randomized control policies are not superior to pure control policies when we restrict ourselves to a Bayes cost ordering of Γ (see [1]).

In this paper we will consider the synthesis of Bayes control policies with respect to a given a priori distribution $\theta_0^* \in \Theta^*$. If it is known that the actual value of θ in the system is a particular realization of a random vector with distribution function θ_0^*, then the Bayes policy possesses characteristics that one would intuitively look for in an "optimal" control. If there is no such a priori knowledge, a minimax policy might seem more appropriate. Result 1 indicates that minimax and Bayes policies are intimately related, and, it can be shown that in many cases all "good" control policies are Bayes with respect to some a priori distribution. The Bayes ordering is a mathematical restriction, of course, as evidenced by the elimination of randomized control policies from consideration. This can, however, be viewed as a great practical advantage of Bayes rules since the complexity of the mechanization of a randomized policy could be prohibitive.

The reader should observe at this point the intimate relation between the problem posed in this paper and the fixed sample size game defined in [1] (see, in particular, [1, Definition 3.5.4]). The set of strategies which nature has been allowed are more restrictive than permitted in the general two-person, zero-sum game in the sense that nature is permitted no observation of the control rule, $\bar{u}$, chosen by the engineer while $\bar{u}$ will be an explicit function of z_i. This lack of symmetry provides at least intuitive justification for the assertion that pure control rules are at least as good as randomized control policies.

Let us also mention parenthetically that the minimax cost given in Definition 8 may not be the value of the control process. For questions relating to the value of the process and the maximin control policy the reader is again referred to [1].

4. Evaluation of the Bayes cost. Before we can choose a policy $\bar{u}^{(0)} \in \Gamma$ which minimizes $H(\bar{u}, \theta^*)$, we must write out the Bayes cost explicitly.

DEFINITION 10. The $(s + 1)$-fold Cartesian product of E_n spaces $(E_n \times E_n \times \cdots \times E_n)$ will be denoted by X^s. An element of X^s will be denoted by x^s, and will represent the space-time history of the system state vector from time $t = 0$ to time $t = s\Delta$. Similiarly, the $(s + 1)$-fold Cartesian product of E_k will be denoted by V^s. An element of V^s will be denoted by v^s and will represent the space-time history of the control action from $t = 0$ to $t = s\Delta$. The history of ξ, $\tilde{x}$, and z will be denoted in the same way.

At any time $s\Delta$, the control action will depend upon the information vector z_s. Assume that at time $s\Delta$ the following sequences were known: v^{s-1}, x^s, ξ^s, $\tilde{x}^s$ and θ. Then, using the notation of [2], define

$$\begin{aligned} r_s &= E\{W_s|\, v^{s-1}, x^s, \xi^s, \tilde{x}^s, \theta\} \\ &= W_s(x_s, \bar{u}_s(z_s), \tilde{x}_s). \end{aligned} \tag{9}$$

The expected incremental cost at time $t = s\Delta$ is

$$\begin{aligned} R_s &= E\{r_s\} \\ &= \int W_s(x_s, \bar{u}_s(z_s), \tilde{x}_s) p(v^s, x^s, \tilde{x}^s, \xi^s, \theta)\; d\Omega(x^s, \tilde{x}^s, \xi^s, v^s, \theta), \end{aligned} \tag{10}$$

where $\int d\Omega(\;)$ represents an integration over the whole region of variation of the argument of Ω.

Equation (10) contains a rather unwieldy joint probability density function. It can be broken down into more manageable parts as follows:

$$p(v^s, x^s, \tilde{x}^s, \xi^s, \theta) = p(v^s, \tilde{x}^s, x^s, \xi^s|\, \theta)p(\theta). \tag{11}$$

The conditional density function can be further reduced:

$$\begin{aligned} p(v^s, x^s, \tilde{x}^s, \xi^s|\, \theta) = p(x_0, \tilde{x}_0, v_0, \xi_0|\theta)p(v_1, x_1, \tilde{x}_1, \xi_1|\, \theta, x_0, \tilde{x}_0, v_0, \xi_0) \\ \cdots p(v_i, x_i, \tilde{x}_i, \xi_i|\, \theta, x^{i-1}, \tilde{x}^{i-1}, v^{i-1}, \xi^{i-1}) \\ \cdots p(v_s, x_s, \tilde{x}_s, \xi_s|\, \theta, x^{s-1}, \tilde{x}^{s-1}, v^{s-1}, \xi^{s-1}). \end{aligned} \tag{12}$$

Finally, the product of conditional density functions becomes

$$\begin{aligned} p(v^s, x^s, \tilde{x}^s, \xi^s|\, \theta) = \prod_{i=0}^{s} p(x_i|\, x^{i-1}, \tilde{x}^{i-1}, v^{i-1}, \xi^{i-1}, \theta) \\ \cdot \prod_{i=0}^{s} p(\tilde{x}_i|x^i, \tilde{x}^{i-1}, v^{i-1}, \xi^{i-1}, \theta) \\ \cdot \prod_{i=0}^{s} p(v_i|x^i, \tilde{x}^i, v^{i-1}, \xi^i, \theta) \\ \cdot \prod_{i=0}^{s} p(\xi_i|x^i, \tilde{x}^i, v^{i-1}, \xi^{i-1}, \theta), \end{aligned} \tag{13}$$

where x^{-1}, $\tilde{x}^{-1}$, v^{-1}, and ξ^{-1} are dummy vectors. Using (1), (2), (4) and (5), we can simplify (13) considerably. Many of the conditional density functions listed in (13) are degenerate because a functional relationship exists between the arguments. For economy of notation, we will suppress the explicit listing of the arguments of $\bar{u}_j$, f_j and g_j. Then (10) becomes

$$(14)\qquad \begin{aligned} R_s = \int W_s \prod_{i=0}^{s} \delta(x_i - f_{i-1}) \prod_{i=0}^{s} \delta(\tilde{x}_i - g_{i-1}) \prod_{i=0}^{s} \delta(v_i - \bar{u}_i) \\ \cdot \prod_{i=0}^{s} p(\xi_i \mid \theta, \xi^{i-1}) p(\theta)\, d\Omega(x^s, \tilde{x}^s, \xi^s, v^s, \theta), \end{aligned}$$

where $\delta(\ )$ is the usual delta function and $f_{-1} = x_0$, $g_{-1} = \tilde{x}_0$.

In (14), $p(\theta)$ is the probability density function for θ. Since θ is a constant, its distribution function is degenerate at the true value of $\theta \in \Theta$. Unfortunately, this distribution is not known to the control element at time $t = s\Delta$ because θ is by definition the unknown parameter of the system. Let $p_0(\theta)$ be the density function which corresponds to the a priori distribution for θ. The compensation element can form an estimate of θ at $t = s\Delta$ by use of the Bayes formula

$$(15)\qquad p(\theta \mid z^s) = p_s(\theta) = \frac{p_0(\theta) p(z^s \mid \theta)}{\int p(z^s \mid \eta) p_0(\eta)\, d\Omega(\eta)}.$$

Note that the Bayes formula is expressed in terms of the observation vector z^s since any estimate of $p(\theta)$ which is made by the controller must be in terms of z^s.

An interesting special case of (15) occurs when ξ^{s-1} appears explicitly in z^s but ξ^s does not. This may occur through an auxiliary feedback path or perhaps (1) and (2) can be solved for ξ^{s-1}. The importance of this information rests on the fact that only ξ depends explicitly on θ. In this case all of the rest of the components of z^s become nuisance variables with respect to estimating θ. With an argument much like the one we used to derive (13), we can show that in this case

$$(16)\qquad p_s(\theta) = \frac{p_0(\theta) \prod_{i=0}^{s-1} p(\xi_i \mid \theta, \xi^{i-1})}{\int \prod_{i=0}^{s-1} p(\xi_i \mid \eta, \xi^{i-1}) p_0(\eta)\, d\Omega(\eta)}.$$

For the case where ξ^{N-1} is observable then,

$$H(\bar{u}, \theta_0^*) = \sum_{s=0}^{N} \int W_s \prod_{i=0}^{s} \delta(x_i - f_{i-1}) \prod_{i=0}^{s} \delta(\tilde{x}_i - g_{\iota-1})$$

$$\cdot \prod_{i=0}^{s} \delta(v_i - \bar{u}_i) \tag{17}$$

$$\cdot \frac{\left[\prod_{i=0}^{s-1} p(\xi_i \mid \theta, \xi^{i-1})\right]^2 p_0(\theta)}{\int \prod_{i=0}^{s-1} p(\xi_i \mid \eta, \xi^{i-1}) p_0(\eta)\, d\Omega(\eta)}\, d\Omega(x^s, \tilde{x}^s, \xi^{s-1}, v^s, \theta).$$

5. Synthesis of the Bayes control policy. To evaluate the Bayes control policy with respect to θ_0^*, the dual control technique of [2] can be used. This approach is closely related to the principle of optimality of [3]. Define

$$\alpha_k = \int W_k \prod_{i=0}^{k} \delta(x_i - f_{i-1}) \prod_{i=0}^{k} \delta(\tilde{x}_i - g_{i-1})$$

$$\cdot \frac{\left[\prod_{i=0}^{k-1} p(\xi_i \mid \theta, \xi^{i-1})\right]^2}{\int \prod_{i=0}^{k-1} p(\xi_i \mid \eta, \xi^{i-1}) p_0(\eta)\, d\Omega(\eta)} \cdot p_0(\theta)\, d\Omega(\theta), \tag{18}$$

$$\beta_k = \prod_{i=0}^{k} \delta(v_i - \bar{u}_i).$$

Then,

$$H(\bar{u}, \theta_0^*) = \int \alpha_N \beta_{N-1} \delta(v_N - \bar{u}_N)\, d\Omega(x^N, \tilde{x}^N, \xi^{N-1}, v^N) + \sum_{i=0}^{N-1} R_i. \tag{19}$$

If we minimize the above expression with respect to $\bar{u}_N$, we see that v_N appears only in the α_N factor. Thus,

$$\inf_{\bar{u}_N} H(\bar{u}, \theta_0^*) = \sum_{i=0}^{N-1} R_i$$

$$+ \inf_{\bar{u}_N} \int \alpha_N \beta_{N-1} \delta(v_N - \bar{u}_N)\, d\Omega(x^N, \tilde{x}^N, \xi^{N-1}, v^N). \tag{20}$$

The minimization is to be taken over all allowable $\bar{u}_N$. Since α_N is continuous in v_N, the minimum exists. Define

$$\rho_N = \inf_{\bar{u}_N} \alpha_N. \tag{21}$$

Then,

$$\inf_{\bar{u}_N} H(\bar{u}, \theta_0^*) = \int \rho_N \beta_{N-1}\, d\Omega(x^N, \tilde{x}^N, \xi^{N-1}, v^{N-1}) + \sum_{i=0}^{N-1} R_i. \tag{22}$$

If we rewrite (22) in the form

$$\inf_{\bar{u}_N} H(\bar{u}, \theta_0^*) = \int \beta_{N-2}\left\{\alpha_{N-1} + \int \rho_N \, d\Omega(x_N, \tilde{x}_N, \xi_{N-1})\right\} \cdot \delta(v_{N-1} - \bar{u}_{N-1}) \, d\Omega(x^{N-1}, \tilde{x}^{N-1}, \xi^{N-2}, v^{N-1}) + \sum_{i=0}^{N-2} R_i, \tag{23}$$

and define

$$\rho_{N-1} = \inf_{\bar{u}_{N-1}} \left\{\alpha_{N-1} + \int \rho_N \, d(x_N, \tilde{x}_N, \xi_{N-1})\right\},$$

then it follows that

$$\inf_{\bar{u}_{N-1}} \inf_{\bar{u}_N} H(\bar{u}, \theta_0^*) = \sum_{i=0}^{N-2} R_i + \int \beta_{N-2}\,\rho_{N-1} \, d\Omega(x^{N-1}, x^{N-1}, \xi^{N-2}, v^{N-2}). \tag{24}$$

Consequently, if we denote that $\bar{u}_j$ which minimizes ρ_j by $\hat{u}_j$, we have the following sequential procedure for evaluating the Bayes control policy,

$$\begin{aligned} \rho_{N+1} &= 0, \\ \rho_{N-j} &= \inf_{\bar{u}_{N-j}} \left\{\alpha_{N-j} + \int \rho_{N-j+1} \, d\Omega(x_{N-j+1}, \tilde{x}_{N-j+1}, \xi_{N-j})\right\}, \\ \inf_{\bar{u} \in \Gamma} H(\bar{u}, \theta_0^*) &= \rho_0, \\ \hat{u} &= \{\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_N\}. \end{aligned} \tag{25}$$

6. Examples. To illustrate the development of the preceding sections, two examples will be considered here. The first example is chosen from [4]. This is a stochastic control problem in which there are no unknown system parameters, and the concepts of minimax policies and Bayes policies coalesce into simply an "optimal" policy. The reason for presenting this example is to show how the above work relates to some already published results.

Let the system be described by the equation

$$\begin{aligned} x_{j+1} &= \Phi_j x_j + \Delta_j v_j, \qquad 0 \leqq j \leqq N, \\ x_0 &= x(0), \\ \tilde{x}_j &\equiv 0, \\ z_j &= \begin{bmatrix} x_j \\ v^{j-1} \end{bmatrix}. \end{aligned} \tag{26}$$

The matrices Φ_i and Δ_j are matrices with random elements. These elements are assumed to be completely described statistically, and, therefore, θ_0^* is degenerate at the true value of θ, $\tilde{\theta}$. What is more, we will assume that these elements are independent from one time increment to the next, and that the means and covariances of the random elements are finite.

It will be the job of the compensation element to generate the scalar control action v in such a way that

$$H(\bar{u}, \tilde{\theta}) = E\left\{\sum_{j=0}^{N} x_j{}^T Q x_j + v_j{}^2\right\} \tag{27}$$

is minimized. Q is a positive symmetric matrix.

For this problem the form of α_k becomes quite simple,

$$\alpha_k = (x_k{}^T Q x_k + v_k{}^2) \prod_{i=0}^{k} \delta(x_i - \Phi_{i-1} x_{i-1} - \Delta_{i-1} v_{i-1}). \tag{28}$$

Define

$$\delta(x_i - \Phi_{i-1} x_{i-1} - \Delta_{i-1} v_{i-1}) = \delta_i .$$

Then, if ρ_{j+1} has the form

$$\rho_{j+1} = x_{j+1}^T P_{j+1} x_{j+1} \prod_{i=0}^{j+1} \delta_i , \tag{29}$$

we will obtain the following form for ρ,

$$\begin{aligned} \rho_j &= \inf_{\bar{u}_j}\left\{\alpha_j + \int \rho_{j+1}\, d\Omega(x_{j+1}, \xi_j)\right\} \\ &= \inf_{\bar{u}_j}\left\{\left[x_j{}^T Q x_j + v_j{}^2 + \int \{(\Phi_j x_j + \Delta_j v_j)^T P_{j+1}(\Phi_j x_j + \Delta_j v_j)\}\right.\right. \\ &\qquad \left.\left. \cdot p(\Phi_j, \Delta_j)\, d\Omega(\Phi_j, \Delta_j)\right] \prod_{i=0}^{j} \delta_i\right\}. \end{aligned} \tag{30}$$

We have used the notation $p(\Phi_j, \Delta_j)$ to represent the joint probability density function for the elements of the Φ_j and Δ_j matrices. Let us define

$$\int f(\Phi_j, \Delta_j) p(\Phi_j, \Delta_j)\, d\Omega(\Phi_j, \Delta_j) = \overline{f(\Phi_j, \Delta_j)}. \tag{31}$$

The minimum of the quadratic form in v_j can be found quite simply. The optimal control rule is given by

$$\begin{aligned} \hat{u}_j &= a_j{}^T x_j , \\ a_j{}^T &= -[\overline{\Delta_j{}^T P_{j+1} \Delta_j} + 1]^{-1} \overline{\Delta_j{}^T P_{j+1} \Phi_j}, \end{aligned} \tag{32}$$

and

$$\begin{aligned} \rho_j &= x_j{}^T P_j x_j \prod_{i=0}^{j} \delta_i , \\ P_j &= Q + \overline{\Phi_j{}^T P_{j+1} \Phi_j} - a_j(\overline{\Delta_j{}^T P_{j+1} \Delta_j} + 1) a_j{}^T. \end{aligned} \tag{33}$$

From (28) it is clear that

$$P_{N+1} = 0. \tag{34}$$

Thus (32), (33) and (34) provide a recurrence formula which can be solved for $\hat{u} = \{\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_N\}$.

The second example to be considered is much more interesting from the conceptual point of view. In this system there is a random parameter for which the complete statistical characterization is not known. Thus, the control element must "learn" about this parameter as the process unfolds. The equation which describes the system is the scalar difference equation,

$$\begin{aligned} x_{j+1} &= \xi_j x_j + v_j, 0 \leqq j \leqq N, \\ x_1 &= x(1), \\ \tilde{x}_j &\equiv 0. \end{aligned} \tag{35}$$

The observable information vector is given by the formula

$$z_j = \begin{bmatrix} x^j \\ v^{j-1} \end{bmatrix}. \tag{36}$$

The object of the control policy is to minimize the expected value of a measure of the final value of the state variable,

$$H(\bar{u}, \theta_0^*) = E\{x_N^2\}. \tag{37}$$

It will be assumed that ξ^N is a sequence of independent random variables with probability density

$$p(\xi_j \mid \theta, \xi^{j-1}) = \frac{1}{\sqrt{\pi}} \exp\{-(\xi_j - \theta)^2\}. \tag{38}$$

Here θ plays the role of the unknown parameter for the system. Θ will be assumed to be the real line. Note the essential difference between this example and the one preceding. If Θ contained only one point, $\tilde{\theta}$, this system would be contained in the set of systems described by (26). For every different $\tilde{\theta}$, we would arrive at a different optimal control policy. Thus, the problem of finding an optimal policy does not have a solution in the same sense it had in the first example. As discussed earlier, if there exists some a priori information on θ, a Bayes policy would seem appropriate. For this example, however, it will be assumed that no such a priori information exists, and a minimax policy is sought. The approach suggested in Result 1 will be pursued.

To apply this theory, the form of the Bayes rules for various $\theta_\epsilon^* \in \Theta^*$ must be considered. The set of all probability distributions over the real line is clearly a very large set. If one is fortunate, a sequence of $\theta_\epsilon^* \in \Theta^*$ which are called for by Definition 7 can be chosen from a small subset of

Θ^*. With this in mind, let us investigate the Bayes policies for $\theta_0^* = \eta(\mu, \sigma^2)$. That is,

$$p_0(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}. \tag{39}$$

It is shown in Appendix 1 that a Bayes policy with respect to θ_0^* is given by

$$\hat{u}_j = \left[-\frac{2a_{3,j}\left(2\sum_{i=0}^{j-1}\frac{x_{i+1}-v_i}{x_i} - \frac{\mu}{2\sigma^2}\right)}{a_{1,j}} + \frac{a_{2,j}\left(\sum_{i=0}^{j-1}\frac{x_{i+1}-v_i}{x_i} - \frac{\mu}{2\sigma^2}\right)}{a_{1,j}} \right] x_j, \qquad 1 \leqq j \leqq N. \tag{40}$$

One method of finding the minimax control policy involves finding an equalizer rule. One guess at an equalizer would be simply to use an unbiased estimate for θ in the formula for the optimal control policy where θ is known. Denoting this policy by $\bar{\omega}$,

$$\bar{\omega}_j = -\left\{\frac{1}{j}\sum_{i=0}^{j-1}\frac{x_{i+1}-v_i}{x_i}\right\} x_j, \qquad 1 \leqq j \leqq N. \tag{41}$$

If we combine (40) and (A.14), we see that

$$\lim_{\sigma^2\to\infty} \hat{u} = \bar{\omega}. \tag{42}$$

This, in itself, is not sufficient to prove that $\bar{\omega}$ is extended Bayes because we must prove

$$\lim_{\sigma^2\to\infty} H(\hat{u}, \theta_0^*(\sigma^2)) = \lim_{\sigma^2\to\infty} H(\bar{\omega}, \theta_0^*(\sigma^2)). \tag{43}$$

The notation $\theta_0^*(\sigma^2)$ has been used to indicate the dependence of the a priori distribution for θ on the parameter σ^2. Equation (A.17) shows that

$$\lim_{\sigma^2\to\infty} H(\hat{u}, \theta_0^*(\sigma^2)) = N_1 x(1)^2, \tag{44}$$

where N_1 is a uniformly bounded function of σ^2 for $\sigma^2 \in (0, \infty)$. The criterion function is continuous in v, and, therefore,

$$\lim_{\sigma^2\to\infty} H(\bar{\omega}, \theta_0^*(\sigma^2)) = N_1 x(1)^2. \tag{45}$$

Equations (44) and (45) prove that $\bar{\omega}$ is an extended Bayes control rule. We must now show that it is an equalizer. Before treating this question in detail, let us consider the following heuristic argument. Because of the manner in which we formulated the problem, ρ_1 provides a measure of the expected cost of the process conditioned on x_1, x_0 and v_0. In general, one might expect that ρ_1 would be implicitly dependent on ξ_0, and thus, explicitly on $(x_1 - v_0)/x_0$ since ξ_0 provides some indication of the true value of θ. Instead, as the a priori information approaches the uniform "distribution" over the real line, ρ_1 becomes independent of ξ_0. Consequently, for every ξ_0, ρ_1 is the same, and we might begin to suspect that the Bayes cost of $\hat{u}$ is becoming less dependent on the true value of θ as $\sigma^2 \to \infty$.

The simplest way of showing that $\bar{\omega}$ is an equalizer rule is to argue as follows:

$$h(x, \bar{\omega}, \bar{x}) = \left[\prod_{i=1}^{N-1}\left(\xi_i - \frac{1}{i}\sum_{i=0}^{i-1}\xi_j\right)x(1)\right]^2. \tag{46}$$

Therefore,

$$H(\bar{\omega}, \theta) = x(1)^2 E\left\{\prod_{i=1}^{N-1}\left(\xi_i - \frac{1}{i}\sum_{j=0}^{i-1}\xi_j\right)^2\right\}. \tag{47}$$

It is easy to see that $\{\xi_1 - (1/i)\sum_{j=0}^{i=1}\xi_j\}$ is a sequence of independent random variables with zero mean and variance equal to $\frac{1}{2}(1 + (1/i))$. Therefore,

$$H(\bar{\omega}, \theta) = \left(\frac{1}{2}\right)^{N-1} x(1)^2 \prod_{i=1}^{N-1}\left(1 + \frac{1}{i}\right) \tag{48}$$

for all $\theta \in \Theta$. Therefore, $\bar{\omega} \in \Gamma$ is a minimax control policy.

7. Conclusions. Since we obtain such a nice solution to the second example, the reader might wonder what would happen if the initial condition is placed on x_0 rather than x_1. For this problem the solution is quite simple. All $\bar{u} \in \Gamma$ are minimax. The truth of this assertion follows from the fact that even if we have a good estimate of θ, the expected cost of the process is proportional to x_1^2. If ξ_0 is not measured before the initiation of control, nature is permitted sufficient freedom to make $E\{x_1^2\} = \infty$ for all v_0.

It is clear that by describing the motion of the control process with a set of difference equations we have introduced an important mathematical restriction on the class of problems which can be treated. In most physical systems, however, this does not seem to be an unnatural method of description. In particular, if the loop contains inertial elements and if the control energy is bounded, it is intuitively clear that a discrete time model of the process will be adequate if the time increment is chosen appropriately. In the same way, if the control policy approaches some limiting form as $N \to \infty$, the results of this technique may be suitable for optimization over an infinite period.

Appendix 1. From (18),

$$\text{(A.1)} \qquad \alpha_N = \int \frac{x_N^2 \prod_{i=0}^{N} \delta_i \left[\prod_{i=0}^{N-1} p(\xi_i \mid \theta, \xi^{i-1})\right]^2}{\int \prod_{i=0}^{N-1} p(\xi_i \mid \eta, \xi^{i-1}) p_0(\eta)\, d\Omega(\eta)}\, d\Omega(\theta).$$

Since α_N is independent of v_N, v_N is arbitrary and $\alpha_N = \rho_N$. Also $W_j = 0$ for $j < N$ and thus $\alpha_j = 0$ for $j < N$. Using (38) and (39), one can show that

$$\text{(A.2)} \qquad \begin{aligned} &\int \frac{\left[\prod_{i=0}^{k} p(\xi_i \mid \theta, \xi^{i-1})\right]^2 p_0(\theta)\, d\Omega(\theta)}{\int \prod_{i=0}^{k} p(\xi_i \mid \eta, \xi^{i-1}) p_0(\eta)\, d\Omega(\eta)} \\ &\quad = K_k \exp\left\{ -\sum_{i=0}^{k} \xi_i^2 + \frac{\left(2\sum_{i=0}^{k} \xi_i - \frac{\mu}{2\sigma^2}\right)^2}{2(k+1) + \frac{1}{2\sigma^2}} - \frac{\left(\sum_{i=0}^{k} \xi_i - \frac{\mu}{2\sigma^2}\right)^2}{k+1+\frac{1}{2\sigma^2}} \right\}, \end{aligned}$$

where K_k is a continuous function of k and σ^2 which is uniformly bounded as a function of σ^2 in the interval $(0, \infty)$ for all k. If we substitute (A.2) into (A.1) and perform the required integration, we obtain the result

$$\text{(A.3)} \qquad \begin{aligned} \int \rho_N\, d\Omega(x_N, \xi_{N-1}) = K_N \prod_{i=0}^{N-1} \delta_i \Biggl\{ &\frac{x_{N-1}^2}{2a_{1,N-1}} \\ &+ \left[v_{N-1} + x_{N-1} \left(\frac{2a_{3,N-1}\left(2\sum^{N-2} \xi_i - \frac{\mu}{2\sigma^2}\right) - a_{2,N-1}\left(\sum^{N-2} \xi_i - \frac{\mu}{2\sigma^2}\right)}{a_{1,N-1}} \right)\right]^2 \Biggr\} \\ &\cdot \exp\{\lambda_{N-1}\}, \end{aligned}$$

where we define

$$\text{(A.4)} \qquad \begin{aligned} \lambda_j = &-\sum_{i=0}^{j-1} \xi_i^2 + \tilde{a}_{3}, j \left(2\sum^{j-1} \xi_i - \frac{\mu}{2\sigma^2}\right)^2 - \tilde{a}_{2,j}\left(\sum^{j-1} \xi_i - \frac{\mu}{2\sigma^2}\right)^2 \\ &+ \frac{\left[2a_{3,j}\left(2\sum^{j-1} \xi_i - \frac{\mu}{2\sigma^2}\right) - a_{2,j}\left(\sum^{j-1} \xi_i - \frac{\mu}{2\sigma^2}\right)\right]^2}{a_{1,j}} \\ &+ a_{4,j}\left(2\sum^{j-1} \xi_i - \frac{\mu}{2\sigma^2}\right)\left(\sum^{j-1} \xi_i - \frac{\mu}{2\sigma^?}\right). \end{aligned}$$

The $a_{i,j}$ in (A.4) satisfy the following recurrence formula:

$$a_{1,N-1} = 1 + \frac{1}{N + \frac{1}{2\sigma^2}} - \frac{4}{2N + \frac{1}{2\sigma^2}},$$

$$\tilde{a}_{2,N-1} = a_{2,N-1} = \frac{1}{N + \frac{1}{2\sigma^2}},$$

$$\tilde{a}_{3,N-1} = a_{3,N-1} = \frac{1}{2N + \frac{1}{2\sigma^2}},$$

$$a_{4,N-1} = 0,$$

$$a_{1,j-1} = 1 - 4\tilde{a}_{3,j} + \tilde{a}_{2,j} + \frac{8a_{3,j}\,a_{2,j} - 16a_{3,j}^2 - a_{2.j}^2}{a_{1,j}} - 2a_{4.j}. \tag{A.5}$$

$$a_{2,j-1} = \tilde{a}_{2,j} - a_{4,j} + \frac{4a_{3,j}\,a_{2,j} - a_{2,j}^2}{a_{1,j}},$$

$$a_{3,j-1} = \tilde{a}_{3,j} + \frac{a_{4,j}}{4} + \frac{4a_{3,j}^2 - a_{3,j}\,a_{2,j}}{a_{1,j}},$$

$$\tilde{a}_{2,j-1} = \tilde{a}_{2,j} - \frac{a_{2,j}^2}{a_{1,j}},$$

$$\tilde{a}_{3,j-1} = \tilde{a}_{3,j} + \frac{4a_{3,j}^2}{a_{1,j}},$$

$$a_{4,j-1} = a_{4,j} - \frac{4a_{3,j}\,a_{2,j}}{a_{1,j}}.$$

Returning to (A.3), we see that only one factor depends on v_{N-1}. Thus the optimal control action at $t = (N - 1)\Delta$ is

$$\hat{u}_{N-1} = -\frac{2a_{3,N-1}\left(2\sum^{N-2}\xi_i - \frac{\mu}{2\sigma^2}\right) - a_{2,N-1}\left(\sum^{N-2}\xi_i - \frac{\mu}{2\sigma^2}\right)}{a_{1,N-1}}\,x_{N-1}, \tag{A.6}$$

and

$$\rho_{N-1} = N_2 x_{N-1}^2 \prod_{i=0}^{N-1} \delta_i \exp\{\lambda_{N-1}\}. \tag{A.7}$$

Now assume ρ_j takes the form

$$\rho_j = N_j x_j^2 \prod_{i=0}^{j} \delta_i \exp\{\lambda_j\}. \tag{A.8}$$

Performing the indicated integrations, we find that

$$\text{(A.9)}\quad \int \rho_j \, d\Omega(x_j\,,\, \xi_{j-1}) = K_j \prod_{i=0}^{j-1} \delta_i \left\{ \frac{x_{j-1}^2}{2a_{1,j-1}} + \left[v_{j-1} + x_{j-1} \left(\frac{2a_{3,j-1}\left(2 \sum^{j-2} \xi_i - \frac{\mu}{2\sigma^2}\right) - a_{2,j-1}\left(\sum^{j-2} \xi_i - \frac{\mu}{2\sigma^2}\right)}{a_{1,j-1}} \right) \right]^2 \right\} \cdot \exp \{\lambda_{j-1}\}.$$

We can see directly from (A.9) that a Bayes control policy is given by

$$\text{(A.10)}\quad \hat{u}_j = - \frac{2a_{3,j}\left(2 \sum^{j-1} \xi_i - \frac{\mu}{2\sigma^2}\right) - a_{2,j}\left(\sum^{j-1} \xi_i - \frac{\mu}{2\sigma^2}\right)}{a_{1,j}} \, x_j\,,$$

and that

$$\text{(A.11)}\quad \rho_{j-1} = N_{j-1} x_{j-1}^2 \prod_{i=0}^{j-1} \delta_i \exp \{\lambda_{j-1}\}.$$

By induction it follows that a Bayes control policy is given by (A.10) for all integers j in the interval $1 \leqq j \leqq N$. Note that many Bayes policies are possible because v_N can be chosen arbitrarily.

Let us next consider the Bayes cost of the control process. From (A.11),

$$\text{(A.12)}\quad \rho_1 = N_1 x_1^2 \delta_1 e^{\lambda_1}.$$

It is interesting to note what happens to this index of performance as $\sigma^2 \rightarrow \infty$. From (A.4),

$$\text{(A.13)}\quad \lim_{\sigma^2 \to \infty} \lambda_1 = \xi_0^2 \left(-1 + \lim_{\sigma^2 \to \infty} \left\{ 4\tilde{a}_{3,1} - \tilde{a}_{2,1} + 2a_{4,1} + \frac{16a_{3,1}^2 + a_{2,1}^2 - 8a_{2,1}a_{3,1}}{a_{1,1}} \right\} \right).$$

It can be shown using (A.5) that

$$\text{(A.14)}\quad \lim_{\sigma^2 \to \infty} a_{1,j} = \frac{j}{j+1}\,, \quad \lim_{\sigma^2 \to \infty} a_{2,j} = \frac{1}{j+1}\,, \quad \lim_{\sigma^2 \to \infty} a_{3,j} = \frac{1}{2(j+1)}\,.$$

Another identity which is contained in (A.4) is

$$\text{(A.15)}\quad 4\tilde{a}_{3,j} - \tilde{a}_{2,j} + 2a_{4,j} = 4a_{3,j} - a_{2,j}\,.$$

Combining these equations,

$$\text{(A.16)}\quad \lim_{2 \to \infty} \lambda_1 = 0,$$

and

$$\lim_{\sigma^2 \to \infty} \rho_1 = N_1 x_1^2 \, \delta(x_1 - x(1)). \tag{A.17}$$

REFERENCES

[1] D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*, Wiley, New York, 1954.

[2] A. A. Feldbaum, *Dual control theory II*, Automat. Remote Control, 21 (1960), pp. 1033–1039.

[3] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, 1961.

[4] T. L. Gunckel and G. F. Franklin, *A general solution for linear, sampled-data control*, ASME J. Basic Engrg., 85 (1963), pp. 197–203.

[5] T. Ferguson, unpublished notes.