J.SIAM CONTROL
Ser. A, Vol. 1, No. 1
Printed in U.S.A., 1962

INTRODUCTION*

BROCKWAY McMILLAN†

With this issue of its *Journal*, the Society for Industrial and Applied Mathematics initiates a new series devoted to articles on the mathematical and scientific problems of control. By thus centralizing what it publishes in the field of control, the Society hopes to stimulate mathematical interest and research in the theory, and to facilitate the dissemination of applicable results to those who have use for them.

Many of us in the Society have felt that control theory both needs, and merits, more explicit attention by mathematicians and by engineers of mathematical bent. It needs attention because of the pervasive importance of control systems to modern technological society. Many evident applications presently suffer for lack of adequate theory; I suspect that an adequately fundamental theory would expose many more applications not now evident. The field merits attention because it shows promise of a rich mathematical harvest. It is a field well illuminated by heuristic guides. Observable physical phenomena, and special problems already solved in the folklore or the literature, abound to support the intuition. To those who do not fear mathematical phenomena that have physical or engineering counterparts, it can be a challenging and, I believe, rewarding domain to explore.

Control systems involving plants as diverse in form as amplifiers, aircraft, chemical processes, and commercial enterprises are now subject to engineering design and manipulative control guided to some degree by mathematical theory. Beyond this already great diversity, there are other control systems of importance to society that require understanding. To mention some extreme examples, society itself, or its economic under structure, is a vast control system. I have little doubt that a mathematical theory of control will some day contribute to our understanding of this system, and perhaps even to a degree of manipulative control over it. The behavior that is characteristic of biological organisms, specifically of man, is that of a control system; from the point of view of the meteorologist, the earth's atmosphere is also a control system. Whether a general theory of control, as control, will ever contribute usefully to understanding or manipulation of these latter phenomena is much more speculative but not evidently impossible.

If one were to poll those who design or manipulate control systems about the need for further theory, the positive replies would undoubtedly emphasize a need for better design methods. Herein lies a trap for both engineer and mathematician. Theory creates understanding. Better designs may result from a theory because the theory is penetrating, is revealing

* Invited by the editors.

† Assistant Secretary of the Air Force for Research and Development, Washington, D. C.

or definitive about the possibilities available, or indicates what is needed to realize the full potential of the plant to be controlled. But to set as a primary objective of mathematical effort the development of explicit design procedures may so limit the problem that nothing results. Furthermore, it is delusive to expect that theory, even if successful, will automatically lead to design procedures that are simple. Theory may show the way to solve a hard problem, but it cannot be expected to eliminate the need for the data or the computations that are intrinsic to the problem.

To put the thoughts of the last paragraph into direct form, one can say, first, to the engineer: don't expect too much. It is characteristic of, and a virtue of, fundamental theories that they are general. Don't expect a fundamental result to solve a detailed problem. Be satisfied if it tells you the data you need, the calculations to make, and some general facts about the result you will get. And, incidentally, don't be surprised if some of the data you need are data you can't get. This may mean that the theory has not been general enough.

Similarly, to the mathematician, one can say: don't expect too much. It is you who must find the right problems. Don't expect engineers to give you good problems ready made. They will give you problems—specific, complex, computational, unrewarding. It is you who must develop the intuition to see the phenomenon behind its manifestation. It is you who must develop the practical sense to make the right idealizations and to establish the right criteria of performance.

I have emphasized mathematical opportunities and the needs of technology for better theory and better understanding. Let me close by emphasizing what I think is an important specific need of the mathematician—indeed even of the mathematician already working in the field. This is the need for a comprehensive synthesis from a mathematical point of view of the present state of knowledge. We need, and the subject is ripe for, an inclusive account of the several important streams of mathematical thought now flowing, an account which relates these to each other and to the problems of technology that gave them rise. From this, alone, I suspect, one would perceive mathematical gaps which would be usefully and rather easily closed. Perhaps opportunities for useful generalization would also become evident. Such an account would also provide a framework within which to formulate further problems of technology, many of which, I am sure, have not been sensed by mathematicians in enough generality to be interesting.

Perhaps the greatest direct contribution the Society could make to the further development of the theory of control would be to stimulate and publish a sympathetic, consistent, and unified synthesis of the field from a mathematician's point of view. One can hope that by centralizing and encouraging publication on control, the Society may succeed in doing this.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 1
Printed in U.S.A., 1962

STABILITY AND CONTROL*

J. P. LASALLE†

1. Introduction. It is reasonable to assume today, as perhaps was not so in this country a few years ago, some knowledge of Liapunov's concepts of stability and of his direct or second method for the study of the stability of dynamical systems. Even so, I would like to review first his definitions of stability and his principal theorems. It turns out, if we discuss first the general stability problem as it arises in the theory of automatic or feedback control systems, that we are led rather naturally to Liapunov's stability method. Following this natural course, we will find ourselves reversing the usual order of presentation. We discuss first the theorems and afterwards the nature of the stabilities assured by the theorems. Next, we shall look at some simple illustrations of the application of the Liapunov method and then consider some linear and nonlinear control problems.

It is, of course, impossible for us today to consider any of these topics in great detail, and for this reason I have included a list of references. Although this list contains a few references which are there only for historical reasons, it is designed primarily to tell you where you might begin to obtain more detailed information about the topics I intend to discuss.

2. Stability of control systems and Liapunov's direct method ([1]–[5] and [7]). To make things simpler for ourselves, let us assume at the outset that all of our functions have continuous first partial derivatives. In many places this is far more than is required but is satisfactory for our purposes.

The behavior of the control system is, we assume, described by a system of differential equations of the form

$$\dot{x}_i = f_i(x_1, \cdots, x_n, u_1, \cdots, u_r), \qquad i = 1, \cdots, n,$$

where $\dot{x}_i = dx_i/dt$. Making use of vector notation, we can write this system in the more manageable form

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

where

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ \vdots \\ u_r \end{pmatrix}, \quad \text{and} \quad f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}.$$

* Received by the editors January 5, 1962. This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under contract No. AF 49(638)-382, and in part by the U. S. Army, Army Ballistic Missile Agency under contract DA-36-034-ORD-3514 RD.

† RIAS, Baltimore 12, Maryland.

We look upon x as the "error" in control and u as the control function we wish to select. The uncontrolled system is

$$\dot{x} = f(x, 0), \tag{2}$$

and zero error is an equilibrium state: $f(0, 0) = 0$. The control system is to be automatic—another way of saying this is that we want feedback control—and u is to be a function of the state of the system x. Thus, we will have

$$\dot{x} = f(x, u(x)) = F(x), \qquad F(0) = 0. \tag{3}$$

The class of allowable control functions u will be limited. They may be restricted in range and may be limited to being special types of functions. In fact, it can be that we are able to choose only a finite number of parameters that enter into the description of the control u.

An early example of this was the problem considered by Minorsky [17] in 1923. He was concerned with the design of an automatic steering device for a large ship (the New Mexico). The differential equations for the turning moment of the ship were approximated by

$$a\ddot{\theta} + b\dot{\theta} - c\theta = 0, \qquad a > 0, b > 0, c > 0.$$

The uncontrolled motion was unstable and the controlled system was

$$a\ddot{\theta} + b\dot{\theta} - c\theta = u(\theta, \dot{\theta}, \ddot{\theta})$$

where

$$u(\theta, \dot{\theta}, \ddot{\theta}) = -(\alpha\ddot{\theta} + \beta\dot{\theta} + \gamma\theta).$$

The problem was a completely linear one and the equations of the controlled motion could be written

$$(a + \alpha)\ddot{\theta} + (b + \beta)\dot{\theta} + (\gamma - c)\theta = 0.$$

The coefficients α, β, γ of control were selected to stabilize the ship ($\gamma > c$), to decrease the moment of inertia ($a + \alpha > 0$ with α as negative as possible), and to increase the resistance to turning ($\beta > 0$). The question of stability offers no difficulty here and the other parameters could be selected intuitively or by experimentation. Such simple linear problems offer no difficulties.

Let us return to the general problem of the control system (3). Although this is not the only way of proceeding, we assume that what we want to do first is to assure that the control tends to keep the error small and also tends to reduce the error monotonically. With this point of view we have a natural way of presenting Liapunov's direct method. Let $V(x)$ be a measure of the error x. $V(x)$ is real-valued and to be a reasonable

measure, at least locally, it must be positive definite. This means that in some neighborhood N of the origin (perhaps a sphere of radius r around the origin)

$$V(x) > 0, \qquad x \neq 0, \tag{4}$$
$$V(0) = 0.$$

Define now for the system (3)

$$\dot{V}(x) = (\text{grad } V) \cdot F(x).$$

This function can be computed directly from the differential equation without a knowledge of the solutions and is why Liapunov's method is said to be *direct.*

If $x(t)$ is a solution of (3), then

$$\frac{d}{dt}[V(x(t))] = \dot{V}(x(t)), \tag{5}$$

the rate of change of V along solutions. If the system is to tend to keep the error small, we want

$$\dot{V}(x) \leqq 0, \qquad x \text{ in } N. \tag{6}$$

If we want the system to reduce the error monotonically, we want

$$\dot{V}(x) < 0, \qquad x \text{ in } N, x \neq 0. \tag{7}$$

LIAPUNOV'S STABILITY THEOREM. *If V satisfies* (4) *and* (6), *then the origin is a stable equilibrium state of* (3).

LIAPUNOV'S ASYMPTOTIC STABILITY THEOREM. *If V satisfies* (4) *and* (7), *then the origin is asymptotically stable.*

THEOREM ON ASYMPTOTIC STABILITY IN THE LARGE. *If V satisfies* (4) *and* (7) *for all $x \neq 0$ and if, in addition, $V(x) \to \infty$ as $\| x \| \to \infty$, then the origin is asymptotically stable in the large.* ($\| x \|$ denotes the Euclidean length of the vector x).

THEOREM ON THE SIZE OF THE REGION OF ASYMPTOTIC STABILITY. *If the region R defined by $V(x) \leqq c$ is bounded and if* (4) *and* (7) *hold for all x in R, $x \neq 0$, then R is contained in the region of asymptotic stability.*

We shall explain in a moment the precise meanings of these stabilities. What we have seen is that a system which "tends to keep the error small" corresponds to stability and that one which does this and at the same time "tends to reduce the error monotonically" corresponds to asymptotic stability. What does this mean precisely? Let $x(t, x^0)$ be the solution of (3) which starts initially at x^0 ($x(0, x^0) = x^0$). Then the origin is said to be *stable* if given $\epsilon > 0$ there is a $\delta > 0$ such that $\| x^0 \| < \delta$ implies

$\| x(t, x^0) \| < \epsilon$ for all $t > 0$. If, in addition, there is an $\eta > 0$ with the property that $\| x^0 \| < \eta$ implies $x(t, x^0) \to 0$ as $t \to \infty$, then the origin is said to be *asymptotically stable.* The set of all x^0 for which $x(t, x^0) \to 0$ as $t \to \infty$ is called the *region of asymptotic stability.* If the region of asymptotic stability is the whole space, then we say that the origin is *asymptotically stable in the large.*

Mathematically, the above theorems are simple elementary results but they provide us with the only general method we have for studying stability that takes into account the nonlinearities of the system. Described in a general way one approach to the problem of control is that of finding a suitable Liapunov function $V(x)$. Then $\dot{V}$ depends on both x and u. Within the allowable set of control functions we can pick out a subset of controls which give an asymptotically stable system. Then within this subset of stable controls we can on the basis of other criteria or experimentation select the one that is best. Let me add that it is not necessary to proceed in this fashion. We may begin by optimization if we are certain that the optimization implies stability. This can often be established by Liapunov's method [12].

3. The practical significance of Liapunov stability ([3], [5]–[11]). The concepts of stability and asymptotic stability defined above are due to Liapunov and could be called *stabilities under sudden perturbations.* The perturbation suddenly moves the system from its equilibrium state but then immediately disappears. Simple stability says that the effect of this will not be great if the sudden perturbation is not too great. Asymptotic stability states, in addition, that if the sudden perturbation is not too great, the effect of the perturbation will tend to disappear. If the system is asymptotically stable in the large, the effect of the perturbation tends to disappear regardless of the size of the sudden perturbation.

In practice, however, the perturbations are not simply impulses and this led Duboshin, a Russian, to consider what he called *stability under persistent perturbations.* Today this is called simply *total stability.*

Let the unperturbed system be

$$\dot{x} = F(x) \tag{3}$$

and the perturbed system be

$$\dot{x} = F(x) + p(x, t). \tag{3*}$$

The system (3) is said to be *totally stable* if given $\epsilon > 0$ there is an $\eta > 0$ and a $\delta > 0$ such that if $\| x^0 \| < \eta$ and $\| p(x, t) \| < \delta$ for all x and $t \geqq 0$, then

$$\| x^*(t, x^0) \| < \epsilon \quad \text{for all} \quad t > 0;$$

$x^*(t, x^0)$ is the solution of (3^*) satisfying $x^*(0, x^0) = x^0$. This says that if the perturbation is not too large and if the system is not too far from the origin initially it will remain near the origin. One sees easily enough that a system may be stable without being totally stable. Consider an undamped simple harmonic oscillator with one degree of freedom. However, as was shown by Malkin and independently by Gorshin, asymptotic stability of (3) implies total stability. This result is a simple consequence of the converse of Liapunov's theorem on asymptotic stability.

Recently Seibert and Auslander have shed new light upon the nature of these stabilities. They have shown that starting with Liapunov's stability there are a whole hierarchy of stabilities (one for each ordinal number) none of which imply stability under persistent perturbations. In addition, they show that if (3) is asymptotically stable, then it possesses a stability under perturbations which is stronger than total stability. They say that a system is *strongly stable under perturbations* if it is totally stable, and in addition, has the property that there is a $\rho > 0$ such that given $\epsilon > 0$, there is a $\delta = \delta(\epsilon)$ and a $\tau = \tau(\epsilon)$ such that

$$\| x^0 \| < \rho, \qquad \| p(x, t) \| < \delta$$

imply

$$\| x^*(t, x^0) \| < \epsilon \quad \text{for all} \quad t \geqq \tau.$$

It turns out, as they show, that this strong stability under perturbations is equivalent to asymptotic stability. This gives us even greater confidence that asymptotic stability means practical stability.

There is another recent result which establishes an important relation between control and asymptotic stability. Consider now the control system

$$\dot{x} = f(x, v(t)),$$

where we consider the control to be a function of t. Assume that the uncontrolled system

$$\dot{x} = f(x, 0)$$

is asymptotically stable in the large and that the allowable control and the control system is in some sense—technically too extensive to explain here—"proper". Then Markus and Lee [22] have shown that for each initial state x^0 there is a control function $v(t)$ that moves the system to the origin in finite time. If one could go a step further and show that there was a feedback control function $u(x)$ that moves the system from each initial state to the origin in finite time, then this would be much stronger than asymptotic stability and we could expect the system to have an

extremely strong stability under perturbations. Later we return to this question and consider an example to illustrate this point.

4. Asymptotic stability of linear systems ([5] and [12]). Consider the linear system

$$\dot{x}_i = \sum_{j=1}^{n} a_{ij}x_j$$

or in matrix notation

$$\dot{x} = Ax. \tag{8}$$

We know that the question of asymptotic stability is simply the algebraic one of determining whether the characteristic values of A have negative real parts and that there are many criteria for deciding this question. Let us look at one that arises from the Liapunov method. Let Q be a positive definite (symmetric) matrix. Then the quadratic form

$$V(x) = \sum_{i,j=1}^{n} q_{ij}x_ix_j = x'Qx$$

is positive and

$$\dot{V}(x) = x'(A'Q + QA)x.$$

(A' is the transpose of A.) Therefore, if

$$A'Q + QA = -C, \qquad C \text{ positive definite}, \tag{9}$$

then the system (8) is asymptotically stable in the large by the third stability theorem given in § 2. Conversely, if (8) is asymptotically stable, then corresponding to each positive definite matrix C there is a unique positive definite matrix Q satisfying (9). In fact,

$$Q = \int_0^\infty e^{A't}Ce^{At}\,dt.$$

Thus, we see for a linear system with constant coefficients that the problem of stability is reduced to that of determining whether the linear system of equations (9) has a positive definite solution Q.

5. General criterion for asymptotic stability in the large ([23]–[26]). In this second illustration we use an idea due to Hartman to obtain in a simple way what had previously appeared to be a fairly complicated result. We are still considering the system

$$\dot{x} = F(x), \qquad F(0) = 0. \tag{3}$$

Take as the Liapunov function any positive definite quadratic form

$$V = x'Qx.$$

Then

$$\dot{V} = F'(x)Qx + x'QF(x).$$

Note that

$$F(x) = \int_0^1 J(sx)x\, ds$$

where $J(x)$ is the Jacobian matrix of $F(x)$. This identity follows easily from computing $d(F(sx))/ds$. Using this identity, we have

$$\dot{V} = \int_0^1 x'[J'(sx)Q + QJ(sx)]x\, ds.$$

Therefore, as a consequence of the third theorem stated in § 2 on asymptotic stability in the large, we see that *if for some positive definite matrix* Q, $J'(x)Q + QJ(x)$ *is negative definite for all* $x \neq 0$, *then the system* (3) *is asymptotically stable in the large.*

6. Improvement of stability and control of linear systems ([12], [17], [19], [20]). What we wish to do here is illustrate by a rather simple example the way in which Liapunov ideas can be used to select controls that improve stability and at the same time improve performance. We assume that the uncontrolled system

$$\dot{x} = Ax$$

was asymptotically stable to begin with or that it has already been stabilized. We are able to add more control and take the controlled system to be of the form

$$\dot{x} = Ax + Bu,$$

where x is an n-vector, A is a constant $n \times n$ matrix, B is a constant $n \times r$ matrix, and u is an r-vector. Our objective is to select a control u that increases the stability and at the same time improves performance. We consider the case where the performance criterion is the quadratic functional

$$I(x^0, u) = \int_0^\infty [x'Cx + u'Ru]\, dt,$$

where C and R are positive definite matrices. With a specified initial state x^0 and a given control function u, the solution of the controlled system is determined uniquely and the above integral is then a line integral along the solution. We are not sure that this integral exists but will see in a moment for the controls we select that it does exist. $\int_0^\infty x'Cx\, dt$ represents

the "cost" of the error and $\int_0^\infty u'Ru\,dt$ the "cost" of control. Select

$$V(x) = x'Qx$$

where $A'Q + QA = -C$.

Note first of all that with no control ($u = 0$)

$$\dot{V}(x) = -x'Cx,$$

and integrating from $t = 0$ to $t = \infty$ we have, since the system is asymptotically stable,

$$V(x^0) = I(x^0, 0).$$

The value of V at the initial point is equal to the measure of performance with no control. We want then to select a control u which increases stability and at the same time improves performance ($I(x^0, u) < I(x^0, 0)$). For the system with control,

$$\dot{V}(x) = -x'Cx + u'B'Qx + x'QBu = -x'Cx + 2u'B'Qx.$$

Let P be any positive definite $n \times r$ matrix and take

$$-Pu = 2B'Qx;$$

that is,

$$u = -2P^{-1}B'Qx.$$

This is linear control and we are assured that it improves stability since

$$\dot{V}(x) = -x'Cx - u'Pu.$$

However, do we obtain improvement in performance? Again integrating from $t = 0$ to $t = \infty$, we have as before

$$V(x^0) = \int_0^\infty x'Cx\,dt + \int_0^\infty u'Pu\,dt$$

and

$$I(x^0, u) = V(x^0) - \int_0^\infty u'Pu\,dt + \int_0^\infty u'Ru\,dt.$$

Since $V(x^0) = I(x^0, 0)$, we obtain immediately

$$I(x^0, 0) - I(x^0, u) = \int_0^\infty u'(P - R)u\,dt.$$

We then see that we will have improved performance ($I(x^0, u) < I(x^0, 0)$) by using this control if $P > R$, that is, if $P - R$ is positive definite. Thus, if we were to take $P = \lambda R$ we would obtain improvement only if $\lambda > 1$.

We do not draw the conclusion that making λ larger improves performance. With $P = \lambda R$ the solution over which the line integral is being evaluated depends upon λ. We are only sure that an optimal λ is greater than one.

This example is meant only to illustrate what can be done with Liapunov functions, and we leave two questions unanswered. The first is with $P = \lambda R$: *what is an optimal choice of* λ*?* The second and more important is: *can one iterate this scheme and converge to optimal linear control (a linear control that minimizes* $I(x^0, u)$*)?* One may also have some choice of the matrix B, and this we have not taken into consideration.

7. Stability of nonlinear control systems ([2]–[5], [12]–[21]). The most celebrated problem is the one first considered by Lurie and Postnikov, studied in considerable detail by Letov, and put in order mathematically by Lefschetz and Yacubovic.

The control system is described by a system of differential equations of the following form

$$\dot{x} = Ax + bf(\sigma),$$
$$\dot{\sigma} = d'x - rf(\sigma)$$

where A and x are as before, b and d are n-vectors, $f(\sigma)$, σ and r are scalars (real numbers). The uncontrolled system $\dot{x} = Ax$ is assumed to be linear and asymptotically stable and f is any continuous function satisfying

$$\sigma f(\sigma) > 0, \qquad \sigma \neq 0.$$

The σ is the feedback control signal and f is the characteristic, say, of a servomechanism. We then want to determine conditions on the parameters b, c, and r which assure that the system is asymptotically stable in the large for all $f(\sigma)$ of the above type. If b, c, and r have this property, the system is said to be *absolutely stable.* Thus, absolutely stable systems are stable for a whole class of functions f and possess a strong stability relative to perturbations of f.

The general idea behind the solution of this problem is the following: Take

$$V(x, \sigma) = x'Qx + \int_0^\sigma f(\sigma)\, d\sigma$$

where Q is selected as before so that $A'Q + QA = -C$, $C > 0$, $Q > 0$. It then turns out that

$$\dot{V} = -x'Cx + 2f(\sigma)\,(Qb + \tfrac{1}{2}\,d)x - rf^2(\sigma)$$

is a quadratic form in x and $f(\sigma)$. It is then relatively simple to write down conditions on the parameters that assure that $\dot{V}$ is negative definite.

For instance, one may pick $Qb + \frac{1}{2} d = 0$ and $r > 0$. Since C is an arbitrary positive definite matrix, there are many ways of doing this and there has been much discussion of simple kinds of sufficient conditions.

Let us look at a quite different type of nonlinear control. The example is somewhat artificial but I believe illustrates a point. Consider the linear system

$$\dot{x} = Ax + Bu$$

where B is a nonsingular $n \times n$ matrix and the control u is subject to the constraint $\| u \| \leqq 1$. Select

$$V = x'Qx,$$

where Q as before satisfies

$$A'Q + QA = -C, \qquad C > 0.$$

Then

$$\dot{V} = -x'Cx + u'B'Qx + x'QBu.$$

Now the choice of u that minimizes $\dot{V}$ is clearly

$$u = -B'Qx/\| B'Qx \|$$

and

$$\dot{V} = -x'Cx - 2 \| B'Qx \|.$$

Thus for $x \neq 0$,

$$\dot{V} \leqq -2 \| B'Qx \| \leqq -2\alpha V^{\frac{1}{2}}, \qquad \alpha > 0.$$

Taking $W = V^{\frac{1}{2}}$, we obtain

$$\dot{W} \leqq -\alpha < 0, \qquad \text{for } x \neq 0.$$

Therefore, this control reduces W to zero in finite time which means that from each initial state the system is brought to the origin in finite time. This, of course, cannot be done by linear control and the system has, as we shall show, an extremely strong stability under perturbations.

With the above control, consider the perturbed system

$$(*) \qquad \dot{x} = Ax + p(x, t).$$

For this perturbed system

$$\dot{W}_* = \dot{W} + \text{grad } W \cdot p.$$

Since for $x \neq 0$, $\| \text{grad } W \| \leqq k$, we have for all $x \neq 0$ that

$$\dot{W}_* \leqq -\alpha + k \| p \|.$$

Hence, if $\| p \| \leqq \delta < \alpha/k$, the system retains the property that from each initial state the control brings the system to the origin in finite time.

This establishes, at least for this particular case, what we believe intuitively and know from experience: *Feedback control should result in a system that has an exceptionally strong stability under perturbations.* A general result of this nature is certainly of mathematical and theoretical interest and should be of considerable practical significance.

8. Stability of functional differential equations ([29], [30]). In the last five years or so there has been an active research effort, particularly in the Soviet Union on what can be called *functional differential equations.* Such equations include, for example, quite general types of differential equations with delays where the delay may depend on the time and also the state of the system. As another example one might well wish to select a control that depends not only upon the present state of the system but which depends, perhaps, on the integral of the state over an interval of the past. Thus, in the system (1) we might have

$$u = u\left(x(t), \int_{t-\delta}^{t} x(\tau)\, d\tau\right)$$

and

$$\dot{x}(t) = f\left(x(t), u\left(x(t), \int_{t-\delta}^{t} x(\tau)\, d\tau\right)\right).$$

The system (3) then becomes

$$\dot{x} = F(x)$$

where now F is a functional and this is then a functional differential equation. Much of the Liapunov theory has been extended to systems of this type [29] and Hale in [30] has shown that the converse theorems are a powerful mathematical tool for deriving general results concerning these equations. This would appear to be an area of mathematical research of considerable interest and importance.

REFERENCES

[1] A. A. Liapunov, *Problème général de la stabilité du mouvement*, Photographically reproduced as Annals of Mathematics Study No. 17, Princeton University Press, Princeton, N. J. (This is the 1907 French translation of Liapunov's Original paper which was published in Russia in 1892.)

[2] W. Hahn, *Theorie und Anwendung der direkten Methode von Ljapunov*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag, Berlin, 1959.

[3] I. G. Malkin, *Theory of Stability of Motion*, AEC Translation No. 3352, Department of Commerce, Washington, D. C., 1958.

[4] L. Cesari, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag, Berlin, 1959.

[5] J. P. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.

[6] G. N. Duboshin, *The problem of the stability of motion under persistently acting perturbations*, Trudy Gos. Astr. Inst. (Sternberg), 14 (1940), No. 1. (Russian.)

[7] I. G. Malkin, *On stability with persistent perturbations*, Prikl. Mat. Meh., 8 (1944), pp. 241–245. (Russian.)

[8] S. I. Gorshin, *On the stability of motion under persistently acting perturbations*, Izv. Akad. Nauk Kazach. SSR, Ser. Mat. Meh. 2, 56 (1948), pp. 46–73. (Russian.)

[9] P. Seibert, *Prolongations and generalized Liapunov*, RIAS Technical Report 61-7, Baltimore, 1961.

[10] P. Seibert, *Stability under perturbations in generalized dynamical systems*, Proceedings OSR-RIAS Symposium, 1961 (to appear).

[11] J. Auslander and P. Seibert, *Prolongations and generalized Liapunov functions*, Proceedings OSR-RIAS Symposium, 1961 (to appear).

[12] R. E. Kalman and J. E. Bertram, *Control system analysis via the second method of Lyapunov. I. Continuous time systems*, Trans. A.S.M.E., 82 Ser. D. J. Basic Engrg. (1960), pp. 371–393. II. Discrete-time-systems, pp. 394–399.

[13] A. I. Lur'e, *Certain Nonlinear Problems in the Theory of Automatic Control*, Gosteknizdat, 1951. (Russian.) (German translation, Akademie Verlag, 1957. English translation, Her Majesty's Stationery, 1957.)

[14] A. M. Letov, *Stability in Nonlinear Control Systems*, Princeton University Press, Princeton, N. J., 1961. (English translation of 1955 USSR edition with an additional chapter.)

[15] V. A. Yakubovich, *On nonlinear differential equations of automatic control systems with one control unit*, Vestnik Leningrad. Univ. Ser. Mat., Meh. i Astron., 15 (1960), pp. 120–153.

[16] S. Lefschetz, *Some mathematical considerations on nonlinear controls*, Contributions to Differential Equations, Vol. 1 (to appear).

[17] N. Minorsky, *Investigation of Nonlinear Control Systems. Part 1; Continuously Acting Control Systems. Part 2*, Office of Naval Research, Washington, D. C., 1960.

[18] Beihefts zur Regelungstecnik, *Nichtlineare Regelungsvorgae*, Verlag R. Oldenbourg, Muenchen, 1956. (In particular see the papers by W. Hahn and E. Pestel.)

[19] A. A. Feldbaum, *Error Evaluation in Automatic Systems*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1959. (Russian.)

[20] M. A. Aizerman, *Lectures on the Theory of Automatic Control*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1958. (Russian.)

[21] J. P. LaSalle, *Some extensions of Liapunov's second method*, Trans. I.R.E., Circuit Theory, CT-7 (1960), pp. 520–527.

[22] L. Markus and E. B. Lee, *On the existence of optimal controls*, Trans. A.S.M.E., Ser. D. J. Basic Engrg., March 1962, pp. 13–22.

[23] P. Hartman, *On stability in the large for systems of ordinary differential equations*, Can. J. Math. (to appear).

[24] L. Markus and H. Yambe, *Global stability criteria for differential systems*, Osaka Math. J., 12 (1960), pp. 305–317.

[25] C. Olech, *On the global stability of an autonomous system on the plane*, RIAS Technical Report 61-12, Baltimore, 1961.

[26] P. HARTMAN AND C. OLECH, *On global asymptotic stability of solutions of differential equations*, Trans. Amer. Math. Soc. (to appear).

[27] V. A. PLISS, *Some Problems in the Theory of the Stability of Motion in the Large*, Isd. Leningradsk. Univ., 1958.

[28] G. P. SZEGÖ, *A contribution to Liapunov's second method: Nonlinear autonomous systems*, Trans. A.S.M.E. (to appear).

[29] N. N. KRASOVSKII, *Some Problems in the Theory of Stability of Motion*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1959. (Russian.)

[30] J. K. HALE, *Asymptotic Behavior of the Solutions of Differential-Difference Equations*, RIAS Technical Report 61-10, Baltimore, 1961.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 1
Printed in U.S.A., 1962

MINIMUM EFFORT CONTROL SYSTEMS*

LUCIEN W. NEUSTADT†

Abstract. An optimal control problem is considered in which it is desired to transfer a linear control system from one given state to another state. The target state may either be a point or a convex closed set. Optimization is understood in the sense of minimizing the control effort, where effort is defined either as maximum amplitude or as an integral of a certain function of the control. The optimization problem is reduced to the problem of finding the unique minimum of a function of n variables (where n is the order of the system). It is shown that the method of steepest descent is particularly applicable to finding this minimum, and consequently to determining the minimum effort and optimal control.

I. INTRODUCTION

We shall consider control systems whose state at any time t is described by an n-dimensional vector $x(t)$ which satisfies the ordinary differential equation

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

The solution of (1) depends on the choice of the function $u(t)$ which we shall refer to as the *control* function. This control function is also assumed to be a vector whose components we shall denote by $u^1(t), \cdots, u^r(t)$. Throughout this paper we shall assume that the $u^j(t)$ are bounded measurable functions, and that $A(t)$ and $B(t)$, which are $n \times n$ and $n \times r$ matrices, respectively, are continuous in t. We shall denote the columns of $B(t)$ by $b_j(t)$, $j = 1, \cdots, r$.

If $u^*(t)$ is an arbitrary (bounded and measurable) r-dimensional vector function defined for $t_0 \leqq t \leqq t_1$, the solution of (1) with $u(t) = u^*(t)$ exists on the interval $[t_0, t_1]$ for an arbitrary initial condition $x(t_0) = x_0$. (For ease of notation, let us henceforth suppose that $t_0 = 0$.) If $x(t)$ is this solution, and $x(T) = \bar{x}$ where $0 = t_0 \leqq T \leqq t_1$, we shall say that $u^*(t)$ *transfers* x *from* x_0 *to* $\bar{x}$ *in time* T.

Consider the following problem. Suppose that a "response time" T, $T > 0$, is given, and that a real valued function $\mathcal{E}$ is defined, where the domain of $\mathcal{E}$ is the set of all bounded, measurable, r-dimensional vector functions on $[0, T]$. We shall say that $\mathcal{E}(u(t))$ is the *effort* associated with $u(t)$. Then our problem consists of the following: Given an initial state x_0 and a desired final state x_1 (which may depend on T), find a control function $u^*(t)$ which transfers x from x_0 to x_1 in time T, and in so doing minimizes the control effort $\mathcal{E}$; i.e., if $\hat{u}(t)$ is any other control which transfers x from x_0 to x_1 in time T, $\mathcal{E}(\hat{u}(t)) \geqq \mathcal{E}(u^*(t))$. Then, $u^*(t)$ will be called the *minimum effort control*, and $\mathcal{E}(u^*(t))$ the *minimum effort*.

This problem may have no solution, either because there is no control

* Received by the editors April 6, 1962.

† Aerospace Corporation, El Segundo, California.

which transfers x from x_0 to x_1 in time T, or because the minimum of the values $\mathcal{E}(u(t))$ is not achieved (on the required set of $u(t)$). However, for the particular cases which we shall consider, neither of these difficulties arises.

Suppose that a solution to the problem exists for all response times T in some interval. In this case, let $f(T)$ be the function which expresses the minimum effort as a function of T in this interval. In most physical problems, short response times correspond to large efforts, so that a control system designer must "trade-off" the desirable features of rapid action and small effort. Knowledge of $f(T)$ will enable him to make an intelligent choice in designing a maximum effort capability into his system. Since x_0 and x_1 may not be known a priori, and the function f clearly depends on them, it is probably necessary to compute f for a representative set of initial and terminal states. The actual optimal control may be difficult to implement in practice, but even for a non-optimal control law, $f(T)$ provides a definite lower bound on the control effort required to perform a specific task.

In this paper we shall consider that $\mathcal{E}(u(t))$ is defined in one of two ways: either as the maximum control amplitude (see (4)), or as the integral of a certain function of the control (see (18)). We shall show that the minimum effort and minimum effort control can be obtained in both cases by solving a simple variational problem, in which it is required to find the minimum of a functional F of n variables. We shall show that F has continuous first partial derivatives which are easily evaluated. Furthermore, we shall show that F has no extrema other than the desired one, so that the method of steepest descent can be used to compute the minimum. Finally we shall generalize the problem to the case where the target point x_1 is replaced by an arbitrary convex set.

The variational formulation which we shall present and derive was first given by Krasovskii [1, 2]. However, Krasovskii did not exploit his result to derive a computational method. Furthermore, our derivation differs from Krasovskii's. We shall use some geometric arguments similar to the ones introduced by Bellman, Glicksberg, and Gross [3], and later used by LaSalle [4], Gamkrelidze [5], and others. Krasovskii took the viewpoint that $\mathcal{E}$ is a functional on an L^p space, and used some results of Krein on the so-called L-problem to derive the variational form.

We also note two other approaches to the problem in case $\mathcal{E}$ is given by an integral of the form $\int_0^T G(u(t))\,dt$, where G is a sufficiently smooth scalar-valued function. One is from the viewpoint of the classical calculus of variations, using undetermined Lagrange multipliers. This approach has been described by Desoer [7] in a slightly different context. The other

approach, using the Pontryagin maximum principle, has been described in [6]. In both these methods one is eventually faced with the problem of determining n constants—the Lagrange multipliers in the first case, and the initial conditions of an adjoint solution in the second. These constants are analogous to the coordinates of the vector η introduced below.

Krasovskii [1, 2] was interested in the time-optimal problem, in which a maximum value for $\mathcal{E}(u(t))$ is given, and it is desired to find a control $u^*(t)$ which transfers x from x_0 to x_1 (both points are given) in minimal time, subject to the maximum effort constraint. If the function $f(T)$ and the corresponding minimum effort controls $u^*(t)$ have been computed, the time-optimal problem is solved by noting the smallest value of T for which $f(T)$ is not greater than the prescribed maximum effort. A different, more efficient method of computing time-optimal controls has been described by the author [8]. In this method it is only necessary to carry out one maximization similar to the one described below, as a result of which both the minimum time and the time-optimal control are derived. Another method of computing time-optimal controls has been described by Ho [9]. This method is one of successive approximations and requires optimizations at various values (guesses) of the minimal time.

We note that the general solution to (1) with initial condition $x(0) = x_0$ is given by

$$x(t) = X(t)\left[x_0 + \int_0^t X^{-1}(s)B(s)u(s)\,ds\right], \tag{2}$$

where $X(t)$ is the matrix solution to the equation

$$\begin{aligned} \dot{X}(t) &= A(t)X(t), \\ X(0) &= I \quad \text{(the identity matrix)}, \end{aligned} \tag{3}$$

and $X^{-1}(t) = Y(t)$ satisfies the equation

$$\dot{Y}^* = -A^*Y^*, \qquad Y(0) = I,$$

where * denotes transpose.

II. EFFORT DEFINED BY MAXIMUM AMPLITUDE

Let us first consider the case where

$$\mathcal{E}(u(t)) = \max_{1 \leqq j \leqq r} \sup_{0 \leqq t \leqq T} |u^j(t)|. \tag{4}$$

Physically, this might correspond to the maximum thrust available in gas jets of a satellite attitude control system.

Suppose we are trying to "hit" a moving target $x_1(t)$, where the vector function $x_1(t)$ is continuous. Then, for a fixed time $T > 0$, we wish to find a control $u^*(t)$ which transfers x from x_0 to $x_1(T)$ in time T, and has the property that if $\hat{u}(t)$ is any other control which performs the same

transfer, then

$$\max_j \sup_{0 \leqq t \leqq T} | u^{*j}(t) | \leqq \max_j \sup_{0 \leqq t \leqq T} | \hat{u}^j(t) |.$$

Let us define the functions $y(t)$ and $g(t, \eta)$ by

$$y(t) = X^{-1}(t)x_1(t) - x_0, \tag{5}$$

$$g(t, \eta) = \eta \cdot X^{-1}(t)B(t), \tag{6}$$

where η is the n-dimensional (row) vector $(\eta^1, \cdots, \eta^n)$, $y(t)$ is an n-vector, and $g(t, \eta)$ is an r-dimensional row vector whose coordinates we shall denote by $g^j(t, \eta)$, $j = 1, \cdots, r$, so that $g^j = \eta \cdot X^{-1}b_j$. Let us use the notation [sgn $g(t, \eta)$] to indicate the r-dimensional column vector function[1] whose j-th component is sgn $g^j(t, \eta)$. Finally, let

$$F(\eta, T) = \sum_{j=1}^{r} \int_0^T | g^j(t, \eta) | \, dt. \tag{7}$$

We shall say that the control system is normal (following LaSalle [4]) if, for every fixed vector $\eta \neq 0$ and every $j = 1, \cdots, r$, the set of $t > 0$ for which $g^j(t, \eta)$ vanishes has measure zero. Then $F(\eta, T) > 0$ for every $\eta \neq 0$, $T > 0$.

THEOREM 1. *Given the normal control system defined by* (1), *the initial point* x_0, *the target point* $x_1(t)$, *the time* $T > 0$, *and the "effort" function defined by* (4), *then there exists a minimum effort control* $u^*(t)$ *which transfers* x *from* x_0 *to* $x_1(t)$ *in time* T. *If* $y(t)$ *defined by* (5) *does not vanish for* $t = T$, *the minimum effort* $\mathcal{E}_{\min} = \mathcal{E}(u^*(t))$ *is given by*

$$\frac{1}{\mathcal{E}_{\min}} = \min_{\eta \in P} F(\eta, T), \tag{8}$$

where F *is given by* (7) *and* P *is the plane* $\eta \cdot y(T) = 1$. *Furthermore, the minimum effort control is unique* (*to within a set of* t *of measure zero*), *and is given by*

$$u^*(t) = \mathcal{E}_{\min} \operatorname{sgn} g(t, \eta^*), \tag{9}$$

where η^* *is any vector in* P *for which the minimum in* (8) *is attained. Finally, if* $y(T) = 0$, *the control* $u^*(t) \equiv 0$ *is the desired minimum effort control.*

Proof. If $y(T) = 0$, (2) and (5) show that $u(t) \equiv 0$ transfers x from x_0 to $x_1(T)$ in time T. Suppose now that $y(T) \neq 0$. Define the set C_T by

$$C_T = \left\{ \int_0^T X^{-1}(t)B(t)u(t) \, dt; \right. \\ \left. u^j(t) \text{ measureable}, \, | u^j(t) | \leqq 1, j = 1, \cdots, r, 0 \leqq t \leqq T \right\}. \tag{10}$$

[1] This function is not well defined when $g^j(t, \eta) = 0$, but by the assumption which follows, the set of t for which this is true has measure zero.

It is easy to show that C_T is a convex, bounded set in n-space which is symmetric with respect to the origin. Also, C_T is closed (for example, see [4, Lemma 2]). Furthermore, C_T does not lie in any linear subspace of dimension less than n, so that it is a convex body. To show this, assume the contrary. Then there is a vector θ orthogonal to this subspace. Therefore, θ is orthogonal to C_T so that $\theta \cdot \zeta = 0$ for all $\zeta \in C_T$. But if

$$\zeta^* = \int_0^T X^{-1}(t)B(t) \operatorname{sgn} g(t, \theta)\, dt,$$

$\zeta^* \in C_T$ and $\theta \cdot \zeta^* = F(\theta, T)$. But $F(\theta, T) > 0$ by our normality assumption, and we have a contradiction. Thus, if ρ is any nonzero vector, there are vectors $\zeta_1, \cdots, \zeta_m$ of C_T and real numbers (not all zero) $\lambda_1, \cdots, \lambda_m$ such that $\sum_{i=1}^m \lambda_i \zeta_i = \rho$. Because of the symmetry of C_T, we may assume that $\lambda_i \geqq 0$, $i = 1, \cdots, m$. Since C_T is convex, $k\rho \in C_T$ where k is the positive number $\sum_{i=1}^m \lambda_i$. We shall use this fact below.

Let us denote the control function sgn $g(t, \eta)$ by $u(t, \eta)$, and define the point $z(T, \eta)$ of C_T for every $\eta \neq 0$ by

$$z(T, \eta) = \int_0^T X^{-1}(t)B(t)u(t, \eta)\, dt. \tag{11}$$

Note that $\eta \cdot z(T, \eta) = F(\eta, T)$. Furthermore, it is easy to show that

$$\eta \cdot z(T, \eta) > \eta \cdot \zeta, \quad \text{for all } \zeta \in C_T, \zeta \neq z(T, \eta). \tag{12}$$

Also, $u(t, \eta)$ gives the only (disregarding sets of measure zero) representation of $z(T, \eta)$ in the form (10). Thus, $z(T, \eta)$ is a boundary point of C_T, and η is the outward normal of a support plane to the set at this point (or, as we shall say, η is a normal to C_T at $z(T, \eta)$).

Now define

$$\alpha = \max_{\beta y(T) \in C_T} \beta.$$

Since $y(T) \neq 0$ and C_T is compact and contains the origin, α is defined, and by what was said earlier, $\alpha > 0$. It is clear that $\alpha y(T)$ is a boundary point of C_T. Let η^* be the normal to C_T at $\alpha y(T)$, so that $\eta^* \cdot \alpha y(T) \geqq \eta^* \cdot \zeta$ for all $\zeta \in C_T$. Then it is easy to see that $\alpha y(T) = z(T, \eta^*)$, so that $F(\eta^*, T) = \eta^* \cdot z(T, \eta^*) > \eta^* \cdot \zeta$ for all $\zeta \in C_T$ distinct from $\alpha y(T)$. Since $F(\eta^*, T) > 0$ and the length of η^* plays no role in our discussion, we may assume, without loss of generality, that $\eta^* \cdot y(T) = 1$. It readily follows from (2), (5), and (11) that the control $u^*(t) = \alpha^{-1} u(t, \eta^*)$ transfers x from x_0 to $x_1(T)$ in time T. Also, $\varepsilon(u^*(t)) = \alpha^{-1}$, so that $\varepsilon_{\min} \leqq \alpha^{-1}$.

Now suppose that $\hat{u}(t)$ is a control which transfers x from x_0 to $x_1(T)$ in time T. We shall show that $\varepsilon(\hat{u}(t)) \geqq \alpha^{-1}$, so that $\varepsilon_{\min} = \alpha^{-1}$. It follows

from (2) and (5) that

$$y(T) = \int_0^T X^{-1}(t)B(t)\hat{u}(t)\,dt,$$

and if we set $\bar{u}(t) = [\mathcal{E}(\hat{u}(t))]^{-1}\hat{u}(t)$ (so that $\mathcal{E}(\bar{u}(t)) = 1$), the vector

$$\frac{1}{\mathcal{E}(\hat{u}(t))}\,y(T) = \int_0^T X^{-1}(t)B(t)\bar{u}(t)\,dt \tag{13}$$

belongs to C_T. Then, by definition of α, $\mathcal{E}(\hat{u}(t)) \geqq \alpha^{-1}$.

Thus, $u^*(t)$ is indeed a minimum effort control. To show it is unique, suppose that $\hat{u}(t)$ is defined as above, and that $\mathcal{E}(\hat{u}(t)) = \alpha^{-1}$. Then the vector (13) is $\alpha y(T) = z(T, \eta^*)$, and because of the unique representation of this vector in the form (10), $\bar{u}(t) = u(t, \eta^*)$ almost everywhere. Hence, $u^*(t) = \hat{u}(t)$ a.e.

Let us verify (8) and show that η^* is a minimizing vector. By what has already been proved, if η is any vector in P (i.e., $\eta \cdot y(T) = 1$)

$$F(\eta, T) = \eta \cdot z(T, \eta) \geqq \eta \cdot \alpha y(T) = \alpha. \tag{14}$$

Furthermore, $\alpha y(T) = z(T, \eta^*)$, and $\eta^* \in P$, so that

$$\alpha = \eta^* \cdot z(T, \eta^*) = F(\eta^*, T). \tag{15}$$

Combining (14) and (15), we obtain the desired result

$$\frac{1}{\mathcal{E}_{\min}} = \alpha = \min_{\eta \in P} F(\eta, T) = F(\eta^*, T). \tag{16}$$

To complete the proof of the theorem we need only show that if η^{**} is any other vector in P at which (8) takes on its minimum, then $u(t, \eta^*) = u(t, \eta^{**})$. But suppose that $F(\eta^{**}, T) = F(\eta^*, T) = \alpha$, and that $\eta^{**} \cdot y(T) = 1$. Then $F(\eta^{**}, T) = \eta^{**} \cdot z(T, \eta^{**}) = \alpha = \eta^{**} \cdot \alpha y(T)$. Because of (12) it follows that $z(T, \eta^{**}) = \alpha y(T) = z(T, \eta^*)$. Since the representation of $z(T, \eta^*)$ in the form (10) is unique, $u(t, \eta^*) = u(t, \eta^{**})$ a.e.

Note that the vector $\eta^* \in P$ which realizes the minimum in (8) need not be unique. This corresponds to the geometric fact that there may be more than one support plane to C_T at $\alpha y(T)$. To each support plane at $\alpha y(T)$ there corresponds an η^*. For each of these η^* the minimum effort control $u^*(t)$ is the same. Let us denote the set of these minimizing vectors, which is a closed convex subset of P, by H.

We claim that the function $F(\eta, T)$ has no extrema on P away from the set H. It is shown in [8] that $F(\eta, T)$, as a function of $\eta^1, \cdots, \eta^n$, is in C^1, and that its gradient is given by

$$\nabla F(\eta) = z(T, \eta) \neq 0.$$

Thus, F has an extremum on P if and only if $\nabla F(\eta) = z(T, \eta) = Ky(T)$ for some constant K, since $y(T)$ is the normal to P. But $z(T, \eta)$ is a boundary point of C_T for every η, and the only multiples of $y(T)$ which are boundary points of C_T are $\pm\alpha y(T)$. However, if $z(T, \hat{\eta}) = -\alpha y(T)$, $\hat{\eta}$ cannot belong to P because $0 < \hat{\eta}\cdot z(T, \hat{\eta}) = -\alpha\hat{\eta}\cdot y(T)$, or $\hat{\eta}\cdot y(T) = -\alpha^{-1}\hat{\eta}\cdot z(T, \hat{\eta}) < 0$, and $\hat{\eta}\cdot y(T) \neq 1$. Therefore, if F has an extremum in P at η, $z(T, \eta) = \alpha y(T)$, which implies that $F(\eta, T) = \eta\cdot z(T, \eta) = \eta\cdot\alpha y(T) = \alpha$; i.e., $\eta \in H$.

Thus, consider finding the minimum of F on P by the method of steepest descent. To do so, we let η be a function of a parameter τ and solve the differential equation

$$\begin{aligned}\frac{d\eta}{d\tau} &= -\nabla F + \frac{[\nabla F\cdot y(T)]y(T)}{\| y(T) \|^2} \\ &\equiv -z(T, \eta) + \frac{[z(T, \eta)]\cdot[y(T)]}{\| y(T) \|^2}\, y(T),\end{aligned} \tag{17}$$

where the right-hand side is the component of ∇F in P. The terms in the right-hand side of (17) can be computed directly in terms of known quantities (see (11), (6), (5), and (3)). The term $\eta\cdot X^{-1}(t)$ which occurs in (6) may be computed as follows. Let $y(t, \eta) = X^{*-1}(t)\eta^T = Y^*\eta^T$ (where η^T denotes the transpose of η); then,

$$\frac{dy(t, \eta)}{dt} = -A^*(t)y(t, \eta), \qquad y(0, \eta) = \eta^T.$$

Equation (17) can be solved on an analog computer, provided that a sampling procedure is used in the computation of $z(T, \eta)$. (The latter cannot be computed instantaneously since an integration is required.)

Since $F \in C^1$, if the steepest descent method converges at all, it must converge to an extremum of F, which must be a point on the desired set H.

Theorem 1 essentially was first proved by Krasovskii [1]. The fact that a control function which transfers x from x_0 to x_1 in time T, for arbitrary x_0, x_1, and $T > 0$, does exist (which is an immediate consequence of Theorem 1) has already been shown by LaSalle [4, Theorem 6]. As LaSalle pointed out, one can even relax the assumption of normality in this existence proof.

Kulikowski [10] considered the same problem of minimizing the function F. To do so he assumed that the components of $X^{-1}(t)b_j(t)$ could be approximated by n-th order polynomials in t, in which case he could compute the minimizing polynomial explicitly. However, this polynomial approximation may in general be inadequate.

III. EFFORT DEFINED BY INTEGRALS

Now consider the case where effort is given by

$$\int_0^T \sum_{j=1}^{r} \lambda_j \mid u^j(t) \mid^p dt, \tag{18}$$

where each λ_j is positive, and $p > 1$. By redefining the $u^j(t)$ and the $b_j(t)$, we can assume that $\lambda_j = 1$ for $j = 1, \cdots, r$. It is also convenient to consider the $(1/p)$-th power in (18). This new definition obviously does not change the minimum effort control. Thus, in this section we shall consider that effort is defined by

$$\mathcal{E}(u(t)) = \left(\int_0^T \sum_{j=1}^{r} \mid u^j(t) \mid^p dt \right)^{1/p}, \tag{19}$$

where p is a fixed number greater than one.

Physically, the most interesting case is when $p = 2$, in which case effort may correspond to energy, power, etc. Furthermore, if $p = 2$, the term $\sum_j \lambda_j [u^j(t)]^2$ in (18) can be replaced by an arbitrary positive definite quadratic form, since a linear transformation on the u^j, which does not change the form of (1), will put the effort in the form of (19).

Let us define the functions $y(t)$, $g(t, \eta)$, $g^j(t, \eta)$, and sgn $g(t, \eta)$ as in section II, but let us redefine the function F by

$$F(\eta, T) = \left(\sum_{j=1}^{r} \int_0^T \mid g^j(t, \eta) \mid^q dt \right)^{1/q}, \tag{20}$$

where q is related to p by the relation $p^{-1} + q^{-1} = 1$. If the control system is normal, $F(\eta, T) > 0$ for all $\eta \neq 0$, $T > 0$.

We then have the following theorem.

THEOREM 2. *Given the normal control system defined by* (1), *the initial point* x_0, *the target point* $x_1(t)$, *the time* $T > 0$, *and the effort function defined by* (19), *then there exists a minimum effort control* $u^*(t)$ *which transfers* x *from* x_0 *to* $x_1(T)$ *in time* T. *If* $y(t)$ *defined by* (5) *does not vanish for* $t = T$, *the minimum effort* $\mathcal{E}_{\min} = \mathcal{E}(u^*(t))$ *is given by*

$$\frac{1}{\mathcal{E}_{\min}} = \min_{\eta \in P} F(\eta, T), \tag{21}$$

where F *is given by* (20), *and* P *is the plane* $\eta \cdot y(T) = 1$. *Furthermore, the minimum effort control is unique to within a set of* t *of measure zero and is given by*

$$u^{*j}(t) = \mu \mid g^j(t, \eta^*) \mid^{q/p} \operatorname{sgn} g^j(t, \eta^*), \tag{22}$$

where

$$\mu = \mathcal{E}_{\min} [F(\eta^*, T)]^{-q/p},$$

and η^* is any vector in P at which the minimum in (21) *is attained. Finally, if $y(T) = 0$, the control $u^*(t) \equiv 0$ is the desired minimum effort control.*

Proof. If $y(T) = 0$, it is clear that $u^*(t) \equiv 0$ is the desired control. Thus, we shall assume that $y(T) \neq 0$. Define the set C_T, which corresponds to the set C_T of Theorem 1, by

$$(23) \quad C_T = \left\{ \int_0^T X^{-1}(t)B(t)u(t)\,dt;\ u(t) \text{ measurable}, \mathcal{E}(u(t)) \leqq 1 \right\}.$$

Using the Hölder inequality and the boundedness of $X^{-1}(t)$ and $B(t)$, we easily conclude that C_T is bounded. The set is convex because the function $\varphi(u) = \varphi(u^1, \cdots, u^r) = \sum_{j=1}^r | u^j |^p$ is convex. It is also obvious that C_T is symmetric with respect to the origin. As in Theorem 1 (making obvious minor modifications) we can show that if ρ is any nonzero vector, there is a positive number k such that $k\rho \in C_T$.

Let us show that C_T is closed. Suppose that $\{x_n\}$ is a sequence of points in C_T, with $x_n \to x_\infty$ as $n \to \infty$, and suppose that x_n is represented in the form (23) by the function $u_n(t)$. We shall show that there is a measurable function $u_\infty(t)$ such that $\mathcal{E}(u_\infty(t)) \leqq 1$, and

$$(24) \qquad x_\infty = \int_0^T X^{-1}(t)B(t)u_\infty(t)\,dt,$$

which will prove that $x_\infty \in C_T$ and that C_T is closed. Let us consider the scalar functions $u_n{}^j(t)$ $(j = 1, \cdots, r; n = 1, 2, \cdots)$ as elements of $L^p(0, T)$. Clearly,

$$\int_0^T | u_n{}^j(t) |^p\,dt \leqq [\mathcal{E}(u_n(t))]^p \leqq 1$$

for all n and j, so that the functions $u_n{}^j(t)$ are uniformly bounded in the L^p norm. Hence, for each j, we can find functions $u_\infty{}^j(t)$ in L^p such that a subsequence of $\{u_n{}^j(t)\}_n$ converges weakly to $u_\infty{}^j(t)$ [11, p. 130]. Without loss of generality we shall assume that $u_n{}^j(t) \to u_\infty{}^j(t)$ as $n \to \infty$ weakly for each j. If we let $u_\infty(t)$ be the function with components $u_\infty{}^j(t)$, (24) follows from the definition of weak convergence. Thus we need only prove that $\mathcal{E}(u_\infty(t)) \leqq 1$. Let q be defined by $p^{-1} + q^{-1} = 1$, and let

$$A_n = \sum_{j=1}^r \int_0^T u_n{}^j(t) \,|\, u_\infty{}^j(t) \,|^{p/q} \operatorname{sgn} u_\infty{}^j(t)\,dt.$$

By definition of weak convergence, as $n \to \infty$,

$$A_n \to \sum_{j=1}^r \int_0^T | u_\infty{}^j(t) |^p\,dt = [\mathcal{E}(u_\infty(t))]^p,$$

where we have used the fact that $pq^{-1} + 1 = p$. But it follows from the

Hölder inequalities for sums and integrals and the fact that $\mathcal{E}(u_n(t)) \leqq 1$ for all n, that

$$A_n \leqq \int_0^T \sum_j |u_n{}^j(t)|\,|u_\infty{}^j(t)|^{p/q}\,dt \leqq \int_0^T \Big(\sum_j |u_n{}^j|^p\Big)^{1/p}\Big(\sum_j |u_\infty{}^j|^p\Big)^{1/q}dt$$
$$\leqq \left(\int_0^T \sum_j |u_n{}^j|^p\,dt\right)^{1/p}\left(\int_0^T \sum_j |u_\infty{}^j|^p\,dt\right)^{1/q} \leqq \left(\int_0^T \sum_j |u_\infty{}^j|^p\,dt\right)^{1/q}$$
$$= [\mathcal{E}(u_\infty(t))]^{p/q}.$$

Therefore,

$$[\mathcal{E}(u_\infty(t))]^p = \lim_{n\to\infty} A_n \leqq [\mathcal{E}(u_\infty(t))]^{p/q},$$

or, since $p - pq^{-1} = 1$, $\mathcal{E}(u_\infty(t)) \leqq 1$. Thus C_T is closed.

Now redefine the function $u(t, \eta)$ for every $\eta \neq 0$ by (see (20) and (6)),

$$u^j(t, \eta) = [F(\eta, T)]^{-q/p}\,|g^j(t, \eta)|^{q/p}\operatorname{sgn} g^j(t, \eta), \tag{25}$$

and define $z(T, \eta)$ as in (11), where $u(t, \eta)$ is given by (25). It is easily seen that $\mathcal{E}(u(t, \eta)) = 1$, so that $z(T, \eta) \in C_T$. Furthermore, as in Theorem 1, we have the identity $\eta \cdot z(T, \eta) = F(\eta, T)$. This can be verified immediately by direct substitution.

Let us prove that (12) is satisfied by every point ζ in C_T distinct from $z(T, \eta)$, and that the representation of $z(T, \eta)$ in the form (23) is unique. Thus, suppose that

$$\zeta = \int_0^T X^{-1}(t)B(t)\hat{u}(t)\,dt,$$

where

$$\sum_{j=1}^r \int_0^T |\hat{u}^j(t)|^p\,dt \leqq 1.$$

Then,

$$|\eta\cdot\zeta| \leqq \int_0^T \sum_{j=1}^r |g^j(t, \eta)\hat{u}^j(t)|\,dt$$
$$\leqq \int_0^T \Big(\sum_j |g^j|^q\Big)^{1/q}\Big(\sum_j |\hat{u}^j|^p\Big)^{1/p}dt$$
$$\leqq \left(\int_0^T \sum_j |g^j|^q\,dt\right)^{1/q}\left(\int_0^T \sum_j |\hat{u}^j|^p\,dt\right)^{1/p} \tag{26}$$
$$\leqq \left(\int_0^T \sum_j |g^j(t, \eta)|^q\,dt\right)^{1/q} = F(\eta, T) = \eta\cdot z(T, \eta).$$

The second inequality in (26) follows from the Hölder inequality for sums. Equality in the first and second inequalities holds if and only if

$$\hat{u}^j(t) = K \mid g^j(t, \eta)\mid^{q/p} \operatorname{sgn} g^j(t, \eta)$$

a.e., where K is a constant. If $\hat{u}(t)$ has this form, the third inequality, which follows from the Hölder inequality for integrals, becomes an equality. Equality holds in the last inequality if and only if $\varepsilon(\hat{u}(t)) = 1$. In summary, $\eta \cdot \zeta = \mid \eta \cdot \zeta \mid = \eta \cdot z(T, \eta)$ if and only if $\hat{u}(t) = u(t, \eta)$ a.e., i.e., if and only if $\zeta = z(T, \eta)$. Hence, $\eta \cdot \zeta < \eta \cdot z(T, \eta)$ if $\zeta \in C_T$, $\zeta \neq z(T, \eta)$, and $u(t, \eta)$ gives the only (disregarding sets of measure zero) representation of $z(T, \eta)$ in the form (23).

The remainder of the proof is an almost literal repetition of the corresponding part of the proof of Theorem 1 (beginning immediately after (12)), and we shall omit it.

COROLLARY. *The condition in Theorem 2 that the control system is normal can be replaced by the following hypothesis: For every vector* $\eta \neq 0$, *and every* $T > 0$, *the set of* $t \in [0, T]$ *for which* $g(t, \eta) \neq 0$ *has positive measure.*

Proof. The normality was used in both Theorems 1 and 2 to show that $F(\eta, T) > 0$, if η and T do not vanish. It is easily seen that the alternative hypothesis is sufficient to prove this relation. In Theorem 1 normality is necessary to prove (12) and the unique representation of $z(T, \eta)$. However, the same results can be obtained in Theorem 2 with the weaker hypothesis.

The case where $p = 1$ is of particular interest in applications. However, this case presents some difficulties which do not arise when $p > 1$. In the first place there is generally no minimum effort control in the strict sense. To realize such a control one must allow delta functions. Let $\delta(t)$ denote the unit delta function at $t = 0$.

Define

$$F(\eta, T) = \max_{0 \leqq t \leqq T} \max_{1 \leqq j \leqq r} \mid g^j(t, \eta)\mid,$$

and assume that for each j and each $\eta \neq 0$ there are a finite number of times $\tau_i{}^j \in [0, T]$ at which $\mid g^j(t, \eta)\mid = F(\eta, T)$. This number may depend on η, and may be zero. Of course, the values of $\tau_i{}^j$ depend on η.

Then, the minimum effort is again given by (21), and if $\eta^* \in P$ is a minimizing vector, the minimum effort control $u^*(t)$ is given by

$$u^{*j}(t) = \sum_i \mu_i{}^j \delta(t - \tau_i{}^j),$$

where the $\tau_i{}^j$ correspond to $\eta = \eta^*$, $\operatorname{sgn} \mu_i{}^j = \operatorname{sgn} g^j(\tau_i{}^j, \eta^*)$ or $\mu_i{}^j = 0$, and $\sum_{i,j} \mid \mu_i{}^j \mid = \varepsilon_{\min}$. If the set of $\tau_i{}^j$ is empty for some j, $u^{*j}(t) \equiv 0$.

A computational difficulty arises because η^* does not completely determine $u^*(t)$. Indeed, the $\tau_i{}^j$ are determined, but the $\mu_i{}^j$ in general are

not. Furthermore, the minimum effort control may even not be unique. In addition, the steepest descent method described below cannot be applied since $F(\eta, T)$ may not be differentiable.

The derivation of the above relations is very similar to the proof of Theorem 2, but since the relations do not serve our purpose, these derivations are omitted. Note that the corresponding set C_T is not closed, which accounts for the absence of true minimum effort controls.

Most of the results of Theorem 2, as well as the relations for the case $p = 1$, were first derived by Krasovskii [2].

Let us note that the method of steepest descent can be applied to the computation of the minimum described in Theorem 2. It is more convenient to deal with the function $G(\eta, T) = [F(\eta, T)]^q$, than with F. Clearly F and G have the same minima. Now,

$$\frac{\partial G}{\partial \eta^i} = \sum_{j=1}^{r} \frac{\partial}{\partial \eta^i} \int_0^T |\, g^j(t, \eta)\,|^q\, dt = \sum_{j=1}^{r} \int_0^T \frac{\partial}{\partial \eta^i} |\, g^j(t, \eta)\,|^q\, dt,$$

and

$$g^j(t, \eta) = \eta \cdot X^{-1}(t) b_j(t) = \sum_{k=1}^{n} \eta^k y_j{}^k(t),$$

where $y_j{}^k$ is the k-th component of $X^{-1}b_j$, i.e., $y_j{}^k(t) = e^k \cdot X^{-1}(t)b_j(t)$, where e^k is the k-th (row) coordinate vector. Then it is easily seen that

$$\frac{\partial}{\partial \eta^i} |\, g^j(t, \eta)\,|^q = q y_j{}^i(t)\, |\, g^j(t, \eta)\,|^{q-1} \operatorname{sgn} g^j(t, \eta),$$

so that

$$\begin{aligned} \frac{\partial G}{\partial \eta^i} &= q e^i \cdot \sum_{j=1}^{r} \int_0^T X^{-1}(t) b_j(t)\, |\, g^j(t, \eta)\,|^{q/p} \operatorname{sgn} g^j(t, \eta)\, dt \\ &= q[F(\eta, T)]^{q/p} e^i \cdot \int_0^T X^{-1}(t) B(t) u(t, \eta)\, dt \end{aligned}$$

(see (25)). Thus,

$$\nabla G = M z(T, \eta),$$

where $M = q[F(\eta, T)]^{q/p}$; i.e., as in section II, the minimum of G, and therefore of F, on P can be found by solving the equation (dropping the positive "gain factor" M)

$$\frac{d\eta}{d\tau} = -z(T, \eta) + \frac{z(T, \eta) \cdot y(T)}{\| y(T) \|^2}\, y(t). \tag{27}$$

Furthermore, the remarks that were made in section II with respect to uniqueness of extrema apply here as well.

IV. THE PROBLEM WHERE THE TERMINAL STATE IS A SET

Let us now consider the same problem described above, where the target point $x_1(t)$ is replaced by a fixed target set Ω. Thus, for each fixed time T, we wish to find the control $u^*(t)$ which transfers x from x_0 to some point of Ω in time T, and has the property that if $\hat{u}(t)$ is any other control which transfers x from x_0 to a point of Ω in time T, $\mathcal{E}(\hat{u}(t)) \geqq \mathcal{E}(u^*(t))$. The arguments of this section apply to the case where effort is defined either as in section II, or in section III (with $p > 1$). We shall assume that the set Ω is closed and convex. A problem in this formulation arises, for example, in the regulation of plants with numerator dynamics [12].

If Ω is compact, the minimum effort and minimum effort control may be computed in almost the exact same way as in the case when Ω consists of a single point. It is only necessary to replace the plane P in (8) and (21) by the set Q defined as follows.

Let Y_T be the set $X^{-1}(T)\Omega - x_0$. Then Q is defined as the set of all vectors η for which

$$\min_{y \in Y_T} \eta \cdot y = 1. \tag{28}$$

Thus, we have the following theorem.

THEOREM 3. *Given the normal control system defined by* (1), *the initial point* x_0, *the compact convex target set* Ω, *the time* $T > 0$, *and the effort function defined by* (4) *or* (19), *then there exists a minimum effort control* $u^*(t)$ *which transfers* x *from* x_0 *to* Ω *in time* T. *If the origin does not belong to the set* $Y_T = X^{-1}(T)\Omega - x_0$, *the minimum effort* $\mathcal{E}_{\min} = \mathcal{E}(u^*(t))$ *is given by*

$$\frac{1}{\mathcal{E}_{\min}} \min_{\eta \in Q} F(\eta, T), \tag{29}$$

where F *is given by* (7) *or* (20), *and* Q *is defined by* (28). *Furthermore, the minimum effort control is unique to within a set of* t *of measure zero and is given by* (9) *or* (22), *where* η^* *is any vector in* Q *for which the minimum in* (29) *is attained. Finally, if* $0 \in Y_T$, $u^*(t) \equiv 0$ *is the desired minimum effort control.*

Proof. If $0 \in Y_T$, it immediately follows from (2) and (5), and the definition of Y_T, that $u^*(t) \equiv 0$ transfers x from x_0 to Ω in time T. Therefore, assume that $0 \notin Y_T$.

Now consider the set C_T defined in Theorem 1 (or 2), and let $\Gamma = \{\beta : (\beta Y_T) \cap C_T \text{ is not empty}\}$. Since Ω is compact, Y_T is also compact, and since $0 \notin Y_T$ there is a neighborhood V of 0 which does not meet Y_T. Thus, there is a positive ϵ such that $\| y \| > \epsilon$ for all $y \in Y_T$. Since C_T is bounded, there is a number λ such that $\| \zeta \| < \lambda$ for all $\zeta \in C_T$. There-

fore, if $\beta \in \Gamma$, $\beta < \lambda/\epsilon$. Let α be the least upper bound of the numbers $\beta \in \Gamma$. Note that $\alpha > 0$. Let us show that $\alpha \in \Gamma$. Suppose that $\beta_n \to \alpha$, $\beta_n \in \Gamma$, and let y_n be a point of Y_T such that $\beta_n y_n \in C_T$. Since C_T is compact, there is a point $\zeta^* \in C_T$ such that a subsequence of $\{\beta_n y_n\}$ converges to ζ^*. We shall suppose that $\beta_n y_n \to \zeta^*$. Then, since Y_T is bounded, it is also true that $\alpha y_n \to \zeta^*$, or $y_n \to (\alpha^{-1}\zeta^*)$. But since Y_T is closed, $(\alpha^{-1}\zeta^*) \in Y_T$ or $\zeta^* \in (\alpha Y_T) \cap C_T$, i.e., $\alpha \in \Gamma$. Let $\zeta^* = \alpha y^*$ where $y^* \in Y_T$.

Note that αY_T cannot meet the interior of the convex body C_T. For if $\zeta = \alpha y$, where $y \in Y_T$, is an interior point of C_T, $(1 + \epsilon)\zeta = (1 + \epsilon)\alpha y$ also belongs to C_T for some $\epsilon > 0$, i.e., $\alpha' \in \Gamma$ where $\alpha' = (1 + \epsilon)\alpha > \alpha$, contradicting the definition of α. Thus, in particular, ζ^* is a boundary point of C_T (as well as of αY_T).

Since C_T is a convex body which meets the convex set αY_T only at boundary points of C_T, there is a plane which supports both αY_T and C_T at their common boundary point ζ^* (see, for example, [13, p. 142, problem 1]). In other words, there is a vector η^* such that $\eta^* \cdot \zeta \leqq \eta^* \cdot \zeta^* \leqq \eta^* \cdot \alpha y$ for all $\zeta \in C_T$ and all $y \in Y_T$. Then $\zeta^* = z(T, \eta^*)$ (defined either as in Theorems 1 or 2), and $\eta^* \cdot \zeta < \eta^* \cdot \zeta^*$ for all $\zeta \in C_T$ distinct from ζ^* (see (12)). As a result, we can conclude that ζ^* is the only point common to αY_T and C_T. Furthermore, we may assume, without loss of generality, that $\eta^* \cdot \zeta^* = \alpha$, so that $\alpha \leqq \alpha \eta^* \cdot y$ for all $y \in Y_T$, or $1 = \min_{y \epsilon Y_T} \eta^* \cdot y$, i.e., $\eta^* \in Q$.

We may now proceed as in Theorem 1, and show that $\mathcal{E}_{\min} = \alpha^{-1}$, and that the unique minimum effort control is given by $\alpha^{-1} u(t, \eta^*)$. To prove relation (29) we note that if η is any vector in Q, i.e., η satisfies (28), then

$$\eta \cdot z(T, \eta) \geqq \eta \cdot \zeta^* \geqq \inf_{y \epsilon Y_T} \eta \cdot \alpha y = \alpha,$$

because $\zeta^* \in (\alpha Y_T) \cap C_T$ (see (12)). Together with the relations $\eta^* \in Q$, $\eta^* \cdot \zeta^* = \alpha$, $\zeta^* = z(T, \eta^*)$, $\eta \cdot z(T, \eta) = F(\eta, T)$, and $\alpha^{-1} = \mathcal{E}_{\min}$, this proves (29), and shows that η^* is a minimizing vector. Finally, suppose that $\eta^{**} \in Q$ and that $F(\eta^{**}, T) = \alpha$, so that $\eta^{**} \cdot z(T, \eta^{**}) = \alpha$. Since $\eta^{**} \in Q$, and $\zeta^* = \alpha y^*$ where $y^* \in Y_T$, $\eta^{**} \cdot \zeta^* = \alpha \eta^{**} \cdot y^* \geqq \alpha$. But $\zeta^* \in C_T$, so that it follows from (12) that $z(T, \eta^*) = \zeta^* = z(T, \eta^{**})$. Since the representation of $z(T, \eta^*)$ in the form (10) (or (23)) is unique, $u(t, \eta^*) = u(t, \eta^{**})$ a.e., and the proof of Theorem 3 is complete.

COROLLARY. *If $\mathcal{E}$ is defined by* (19), *the normality assumption can be replaced by the condition in the corollary of Theorem* 2.

If the set Ω is unbounded (although convex and closed), it can be replaced by its intersection with some closed sphere of sufficiently large radius. This intersection Ω' will be compact and convex, and if the sphere

is large enough, the problem of attaining Ω' will be equivalent to attaining Ω, and Theorem 3 can be applied. In practice it should not be difficult to obtain a reasonable estimate for this radius.

The method of steepest descent may also be applied to finding the minimum in (29). In this case, one is interested in finding the minimum of F not on the plane P, but on the surface Q. If Q is piecewise smooth, corresponding to Ω sufficiently smooth, (27) may be used, except that $y(T)$, the normal to P, must be replaced by the normal to Q, which in general will vary with η. The ease with which this normal can be computed as a function of η depends on the nature of Ω.

For example, suppose that Ω is a convex polyhedron of dimension n or less. Then Y_T is also a convex polyhedron, say with vertices $y_1, \cdots, y_s$. It is easily seen that (28), which defines Q, is equivalent to

$$\min_{1 \leqq i \leqq s} \eta \cdot y_i = 1. \tag{30}$$

Let H_i be the half-space defined by $\{\eta : \eta \cdot y_i \geqq 1\}$, and let Z_i be the plane $\{\eta : \eta \cdot y_i \geqq 1\}$. If $Z = \bigcap_{i=1}^{s} H_i$, Z is an unbounded convex "polyhedron" whose faces lie in the H_i. It follows from (30) that Q is the boundary of Z, i.e., is made up of these plane faces. Then, in the steepest descent procedure, η will either vary within one face, or proceed from face to face, possibly moving along the intersection of two or more faces during a part of its motion. It can be shown, as in section II, that F has no local extrema on any of the faces of Q or on the intersections of these faces other than the desired one. Thus, if the steepest descent process converges, it tends to the desired minimum of F on Q.

REFERENCES

[1] N. N. Krasovskii, *On the theory of optimal control*, Avtomat. i Telmeh., 18 (1957), pp. 960–970. (English translation in Automation and Remote Control, vol. 18, pp. 1005–1016.)

[2] N. N. Krasovskii, *On the theory of optimal control*. Prikl. Mat. Meh., 23 (1959), pp. 625–639. (English translation in J. Appl. Math. Mech., vol. 23, pp. 899–919.)

[3] R. Bellman, I. Glicksberg and O. Gross, *On the "bang-bang" control problem*, Quart. Appl. Math., 14 (1956), pp. 11–18.

[4] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. 5, Princeton University Press, 1960, pp. 1–24.

[5] R. V. Gamkrelidze, *The theory of time-optimal processes in linear systems*, Izv. Akad. Nauk SSSR, Ser. Mat., 22 (1958), pp. 449–474. (English translation in report no. 61-7, University of California, Department of Engineering, Los Angeles, California.)

[6] V. G. Boltyanskii, R. V. Gamkrelidze and L. S. Pontryagin, *The theory of optimal processes*, Izv. Akad. Nauk SSSR, Ser. Mat., 24 (1960), pp. 3–42. (English translation in American Math. Society Translations, Series 2, vol. 18, 1961, pp. 341–382.)

[7] C. A. DESOER, *The bang bang servo problem treated by variational techniques*, Information and Control, 2 (1959), pp. 333–348.

[8] L. W. NEUSTADT, *Synthesizing time optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–493.

[9] Y. C. HO, *A successive approximation technique for optimal control systems subject to input saturation*, Trans. A.S.M.E. Ser. D., J. Basic Engrg., 84 (1962), pp. 33–40.

[10] R. KULIKOWSKI, *Synthesis of a class of optimum control systems*, Bull. Acad. Polon. Sci. Sér. Sci. Tech., 7 (1959), pp. 663–671.

[11] S. BANACH, *Théorie des opérations linéaires*, Hafner, New York.

[12] E. B. LEE, *On the time-optimal regulation of plants with numerator dynamics*, Trans. I.R.E. on Automatic Control (Correspondence), AC-6 (1961), pp. 351–352.

[13] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York 1958.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 1
Printed in U.S.A., 1962

VECTOR LYAPUNOV FUNCTIONS*

RICHARD BELLMAN†

1. Introduction. One of the most versatile techniques in the theory of nonlinear differential equations is the second method of Lyapunov. It depends crucially upon the fact that a function satisfying the scalar inequality

$$\frac{du}{dt} \leqq ku, \qquad u(0) = c, \tag{1.1}$$

is majorized by the solution of the equation

$$\frac{dv}{dt} = kv, \qquad v(0) = c. \tag{1.2}$$

It is natural to ask whether it might be more convenient in some circumstances to use a vector Lyapunov function rather than a scalar function. If it is, then we require an analogue of the foregoing majorization relation. It turns out that one exists; see [1, 2, 3]. We shall discuss it in the next section and give an application.

In the course of discussion of this result with J. P. LaSalle, he pointed out that J. K. Hale had independently arrived at the use of a vector Lyapunov function. His results will be presented elsewhere.

2. A lemma concerning nonnegative matrices. Let A be a constant matrix and e^{At} denote the matrix exponential, the solution of the matrix equation

$$\frac{dX}{dt} = AX, \qquad X(0) = I. \tag{2.1}$$

In order that all the elements of e^{At} be nonnegative for $t \geqq 0$, it is necessary and sufficient that $a_{ij} \geqq 0$, $i \neq j$. For various simple proofs of this useful result, see [2], [3].

From this we readily derive the following lemma.

LEMMA. *If* $a_{ij} \geqq 0$, $i \neq j$, *then*

$$\frac{dx}{dt} \leqq Ax, \qquad x(0) = c, \tag{2.2}$$

* Received by the editors May 29, 1962. This study was 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 Technology, Hq USAF. Views or conclusions contained herein should not be interpreted as representing the official opinion or policy of the United States Air Force.

† The RAND Corporation, Santa Monica, California.

implies $x \leqq y$ *where*

$$\frac{dy}{dt} = Ay, \qquad y(0) = c. \tag{2.3}$$

Here $x \leqq y$ implies component-by-component majorization.

3. An application. As an application of the foregoing result, suppose that u, v are two functions of t satisfying the inequalities

$$0 \leqq u \leqq k_1, \qquad 0 \leqq v \leqq k_2, \tag{3.1}$$

and the differential equations

$$\begin{aligned} \frac{du}{dt} &= -a_{11}u + a_{12}v + b_1 uv, \qquad u(0) = c_1, \\ \frac{dv}{dt} &= a_{21}u - a_{22}v + b_2 uv, \qquad v(0) = c_2, \end{aligned} \tag{3.2}$$

where $a_{ij} \geqq 0$, $b_1, b_2 \geqq 0$, $c_1, c_2 \geqq 0$. If the characteristic roots of

$$A = \begin{vmatrix} -a_{11} & a_{12} \\ a_{21} & -a_{22} \end{vmatrix} \tag{3.3}$$

have negative real parts and c_1, c_2 are small enough, then Poincaré-Lyapunov theory asserts that the solution of (3.2) approaches zero.

Using the lemma given above, we can obtain a nonlocal result. From (3.2) and (3.1) we have

$$\begin{aligned} \frac{du}{dt} &\leqq -a_{11}u + a_{12}v + b_1 k_2 u, \\ \frac{dv}{dt} &\leqq a_{21}u - a_{22}v + b_2 k_1 v, \end{aligned} \tag{3.4}$$

whence the solution is majorized by the solution of

$$\begin{aligned} \frac{dw}{dt} &= -a_{11}w + a_{12}z + b_1 k_2 w, \qquad w(0) = c_1, \\ \frac{dz}{dt} &= a_{21}w - a_{22}z + b_2 k_1 z, \qquad z(0) = c_2, \end{aligned} \tag{3.5}$$

i.e., $0 \leqq u \leqq w$, $0 \leqq v \leqq z$. The solutions of (3.5) approach zero as $t \to \infty$ if

$$B = \begin{vmatrix} -a_{11} + b_1 k_2 & a_{12} \\ a_{21} & -a_{22} + b_2 k_1 \end{vmatrix} \tag{3.6}$$

is a stability matrix.

It is easy to see how this method may be extended to treat systems of any order. Furthermore, it may be used for any system of equations for which a corresponding positivity relation holds. For some results of this type see [3] and the references given there.

4. Lyapunov functions. Given a vector system of differential equations

$$\begin{aligned} \frac{dx}{dt} &= Ax + By + g(x, y), \qquad x(0) = a, \\ \frac{dy}{dt} &= Cx + Dy + h(x, y), \qquad y(0) = b, \end{aligned} \tag{4.1}$$

where x and y are m and n dimensional vectors respectively, and A, B, C, D are matrices of appropriate dimensions, we form the nonnegative functions u and v by means of the quadratic forms

$$u = (x, Rx), \qquad v = (y, Sy), \tag{4.2}$$

where R and S are suitably chosen positive-definite matrices. Given a priori knowledge of the bounds on the components of x and y, we have numbers corresponding to k_1 and k_2, and can obtain various conditions for stability of the null solution.

In any given situation, we can group the equations in a number of different ways and obtain a number of sufficient conditions for stability of the equilibrium solution.

REFERENCES

[1] R. Bellman, I. Glicksberg and O. Gross, *On some variational problems occurring in the theory of dynamic programming*, Rend. Circ. Mat. Palermo, 3 (1954), pp. 1–35.

[2] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill Book Company, Inc., New York, 1960.

[3] E. F. Beckenbach and R. Bellman, *Inequalities*, Ergebnisse der Math., Springer, Berlin, 1961.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 1
Printed in U.S.A., 1962

EXTRACTION AND DETECTION PROBLEMS AND REPRODUCING KERNEL HILBERT SPACES*‡

EMANUEL PARZEN†

Summary. Problems involving the extraction, detection and prediction of signals in the presence of noise are among the central problems of statistical communication theory. Over the past few years I have sought to develop an approach to such problems which (i) would simultaneously apply to time series which are stationary or non-stationary, discrete parameter or continuous parameter, univariate or multivariate, and (ii) would distinguish between the statistical and analytical aspects of these problems, and in particular would clarify the role played by various widely employed analytical techniques (such as the Wiener-Hopf equation and eigenfunction expansions).

In developing this approach, two basic concepts are used: the notion of the probability density functional of a time series and the notion of a reproducing kernel Hilbert space. The aim of this paper is to sketch some of the main results which may be obtained by means of this approach.

In sections 1 and 2, it is shown how one may define and obtain a formula for the probability density functional of a normal time series. This formula is used to study the structure of optimum estimators (section 3) and detectors (section 4) by expressing them in a coordinate free way in terms of inner products in a reproducing kernel Hilbert space. Various ways of evaluating such inner products are discussed in sections 5 and 6. In section 7, it is shown how reproducing kernel Hilbert spaces provide a solution to the problems of minimum mean square error linear and non-linear prediction.

1. The probability density functional of a normal time series. Let $\{S(t),\ t \in T\}$ and $\{N(t),\ t \in T\}$ be time series, called respectively the signal process and the noise process. Let Ω be the space of all real valued functions on T. Let P_N and P_{S+N} be probability measures defined on the measurable subsets B of Ω by

$$P_N[B] = \text{Prob}\,[\{N(t),\ t \in T\} \in B] \tag{1.1}$$

$$P_{S+N}[B] = \text{Prob}\,[\{S(t) + N(t),\ t \in T\} \in B]. \tag{1.2}$$

We seek to determine, if it exists, a function p on Ω with the property that

$$P_{S+N}[B] = \int_B p\, dP_N\,. \tag{1.3}$$

The function p may be called the *probability density functional* of P_{S+N}

* Received by the editors June 24, 1962. Prepared under contract Nonr 3440(00) for the Office of Naval Research.

† Stanford University, Stanford, California.

‡ *Editorial note.* To study the control problem in a realistic setting, stochastic considerations are essential. This paper does not directly consider the control problem, but it does describe a recent area of mathematical development which is important to stochastic optimization and control. In particular, the results in linear and nonlinear prediction are immediately applicable to that class of optimal control problems where stochastic and optimization considerations can be separated.

with respect to P_N in order to emphasize that its argument is a function $\{X(t), t \in T\}$. It is also denoted $p(X(t), t \in T)$ and called the probability density functional of the signal plus noise process

$$X(t) = S(t) + N(t), \qquad t \in T, \tag{1.4}$$

with respect to the noise process $\{N(t), t \in T\}$ The function p is often written symbolically as a derivative,

$$p = \frac{dP_{S+N}}{dP_N} \tag{1.5}$$

and called the Radon-Nikodym derivative of P_{S+N} with respect to P_N (see [4], p. 329).

A necessary and sufficient condition that the probability density (1.5) exist is that P_{S+N} be *absolutely continuous* with respect to P_N in the sense that, for every measurable subset A of Ω,

$$P_N[A] = 0 \quad \text{implies} \quad P_{S+N}[A] = 0. \tag{1.6}$$

In order that P_{S+N} *not* be absolutely continuous with respect to P_N it is necessary and sufficient that there exist a set A such that

$$P_N[A] = 0 \quad \text{and} \quad P_{S+N}[A] > 0. \tag{1.7}$$

The probability measures P_N and P_{S+N} are said to be *orthogonal* if there exists a set A such that

$$P_N[A] = 0 \quad \text{and} \quad P_{S+N}[A] = 1. \tag{1.8}$$

One can regard (1.8) as the extreme case of not being absolutely continuous.

The notion of orthogonality derives its importance from detection theory (the theory of testing hypotheses). The simple hypotheses

$$H_0: X(\cdot) = N(\cdot)$$
$$H_1: X(\cdot) = S(\cdot) + N(\cdot)$$

are said to be *perfectly detectable* if there exists a set A such that

$$\begin{aligned} P_N[A] &= \text{Prob}\,[\{X(t), t \in T\} \in A \mid H_0] = 0, \\ P_{S+N}[A] &= \text{Prob}\,[\{X(t), t \in T\} \in A \mid H_1] = 1. \end{aligned} \tag{1.9}$$

Clearly, the hypotheses H_0 and H_1 are perfectly detectable if and only if P_N and P_{S+N} are orthogonal.

Given the probability measures P_N and P_{S+N}, the following questions arise:

(i) to determine whether P_N and P_{S+N} are orthogonal;

(ii) to determine whether P_{S+N} is absolutely continuous with respect to P_N;

(iii) to determine the Radon-Nikodym derivative (1.5) if it exists.

To answer these questions, the natural way to proceed is to approximate the infinite dimensional case by finite dimensional cases. For any finite subset

$$T' = \{t_1, \cdots, t_n\} \quad \text{of} \quad T, \tag{1.10}$$

let $P_{N,T'}$ and $P_{S+N,T'}$ denote the probability distributions of $\{X(t), t \in T'\}$ under P_N and P_{S+N} respectively. *Assume that* $P_{S+N,T'}$ *is absolutely continuous with respect to* $P_{N,T'}$ with Radon-Nikodym derivative denoted

$$p_{T'} = \frac{dP_{S+N,T'}}{dP_{N,T'}}. \tag{1.11}$$

The *divergence* between P_{S+N} and P_N on the basis of having observed $\{X(t), t \in T'\}$ is defined by

$$\begin{aligned} J_{T'} &= \mathrm{E}_{S+N}[\log p_{T'}] - \mathrm{E}_N[\log p_{T'}] \\ &= \int_\Omega \log p_{T'}\, dP_{S+N} - \int_\Omega \log p_{T'}\, dP_N. \end{aligned} \tag{1.12}$$

(The intuitive meaning of the divergence is described in section 4; see (4.8).) Using the theory of martingales it may be shown that

$$0 \leqq J_{T'} \leqq J_{T''} \qquad \text{if} \quad T' \subset T''. \tag{1.13}$$

Consequently, the limit

$$J_T = \lim_{T' \to T} J_{T'} \tag{1.14}$$

exists and is finite or infinite. Further, it may be shown [3] that

(i) if $J_T < \infty$, then P_{S+N} is absolutely continuous with respect to P_N and

$$p = \frac{dP_{S+N}}{dP_N} = \lim_{T' \to T} p_{T'}; \tag{1.15}$$

(ii) if $J_T = \infty$, and both the time series $\{N(t), t \in T\}$ and $\{S(t), t \in T\}$ are normal, then P_{S+N} and P_N are orthogonal.

We next apply these criteria under the following assumptions.

The noise process $\{N(t), t \in T\}$ is a normal process with zero means and covariance kernel

$$K(s, t) = \mathrm{E}[N(s)N(t)] \tag{1.16}$$

which is positive definite in the sense that for every finite subset T'

$= \{t_1, \cdots, t_n\}$ of T, the covariance matrix

$$K_{T'} = \{K(t_i, t_j)\} = \begin{Vmatrix} K(t_1, t_1) & \cdots & K(t_1, t_n) \\ \vdots & & \vdots \\ K(t_n, t_1) & \cdots & K(t_n, t_n) \end{Vmatrix} \tag{1.17}$$

is non-singular, with inverse matrix denoted

$$K_{T'}^{-1} = \{K^{-1}(t_i, t_j)\}. \tag{1.18}$$

(It should be noted that the assumption of positive definiteness is only made for mathematical convenience in the present exposition; it can be omitted.)

In regard to the signal process, two cases are of most interest:

(i) *Sure signal case.* $\{S(t), t \in T\}$ is a non-random function.

(ii) *Stochastic signal case.* $\{S(t), t \in T\}$ is a normal time series, independent of the noise process, with zero means and positive definite covariance kernel

$$R(s, t) = \mathrm{E}[S(s)S(t)]. \tag{1.19}$$

To employ the criterion (1.15), we first need to compute the divergence $J_{T'}$, defined by (1.12). In this paper we consider explicitly only the sure signal case; the stochastic signal case is considered in [10].

In the sure signal case, one can show that

$$\log p_{T'} = (X, S)_{K,T'} - (S, S)_{K,T'} \tag{1.20}$$

where we define for any functions f and g on T

$$(f, g)_{K,T'} = \sum_{s,t \in T'} f(s)K^{-1}(s, t)g(t). \tag{1.21}$$

Consequently

$$J_{T'} = \mathrm{E}_{S+N}[(X, S)_{K,T'}] - \mathrm{E}_N[(X, S)_{K,T'}] = (S, S)_{K,T'}, \tag{1.22}$$

and

$$J_T < \infty \quad \text{if and only if} \quad \lim_{T' \to T} (S, S)_{K,T'} < \infty. \tag{1.23}$$

In words, in the sure signal case, P_{S+N} is absolutely continuous with respect to P_N if and only if $(S, S)_{K,T'}$ approaches a limit as T' tends to T. Fortunately it is possible to characterize those functions $S(\cdot)$ which have this property. To do this, we introduce the notion of a reproducing kernel Hilbert space.

2. Reproducing kernel Hilbert spaces. Let $K(s, t)$ be the covariance kernel of a time series $\{X(t), t \in T\}$. For each t in T, let $K(\cdot, t)$ be the

function on T whose value at s in T is equal to $K(s, t)$. It may be shown (see Aronszajn [1]) that there exists a unique Hilbert space, denoted $H(K; T)$ or $H(K)$, with the following properties:

(i) the members of $H(K; T)$ are real valued functions on T;

(ii) for every t in T,

$$K(\cdot, t) \in H(K; T); \tag{I}$$

(iii) for every t in T and f in $H(K; T)$

$$f(t) = (f, K(\cdot, t))_{K,T} \tag{II}$$

where the inner product between two functions f and g in $H(K; T)$ is written $(f, g)_{K,T}$ or, for brevity, $(f, g)_K$.

Example 2A. Suppose $T = [1, 2, \cdots, n]$ for some positive integer n, and that the covariance kernel K is given by a symmetric positive definite matrix $\{K_{ij}\}$ with inverse $\{K^{ij}\}$. The corresponding reproducing kernel space $H(K; T)$ consists of all n-dimensional vectors $f = (f_1, \cdots, f_n)$ with inner product

$$(f, g)_{K,T} = \sum_{s,t=1}^{n} f_s K^{st} g_t . \tag{2.1}$$

To prove (2.1) one need only verify that the reproducing property holds: for $u = 1, \cdots, n$,

$$(f, K_{\cdot u})_{K,T} = \sum_{s,t=1}^{n} f_s K^{st} K_{tu} = \sum_{s=1}^{n} f_s \delta(s, u) = f_u .$$

The inner product may also be written as a ratio of determinants:

$$(f, g)_{K,T} = -\begin{vmatrix} K_{11} & \cdots & K_{1n} & f_1 \\ \vdots & \cdots & \vdots & \vdots \\ K_{n1} & \cdots & K_{nn} & f_n \\ g_1 & \cdots & g_n & 0 \end{vmatrix} \div \begin{vmatrix} K_{11} & \cdots & K_{1n} \\ \vdots & & \vdots \\ K_{n1} & \cdots & K_{nn} \end{vmatrix} . \tag{2.2}$$

To prove (2.2) one again need only verify the reproducing property. In the case in which the covariance matrix K is singular, one may define the corresponding reproducing kernel inner product in terms of the *pseudo-inverse* of the matrix K.

Example 2B. Let $T = \{t: a \leqq t \leqq b\}$ and let $\{N(t), a \leqq t \leqq b\}$ be the Wiener process; that is, it has independent increments and covariance function

$$K(s, t) = \sigma^2 \min (s, t) \tag{2.3}$$

for some parameter σ^2. Consider the Hilbert space $H(K; T)$ consisting

of all functions f on $a \leqq t \leqq b$ of the form

$$f(t) = f(a) + \int_a^t f'(u)\, du \tag{2.4}$$

for some square integrable measurable function f' on $a \leqq t \leqq b$ (which can be called the L_2-derivative of f), with inner product defined by

$$(f, g)_{K,T} = \frac{1}{\sigma^2}\left\{\frac{1}{a} f(a)g(a) + \int_a^b f'(u)g'(u)\, du\right\}. \tag{2.5}$$

If one defines

$$\begin{aligned} I_t(u) &= 1 \quad \text{if} \quad a \leqq u \leqq t, \\ &= 0 \quad \text{if} \quad t < u \leqq b, \end{aligned} \tag{2.6}$$

one may rewrite (2.4)

$$f(t) = f(a) + \int_a^b f'(a) I_t(u)\, du.$$

Now the covariance kernel $K(s, t)$ may be represented

$$K(s, t) = \sigma^2 a + \sigma^2 \int_a^b I_s(u) I_t(u)\, du.$$

Therefore, for each t in T, $K(\cdot, t)$ belongs to $H(K; T)$ with L_2 derivative

$$\frac{d}{ds} K(s, t) = I_t(s).$$

Further

$$\begin{aligned} (f, K(\cdot, t))_{K,T} &= \frac{1}{\sigma^2}\left\{\frac{1}{a} f(a)\sigma^2 a + \int_a^b f'(u)\sigma^2 I_t(u)\, du\right\} \\ &= f(a) + \int_a^t f'(u)\, du = f(t). \end{aligned}$$

We thus see that the reproducing kernel Hilbert space corresponding to the covariance kernel (2.3) consists of all L_2-differentiable functions on T with inner product given by (2.5).

The relevance of the theory of reproducing kernel Hilbert spaces to the theory of probability density functionals derives from the following fact: it may be shown (using martingale theory) that

$$\lim_{T' \to T} (S, S)_{K,T'} < \infty \quad \text{if and only if} \quad S \in H(K; T). \tag{2.7}$$

Further, if $S \in H(K; T)$, then

$$\lim_{T' \to T} (S, S)_{K,T'} = (S, S)_{K,T}. \tag{2.8}$$

It follows that, in the sure signal case, P_{S+N} is absolutely continuous with respect to P_N if and only if the signal function $\{S(t),\ t \in T\}$ belongs to the reproducing kernel Hilbert space $H(K;\ T)$ corresponding to the covariance kernel K of the noise process $\{X(t),\ t \in T\}$. If $S \in H(K;\ T)$, then the probability density functional is given by

$$(2.9) \qquad p(X(t), t \in T) = \frac{dP_{S+N}}{dP_N} = \exp\left[(X, S)_{K,T} - \frac{1}{2}(S, S)_{K,T}\right]$$

where by $(X,\ S)_{K,T}$ we mean the limit (in the sense both of convergence with probability one and convergence in quadratic mean)

$$(2.10) \qquad (X, S)_{K,T} = \lim_{T' \to T} (X, S)_{K,T'}.$$

It should be emphasized that although we use inner product notation to write $(X,\ S)_{K,T}$, this is not a true inner product between two elements in a Hilbert space, since the sample function $\{X(t),\ t \in T\}$ does not belong to $H(K)$; that is,

$$\lim_{T' \to T} (X, X)_{K,T'} \quad \text{is infinite with probability one.}$$

In practice, it will be clear how to define $(X,\ S)_{K,T}$ by suitably modifying the expression for the inner product between two functions in $H(K)$. Thus, for the covariance kernel given by (2.3), instead of the expression

$$(X, S)_{K,T} = \frac{1}{\sigma^2}\left\{\frac{1}{a}X(a)S(a) + \int_a^b S'(u)X'(u)\ du\right\}$$

suggested by (2.5) one may show that

$$(X, S)_{K,T} = \frac{1}{\sigma^2}\left\{\frac{1}{a}X(a)S(a) + \int_a^b S'(u)\ dX(u)\right\}.$$

There are a variety of ways in which one can determine whether a function S belongs to a reproducing kernel Hilbert space $H(K;\ T)$ and compute the norm $(S,\ S)_{K,T}$ and the random variable $(X,\ S)_{K,T}$. These are discussed in section 5. However one general principle deserves to be stated at this point.

Roughly speaking, a function $g(\cdot)$ belongs to a reproducing kernel Hilbert space $H(K;\ T)$ only if it is at least as "smooth" or "regular" as the functions $K(\cdot,\ t)$, since every function g in $H(K;\ T)$ is either a linear combination

$$g(\cdot) = \sum_{i=1}^{n} c_i K(\cdot,\ t_i)$$

or a limit of such linear combinations. For example, if T is an interval and K is continuous on $T \otimes T$, then every function in $H(K;\ T)$ is continuous;

if K is twice differentiable on $T \otimes T$, then every function in $H(K; T)$ is differentiable.

We are thus led to the following heuristic conclusion: *in order that a signal not be perfectly detectable in the presence of a noise, it is necessary and sufficient that the signal be as "smooth" or as "regular" as the noise.* In the case of a sure signal, the signal is as smooth as the noise if and only if $S \in H(K; T)$ where K is the covariance kernel of the noise. In the case of stochastic signals, the signal is as smooth as the noise if $S \in H(K; T)$ for almost all sample functions of the signal process: a rigorous formulation of this assertion is given in [10].

3. The structure of optimum extractors. In sections 3 and 4 we show how the formula for the probability density functional given by (2.9) may be used to study the structure of optimum extractors and detectors.

To begin with we consider a time series $\{X(t), t \in T\}$ satisfying the model

$$X(t) = \theta g(t) + N(t) \tag{3.1}$$

where

(i) θ is a parameter varying in a known set Φ;

(ii) $\{g(t), t \in T\}$ is a known non-random function;

(iii) the noise process $\{N(t), t \in T\}$ is normal with zero means and known covariance kernel K.

We call $g(\cdot)$ a regression function, and call the signal process

$$S(t) = \theta g(t), \tag{3.2}$$

a *signal of regression type* with a one-dimensional parameter. We assume that $g(\cdot) \in H(K; T)$.

We let $p(X(t), t \in T \mid \theta)$ denote the probability density functional of the signal plus noise process $\{\theta g(t) + N(t), t \in T\}$ with respect to the noise process $\{N(t), t \in T\}$. By (2.9) it follows that

$$p(X(t), t \in T \mid \theta) = \exp [\theta(X, g)_{K,T} - \tfrac{1}{2} \theta^2 (g, g)_{K,T}]. \tag{3.3}$$

For ease of writing in the sequel, we employ the notation

$$V = (X, g)_{K,T}, \qquad G = (g, g)_{K,T} \tag{3.4}$$

and consequently may write

$$p(X(t), t \in T \mid \theta) = \exp [\theta V - \tfrac{1}{2} \theta^2 G]. \tag{3.3}'$$

Estimation of θ. Suppose that the parameter θ is assumed to be a random variable with probability density function $p(\theta)$ with respect to a measure μ on the measurable subsets of the parameter space Φ. Then the condi-

tional probability density $p(\theta \mid X(t),\ t \in T)$ of θ, given $\{X(t),\ t \in T\}$, can be determined by the formula (usually called Bayes' rule)

$$p(\theta \mid X(t), t \in T) = \frac{p(X(t), t \in T \mid \theta)p(\theta)}{\int_\Phi p(X(t), t \in T \mid \theta)p(\theta)\, d\mu}. \tag{3.5}$$

For the sake of brevity in writing, one refers to the conditional probability density $p(\theta \mid X(t),\ t \in T)$ as the posterior density of θ, and refers to the (marginal) probability density $p(\theta)$ as the prior density of θ.

As with any distribution, the mean and variance of the posterior distribution of θ represent rough measures of the center and spread of the distribution; the mean of the posterior distribution (or conditional mean of θ, given $\{X(t),\ t \in T\}$) is denoted by

$$\theta^* = \mathrm{E}[\theta \mid X(t), t \in T] = \int_\Phi \theta\, p(\theta \mid X(t), t \in T)\, d\mu. \tag{3.6}$$

The variance of the posterior distribution (or conditional variance of θ, given $\{X(t),\ t \in T\}$) is denoted by

$$\mathrm{Var}\,[\theta \mid X(t), t \in T] = \int_\Phi (\theta - \theta^*)^2 p(\theta \mid X(t), t \in T)\, d\mu. \tag{3.7}$$

It may happen that one desires a point estimate of θ. The conditional mean θ^* is often regarded as an optimum point estimate since it has the property that it is the minimum mean square error estimate of θ in the sense that

$$\mathrm{E}[|\, \theta - \theta^* \,|^2] \leqq \mathrm{E}[|\, \theta - \varphi(X(t), t \in T) \,|^2] \tag{3.8}$$

for any functional $\varphi(X(t),\ t \in T)$ on the sample space with finite second moment. The mean square estimation error $\mathrm{E}[|\, \theta - \theta^* \,|^2]$ is given by the mean of the conditional variance,

$$\mathrm{E}[|\, \theta - \theta^* \,|^2] = \mathrm{E}[\mathrm{Var}\,[\theta \mid X(t), t \in T]]. \tag{3.9}$$

We shall refer to the conditional mean θ^* as the *Bayes* estimate of θ.

To illustrate these considerations, let us assume that the prior probability density of θ is given by

$$p(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\theta - \bar{\theta})^2\right] \tag{3.10}$$

for some known constants $\bar{\theta}$ and σ^2; in words, θ is normally distributed with mean $\bar{\theta}$ and variance σ^2. The posterior density of θ, given $\{X(t),$

$t \in T\}$, is then given by

$$p(\theta \mid X(t), t \in T) = \frac{\exp[\alpha\theta - \frac{1}{2}\gamma\theta^2]}{\int_{-\infty}^{\infty} \exp\left[\alpha\theta - \frac{1}{2}\gamma\theta^2\right] d\theta} \tag{3.11}$$

$$= \frac{1}{\sqrt{2\pi\gamma}} \exp\left[-\frac{1}{2\gamma}\left(\theta - \frac{\alpha}{\gamma}\right)^2\right]$$

where

$$\alpha = V + \sigma^{-2}\bar{\theta}, \qquad \gamma = G + \sigma^{-2}. \tag{3.12}$$

In words, the conditional distribution of θ, given $\{X(t), t \in T\}$, is normal with mean

$$\mathrm{E}[\theta \mid X(t), t \in T] = \frac{\alpha}{\gamma} \tag{3.13}$$

and variance

$$\mathrm{Var}[\theta \mid X(t), t \in T] = \frac{1}{\gamma}. \tag{3.14}$$

The Bayes estimate θ^* of θ is given by the conditional mean

$$\theta^* = \mathrm{E}[\theta \mid X(t), t \in T] = \frac{V + \sigma^{-2}\bar{\theta}}{G + \sigma^{-2}} \tag{3.15}$$

with mean square estimation error given by

$$\mathrm{E}[|\theta - \theta^*|^2] = \mathrm{E}[\mathrm{Var}[\theta \mid X(t), t \in T]] = \{G + \sigma^{-2}\}^{-1}. \tag{3.16}$$

In order to understand the meaning of (3.13) and (3.14) it should be noted that the variance σ^2 of a normally distributed random variable θ is a measure of the length of the interval in which an observed value of the random variable may be expected to lie (for example, with probability exceeding 0.99, θ will lie in the interval of length 6σ centered at the mean of θ). Now in considering the problem of extracting a signal from noise, we may be very uncertain about the possible range of values of the parameter θ. This corresponds to assuming that the prior variance σ^2 is very large (tending to infinity). Now as $\sigma^2 \to \infty$

$$\theta^* \to G^{-1}V, \tag{3.17}$$

$$\mathrm{E}[|\theta - \theta^*|^2] \to G^{-1}. \tag{3.18}$$

Thus if the prior variance σ^2 of the parameter θ is very large, the Bayes estimate is approximately given by the estimate θ^{**} defined by

$$\theta^{**} = G^{-1}V \tag{3.19}$$

with mean square estimation error

$$\mathrm{E}[|\theta - \theta^{**}|^2] = G^{-1}. \tag{3.20}$$

It is easy to verify that θ^{**} is the *maximum likelihood estimate* in the sense that

$$p(X(t), t \in T \mid \theta^{**}) = \max_{-\infty<\theta<\infty} p(X(t), t \in T \mid \theta). \tag{3.21}$$

It may also be shown [8] that θ^{**} is the minimum variance unbiased estimate and the minimum variance unbiased linear estimate (and that θ^* is the Bayes linear estimate).

We next rewrite the Bayes estimate in a way which is very useful for applications. Define

$$\begin{aligned} \bar{I} &= \{\mathrm{E}[|\theta - \bar{\theta}|^2]\}^{-1} = \sigma^{-2}, \\ I^{**} &= \{\mathrm{E}[|\theta - \theta^{**}|^2]\}^{-1} = G, \\ I^{*} &= \{\mathrm{E}[|\theta - \theta^{*}|^2]\}^{-1}. \end{aligned} \tag{3.22}$$

Intuitively, $\bar{I}$, I^{**}, and I^* represent the "information" contained respectively in the prior estimate $\bar{\theta}$, the maximum likelihood estimate θ^{**}, and the Bayes estimate θ^*. The Bayes estimate may be written

$$\begin{aligned} \theta^* &= (I^*)^{-1}\{I^{**}\theta^{**} + \bar{I}\bar{\theta}\}, \\ I^* &= I^{**} + \bar{I}. \end{aligned} \tag{3.23}$$

In words, the Bayes estimate is a weighted average of the maximum likelihood estimate and of the prior estimate, the weights being proportional to the "information" in each estimate.

One important application of (3.23) is to the problems of satellite orbit tracking and fitting a trend line to economic time series. In these problems one is called upon to publish a succession

$$\theta_1^*, \theta_2^*, \cdots$$

of estimates of a parameter θ at various times

$$T_1, T_2, \cdots.$$

One way of doing this is to let θ_n^{**} be the maximum likelihood estimate of θ based on the data which has become available in the interval T_{n-1} to T_n. As the Bayes estimate of θ_n^* at time T_n one takes

$$\begin{aligned} \theta_n^* &= \frac{I_n^{**}\theta_n^{**} + I_{n-1}^{*}\theta_{n-1}^{*}}{I_n^{**} + I_{n-1}^{*}}, \\ I_n^* &= I_n^{**} + I_{n-1}^{*}. \end{aligned} \tag{3.24}$$

The foregoing results were all derived assuming a one-parameter model of the form of (3.1). They may be readily extended to a multiparameter model of the form

$$X(t) = \sum_{j=1}^{k} \theta_j g_j(t) + N(t) \tag{3.25}$$

where for $j = 1, 2, \cdots, k$, θ_j is a parameter varying in $-\infty < \theta_j < \infty$, and $\{g_j(t), t \in T\}$ is a known function belonging to $H(K; T)$. Define

$$V_j = (X, g)_K, \qquad G_{ij} = (g_i, g_j)_K, \tag{3.26}$$

$$\theta = \begin{Vmatrix} \theta_1 \\ \vdots \\ \theta_k \end{Vmatrix}, \quad V = \begin{Vmatrix} V_1 \\ \vdots \\ V_k \end{Vmatrix}, \quad G = \begin{Vmatrix} G_{11} & \cdots & G_{1k} \\ \vdots & & \vdots \\ G_{k1} & \cdots & G_{kk} \end{Vmatrix}. \tag{3.27}$$

The probability density function of the time series (signal plus noise process) with respect to the noise process is given by

$$p(X(t), t \in T \mid \theta) = \exp [\theta^{\mathrm{tr}} V - \tfrac{1}{2} \theta^{\mathrm{tr}} G \theta] \tag{3.28}$$

where we write "tr" to denote the transpose of a vector or matrix. The maximum likelihood estimate θ^{**} of θ is given by (compare (3.19))

$$\theta^{**} = G^{-1} V \tag{3.29}$$

with mean square estimation error matrix

$$\mathrm{E}[(\theta - \theta^{**})(\theta - \theta^{**})^{\mathrm{tr}}] = G^{-1}, \tag{3.30}$$

assuming G is a non-singular matrix. Define

$$I^{**} = G \tag{3.31}$$

to be the "information" matrix of the maximum likelihood estimate θ^{**}. If the vector parameter θ possesses a prior distribution which is multivariate normal with mean

$$\mathrm{E}[\theta] = \bar{\theta} \tag{3.32}$$

and non-singular covariance matrix

$$\mathrm{E}[(\theta - \bar{\theta})(\theta - \bar{\theta})^{\mathrm{tr}}] = \bar{I}^{-1}, \tag{3.33}$$

then the Bayes estimate θ^* of θ is given by

$$\theta^* = (I^{**} + \bar{I})^{-1}(I^{**}\theta^{**} + \bar{I}\bar{\theta}) \tag{3.34}$$

with mean square estimation error matrix

$$\mathrm{E}[(\theta - \theta^*)(\theta - \theta^*)^{\mathrm{tr}}] = (I^{**} + \bar{I})^{-1}. \tag{3.35}$$

4. The structure of optimum detectors. The basic concepts involved in determining "optimum" detection systems are best introduced by considering the following four detection problems:

I Detecting the presence of a signal of specified shape.

II Detecting the presence of a signal of regression type.

III Detecting the presence of a stochastic signal.

IV Classifying (or decoding) signals.

Some of the results discussed in this section are related to recent work of Kailath [5] and Turin [15].

Case I. Detecting the presence of a signal of specified shape. Given an observed time series $\{X(t), t \in T\}$ one desires to test the simple hypothesis

$H_0 : X(t) = N(t)$, normal noise alone is present,

against the simple alternative hypothesis

$H_1 : X(t) = S(t) + N(t)$, a signal $S(t)$ of prescribed shape is present, by choosing a subset R_1 of the space Ω of possible realizations of the time series which will be the *rejection region* for H_0 ; that is, one says signal is present if $\{X(t), t \in T\}$ belongs to R_1 and one says that noise alone is present if $\{X(t), t \in T\}$ does not belong to R_1 .

While there exist a number of criteria for optimally choosing the rejection region R_1 , it turns out that the optimum rejection region R_1 may in all cases be expressed as the set of observations $\{X(t), t \in T\}$ for which the probability density of the signal plus noise process with respect to the noise process

$$p(X(t), t \in T) = \exp[(X, S)_K - \tfrac{1}{2}(S, S)_K] \tag{4.1}$$

is above a certain threshold value Λ_0 .

The threshold level Λ_0 depends on the criterion employed. In the so-called Bayes case,

$$\Lambda_0 = (1 - p_s)L_{\mathrm{I}}/p_s L_{\mathrm{II}} \tag{4.2}$$

where L_{I} is the cost of a false alarm (of saying that signal is present when in fact noise alone is present), L_{II} is the cost of a detection failure (of saying that noise alone is present when in fact signal is present), and p_s is the prior probability that signal is present. In the so-called Neyman-Pearson case, one chooses a desired level α for the false alarm probability

$$P[\{X(t), t \in T\} \in R_1 \mid H_0] = P[p(X(t), t \in T) \geqq \Lambda_0 \mid H_0] = \alpha. \tag{4.3}$$

It should be noted that in the Bayes case the critical region R_1 has the property that it minimizes the expected cost of an incorrect decision

$$\begin{aligned}(1 - p_s)L_{\mathrm{I}}P[\{X(t), t \in T\} \in R_1 \mid H_0] \\ + p_s L_{\mathrm{II}}\{1 - P[\{X(t), t \in T\} \in R_1 \mid H_1]\}\end{aligned} \tag{4.4}$$

while in the Neyman-Pearson case the critical region R_1 has the property that it minimizes the probability

$$1 - P[\{X(t), t \in T\} \in R_1 \mid H_1] \tag{4.5}$$

that one will fail to detect a signal when it is present, subject to the restriction that the false alarm probability is equal to α.

The probabilities in (4.3) and (4.5) are easily evaluated if one uses the fact that, since the right hand side of (4.1) is a monotone increasing function of $(X, S)_K$, the critical region for testing H_0 against H_1 can be expressed as the set of observations $\{X(t), t \in T\}$ for which $(X, S)_K$ is above a certain threhsold Λ_1. Since $(X, S)_K$ is a linear functional, it follows that it is normally distributed. It may be shown that

$$\begin{aligned} &\text{under } H_0 : \mathrm{E}[(X, S)_K] = 0, \qquad \mathrm{Var}\,[(X, S)_K] = (S, S)_K ; \\ &\text{under } H_1 : \mathrm{E}[(X, S)_K] = (S, S)_K, \qquad \mathrm{Var}\,[(X, S)_K] = (S, S)_K . \end{aligned} \tag{4.6}$$

Since the critical region may be expressed in terms of $(X, S)_K$, we call $(X, S)_K$ an *optimum detector*. The detector $(X, S)_K$, which in the case that T is finite is

$$(X, S)_K = \sum_{t \in T} X(t) \sum_{u \in T} K^{tu} S(u), \tag{4.7}$$

is said to be a *correlation detector* or a *matched filter* since it is obtained by 'correlating' or 'matching' the specified signal shape $S(t)$ with the observed time series $X(t)$.

A convenient measure of how far it is possible to discriminate between two simple hypotheses H_1 and H_0 is provided by the *divergence*, a quantity originating from the concepts of information theory (see [6]) and defined by

$$J(H_0, H_1) = \mathrm{E}_{H_1}[\log p(X(t), t \in T)] - \mathrm{E}_{H_0}[\log p(X(t), t \in T] \tag{4.8}$$

where

$$p(X(t), t \in T) = \frac{dP_{H_1}}{dP_{H_0}}$$

is the probability density of the observations under H_1 with respect to the probability measure corresponding to H_0 (the subscript H_j on an expectation operator indicates that the expectation is taken with respect to the probability corresponding to H_j). For $p(X(t), t \in T)$ given by (4.1)

$$J(H_0, H_1) = \mathrm{E}_{H_1}[(X, S)_K] - \mathrm{E}_{H_0}[(X, S)_K] = (S, S)_K . \tag{4.9}$$

One can regard $(S, S)_K$ as a measure of the *signal to noise ratio*. One thus sees that the larger the signal to noise ratio, the better is one able to discriminate signal plus noise from noise alone.

Case II. Testing for a regression signal. Given an observed time series $\{X(t),\ t \in T\}$, one desires to test the simple hypothesis

$$H_0 : X(t) = N(t), \quad \text{noise alone is present,}$$

against the composite alternative hypothesis

$$H_1 : X(t) = \theta g(t) + N(t) \quad \text{for some } \theta, \quad \text{a signal is present.}$$

It is assumed that g belongs to the reproducing kernel Hilbert space $H(K)$ corresponding to the corvariance kernel K of the noise.

According to the likelihood ratio principle, the rejection region for testing H_0 against H_1 is given by the set of values $\{X(t),\ t \in T\}$ for which the supremum

$$\sup_{-\infty<\theta<\infty} p(X(t),\ t \in T \mid \theta) \tag{4.10}$$

is above a suitable threshold value. If θ^{**} is the maximum likelihood estimate of θ, then the supremum is equal to

$$\exp\left\{\theta^{**}V - \frac{1}{2}\theta^{**2}G\right\} = \exp\left\{\frac{1}{2}\frac{V^2}{G}\right\} = \exp\left\{\frac{1}{2}\theta^{**}I^{**}\theta^{**}\right\}. \tag{4.11}$$

In the case of a multi-parameter model of the form of (3.25), the supremum in (4.10) is equal to

$$\exp\{\tfrac{1}{2}\ V^{\mathrm{tr}}G^{-1}V\} = \exp\{\tfrac{1}{2}\ \theta^{**\mathrm{tr}}I^{**}\theta^{**}\}. \tag{4.12}$$

The probability in (4.3) is easily evaluated using the fact that

$$\text{under } H_0\text{, the quadratic form } V^{\mathrm{tr}}G^{-1}V \text{ is } \chi^2 \text{ distributed with } k \text{ degrees of freedom.} \tag{4.13}$$

It should be noted that in Case I the optimum detector $(X, S)_K$ was a linear function of the observations. In Case II, the optimum detector $V^{\mathrm{tr}}G^{-1}V$ is a quadratic function of the observations $\{X(t),\ t \in T\}$. Further one can write $V^{\mathrm{tr}}G^{-1}V$ as a *generalized correlation detector*

$$V^{\mathrm{tr}}G^{-1}V = (X,\ g^{\mathrm{tr}}G^{-1}V)_K = (X,\ g^{\mathrm{tr}}\theta^{**})_K \tag{4.14}$$

where $g^{\mathrm{tr}} = (g_1, \cdots, g_k)$ is a function of t. Now the maximum likelihood estimate of the value $S(t)$ at t of the signal

$$S(\cdot) = \textstyle\sum_{j=1}^{k}\theta_j\, g_j(\cdot) = g^{\mathrm{tr}}\theta \tag{4.15}$$

is given by the value at t of the function

$$S^{**}(\cdot) = g^{\mathrm{tr}}\theta^{**}. \tag{4.16}$$

Consequently,

$$V^{\mathrm{tr}}G^{-1}V = (X,\ S^{**})_K\,. \tag{4.17}$$

In words, the optimum detector in Case II is of the same form as in Case I except that instead of correlating the prescribed signal shape $S(\cdot)$ with the observed time series $X(\cdot)$ one correlates an estimated signal shape.

Case III. Testing for a stochastic signal. One way in which the problem of detecting a stochastic signal arises is when one considers a regression signal with random regression coefficients. More precisely, one desires to test the simple hypothesis

$$H_0 : X(\cdot) = N(\cdot), \quad \text{noise alone is present,}$$

against the simple alternative

$$H_1 : X(\cdot) = \theta^{\mathrm{tr}} g(\cdot) + N(\cdot), \quad \text{where } \theta \text{ is multivariate normal with mean } \bar{\theta} \text{ and covariance } \bar{I}^{-1}.$$

The rejection region for testing H_0 against H_1 is given by the set of values $\{X(t),\ t \in T\}$ for which the probability density function of signal plus noise with respect to noise

$$p(X(t), t \in T) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(X(t), t \in T \mid \theta) p(\theta)\, d\theta_1 \cdots d\theta_k \tag{4.18}$$

is above a certain threshold value. Define (using the notation defined in (3.27) and (3.33))

$$\alpha = V + \bar{I}\bar{\theta}, \quad I^* = G + \bar{I} \tag{4.19}$$

Then

$$\begin{aligned} p(X(t), t \in T) &= (2\pi)^{-k} |\, \bar{I}\, |^{\frac{1}{2}} \exp\left[-\tfrac{1}{2}\bar{\theta}\ \mathrm{tr}\ \bar{I}\bar{\theta}\right] \\ &\quad \times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left\{\theta^{\mathrm{tr}}\alpha - \frac{1}{2}\theta^{\mathrm{tr}} I^* \theta\right\} d\theta_1 \cdots d\theta_k \\ &= |\, \bar{I}^* \bar{I}^{-1}\, |^{-\frac{1}{2}} \exp\left[\tfrac{1}{2}\{\alpha^{\mathrm{tr}} I^{*-1}\alpha - \bar{\theta}^{\mathrm{tr}}\bar{I}\bar{\theta}\}\right]. \end{aligned} \tag{4.20}$$

Now

$$\begin{aligned} \alpha^{\mathrm{tr}} I^{*-1}\alpha - \bar{\theta}^{\mathrm{tr}}\bar{I}\bar{\theta} &= \theta^{*\mathrm{tr}} I^* \theta^* - \bar{\theta}^{\mathrm{tr}}\bar{I}\bar{\theta} \\ &= (X, S^*) - \bar{\theta}^{\mathrm{tr}}\bar{I}(\theta^* - \bar{\theta}), \end{aligned} \tag{4.21}$$

where $\theta^* = I^{*-1}\alpha$ is the Bayes estimate of θ, and $S^* = \theta^{*\mathrm{tr}} g$ is the Bayes estimate of $S(\cdot) = \theta^{\mathrm{tr}} g(\cdot)$. An optimum detector is given by the last term of (4.21); if $\bar{\theta} = 0$, then it is a generalized correlation detector.

Case IV. Classifying or decoding signals. The output of a certain channel is assumed to be a time series of the form

$$X(t) = S_j(t) + N(t), \quad \text{for some } j = 1, 2, \cdots, k,$$

where the parameter j characterizes the waveform that was the input of the channel. On the basis of the observed time series $\{X(t),\ t \in T\}$, one desires to decide which signal was transmitted, that is, one desires to divide the space Ω of possible realizations of the time series into k regions

$$R_1, R_2, \cdots, R_k$$

such that if $\{X(t), t \in T\}$ belongs to R_j, one decides that the j-th waveform was transmitted (was the channel input). Thus the problem is one of *decoding* a received message in such a way as to maximize the probability of a correct classification (the *encoding* problem is concerned with choosing the transmitted waveforms so as to maximize the probability of correct decoding using the optimum decoder).

For $j = 1, 2, \cdots, k$, let

$$p_j(X(t), t \in T)$$

denote the probability density of the time series $\{S_j(t) + N(t), t \in T\}$ with respect to the noise process $\{N(t), t \in T\}$, and let π_j be the prior probability that $S_j(\cdot)$ was received. The probability of a correct decision (using classification regions $R_1, \cdots, R_k$) is given by

$$\sum_{j=1}^{k} \pi_j \int_{R_j} p_j(X(t), t \in T)\ dP_N .$$

It may be shown (see [13], p. 308) that in order to maximize the probability of a correct decision one should adopt the classification regions $R_1, \cdots, R_k$ defined by

$$R_j = \{(X(t), t \in T) : \pi_j p_j(X(t), t \in T) = \max_{i=1,\cdots,k} \pi_i p_i(X(t), t \in T)\}.$$

In other words, the classification regions which maximize the probability of a correct classification coincide with the classification regions one would obtain by maximizing the posterior probability

$$P[\text{ -th waveform transmitted} \mid \{X(t), t \in T\}] = \frac{\pi_j\, p_j(X(t), t \in T)}{\sum_{i=1}^{k} \pi_i\, p_i(X(t), t \in K)} .$$

In any event, the first step in determining the classification regions is to compute the probability density functionals $p_j(X(t), t \in T)$ which, in the case of normal time series, may be shown to depend on the computation of various inner products in a suitable reproducing kernel Hilbert space.

5. Evaluation of reproducing kernel inner products, orthonormal expansions, and eigenfunction expansions. The developments of sections 2, 3, and 4 show that to determine optimum extractors and detectors for

signals in normal noise, as well as to evaluate the probability density function of signal plus (normal) noise with respect to (normal) noise, it suffices to evaluate the reproducing kernel inner product

$$(h, h)_K \tag{5.1}$$

of various functions h, and the corresponding random variables

$$(X, h)_K \tag{5.2}$$

In this section we discuss various methods of evaluating expressions of the form of (5.1) and (5.2).

Explicit formulae for expressions of the form of (5.1) and (5.2) can be given if the noise process $\{N(t), t \in T\}$ is one of the following types (stated for the continuous parameter univariate case; similar results hold in the discrete parameter and multivariate cases):

(1) T is a finite interval, and the noise process has independent (or orthogonal) increments so that its covariance kernel is of the form

$$K(s, t) = G(\min\{s, t\})$$

for some continuous non-decreasing function $G(u)$.

(2) T is a finite interval, and the noise process is Markov.

(3) T is a finite interval, and the noise process is an autoregressive scheme in the sense that it satisfies the stochastic differential equation

$$\sum_{k=0}^{m} a_k(t)\, N^{(m-k)}(t) = \eta'(t)$$

where $\eta'(t)$ is a white noise (the symbolic derivative of a process $\eta(t)$ with stationary and independent increments), m is a constant (called the order of the scheme), and $a_k(t)$ are non-random functions of time.

(4) T is a semi-infinite interval, $-\infty < t \leqq t_0$, and the noise process can be represented as the response to a white noise input of a filter described by a time-varying impulse response function $W(t, s)$:

$$N(t) = \int_{-\infty}^{t} W(t, s)\, d\eta(s).$$

(5) T is a semi-infinite interval, $-\infty < t \leqq t_0$, and the noise process is stationary (more precisely, covariance stationary) and possesses a rational spectral density function or, more generally, is purely non-deterministic.

In many applications one will not feel justified in assuming a specific model of one of the above type. Nevertheless explicit formulae for expressions of the form (5.1) and (5.2) can be constructed. Indeed they can be constructed in a multitude of ways, as we now show. In the course of our discussion we will also clarify the connection between

(i) reproducing kernel Hilbert spaces,

(ii) orthonormal expansions for time series, and
(iii) expansions of time series in terms of eigenfunctions (often called Karhunen-Loève expansions).

Orthonormal expansions of time series. Using the Gram-Schmidt orthogonalization procedure one may show that every time series $\{X(t), t \in T\}$, for which the Hilbert space $L_2(X(t), t \in T)$ spanned by it is separable, may be written as an infinite series

$$X(t) = \sum_{\nu=1}^{\infty} \eta_\nu \psi_\nu(t), \qquad t \in T, \tag{5.3}$$

where (i) $\{\eta_\nu\}$ is an orthonormal sequence of random variables,

$$\mathrm{E}[\eta_\alpha \eta_\beta] = \delta(\alpha, \beta) = \begin{cases} 1 \text{ if } \alpha = \beta \\ 0 \text{ if } \alpha \neq \beta, \end{cases} \tag{5.4}$$

(ii) the function $\psi_\nu(t)$ is given by (for t in T and $\nu = 1, 2, \cdots$)

$$\psi_\nu(t) = \mathrm{E}[X(t)\eta_\nu], \tag{5.5}$$

(iii) for each t in T

$$\sum_{\nu=1}^{\infty} \psi_\nu^2(t) < \infty,$$

and (iv) for all s and t in T

$$\mathrm{E}[X(s)X(t)] = \sum_{\nu=1}^{\infty} \psi_\nu(s)\psi_\nu(t). \tag{5.7}$$

A representation of the form of (5.3) is called an *orthogonal decomposition* of the time series. It should be noted that the orthogonal decomposition is not unique.

The importance of the representation (5.3) becomes clear if one bears in mind that the time series $X(t)$ is actually a function of two variables $X(t, \omega)$ where t varies in T and ω varies in the space Ω on which $X(t)$ is defined as a random variable. By writing (5.3) in the form

$$X(t, \omega) = \sum_{\nu=1}^{\infty} \psi_\nu(t)\eta_\nu(\omega) \tag{5.3'}$$

one sees that we have succeeded in decomposing the function of two variables $X(t, \omega)$ into a sum of products of functions $\psi_\nu(t)$ and $\eta_\nu(\omega)$ of one variable. In a sense, we have succeeded in isolating the manner in which $X(t, \omega)$ depends on t from the manner in which it depends on ω.

It may be shown that, in terms of an orthonormal decomposition, the reproducing kernel Hilbert space $H(K)$ may be expressed as follows: $H(K)$ consists of all functions $h(t)$ on T which may be represented in the

form

$$h(t) = \sum_{\nu=1}^{\infty} h_\nu \psi_\nu(t), \qquad t \in T, \tag{5.8}$$

for some (necessarily unique) square summable sequence $\{h_\nu\}$. The expressions in (5.1) and (5.2) are then given by

$$(h, h)_{K,T} = \sum_{\nu=1}^{\infty} h_\nu^2, \tag{5.9}$$

$$(h, X)_{K,T} = \sum_{\nu=1}^{\infty} h_\nu \eta_\nu . \tag{5.10}$$

In order for (5.9) and (5.10) to be useful in practice one needs algorithms for determining the sequences $\{h_\nu\}$ and $\{\eta_\nu\}$.

Let G be a Hilbert space whose members are functions on T. Suppose that $\{X(t), t \in T\}$ possesses an orthogonal decomposition which in addition to the properties (5.4)–(5.7) possesses the property that $\{\psi_\nu(\cdot), \nu = 1, 2, \cdots\}$ is an orthogonal sequence of functions in G:

$$(\psi_\alpha , \psi_\beta)_G = 0 \quad \text{if} \quad \alpha \neq \beta. \tag{5.11}$$

If (5.11) holds, it seems plausible that an explicit formula for η_α is given by

$$\eta_\alpha = \frac{(X, \psi_\alpha)_G}{(\psi_\alpha , \psi_\alpha)_G} \tag{5.12}$$

since if one takes the inner product of both sides of (5.9) with ψ_α one obtains

$$(X, \psi_\alpha)_G = \sum_\nu \eta_\nu (\psi_\nu , \psi_\alpha)_G = \eta_\alpha (\psi_\alpha , \psi_\alpha)_G . \tag{5.13}$$

We discuss below the meaning of such expressions as $(X, \psi_\alpha)_G$. Similarly, one may show that

$$h_\alpha = \frac{(h, \psi_\alpha)_G}{(\psi_\alpha , \psi_\alpha)_G} . \tag{5.14}$$

We thus arrive at the following conclusion: *an explicit expression for the reproducing kernel inner products* (5.1) *and* (5.2) *can be obtained in terms of inner products in any Hilbert space G in which there exists an orthogonal sequence* $\{\psi_\nu , \nu = 1, 2, \cdots\}$ *playing the role of coefficients in an orthogonal representation of* $\{X(t), t \in T\}$.

In order to make precise the foregoing discussion we have

(i) to determine which Hilbert spaces G have the property that they contain an orthogonal sequence $\{\psi_\nu , \nu = 1, 2, \cdots\}$ playing the role of coefficients in an orthogonal decomposition of $\{X(t), t \in T\}$,

(ii) to provide an algorithm for computing the orthogonal functions $\{\psi_\nu, \nu = 1, 2, \cdots\}$.

As possible Hilbert spaces G we consider either L_2 spaces $L_2(T, \bar{B}, \mu)$ or reproducing kernel Hilbert spaces whose members are functions on T. The space $L_2(T, \bar{B}, \mu)$ consists of all $\bar{B}$-measurable functions g on T such that

$$\| g \|_\mu^2 = \int_T | g(t) |^2 \mu(dt) < \infty \tag{5.15}$$

where $\bar{B}$ is a σ-field of subsets of T and μ is a measure defined on $\bar{B}$.

The notion of a direct product space plays an important part in our considerations. Given two function spaces G_1 and G_2 consisting of functions defined on T_1 and T_2 respectively, their direct product space, denoted $G_1 \otimes G_2$, is the Hilbert space completion of the set of functions g on $T_1 \otimes T_2$ of the form

$$g(t_1, t_2) = g_1(t_1)g_2(t_2), \tag{5.16}$$

where $g_1 \in G_1$ and $g_2 \in G_2$. The norm of a function in $G_1 \otimes G_2$ of the form of (5.16) is defined by

$$\| g \|_{G_1 \otimes G_2}^2 = \| g_1 \|_{G_1}^2 \| g_2 \|_{G_2}^2 . \tag{5.17}$$

The function g defined by (5.16) is on occasion denoted by $g_1 \otimes g_2$.

It should be noted that if G_1 and G_2 are reproducing kernel Hilbert spaces, with respective reproducing kernels K_1 and K_2 defined on $T \otimes T$, then $G_1 \otimes G_2$ is a reproducing kernel Hilbert space with kernel $K_1 \otimes K_2$, where $K_1 \otimes K_2$ is a function of four real variables defined by

$$K_1 \otimes K_2(s_1, s_2, t_1, t_2) = K_1(s_1, t_1)K_2(s_2, t_2) \tag{5.18}$$

and

$$(g, K_1 \otimes K_2(\cdot, \cdot, t_1, t_2))_{G_1 \otimes G_2} = g(t_1, t_2). \tag{5.19}$$

In the case that $G_1 = G_2 = L_2(T, \bar{B}, \mu)$, $G_1 \otimes G_2$ consists of all ($\bar{B} \otimes \bar{B}$-measurable) functions g on $T \otimes T$ such that

$$\| g \|_{G_1 \otimes G_2}^2 = \int_T \int_T g^2(s, t) \, \mu(ds) \, \mu(dt) < \infty. \tag{5.20}$$

If G_1 and G_2 are equal to the reproducing kernel Hilbert space consisting of all L_2-differentiable functions on the interval $\{t: a \leqq t \leqq b\}$ with norm squared

$$\| g \|_{G_1}^2 = \frac{1}{a} g_1^2(a) + \int_a^b | g_1'(t) |^2 \, dt, \tag{5.21}$$

then $G_1 \otimes G_2$ is a reproducing kernel Hilbert space with norm squared

$$\begin{aligned} \| g \|^2_{G_1 \otimes G_2} &= \frac{1}{a^2} g^2(a, a) + \frac{1}{a} \int_a^b \left| \frac{\partial}{\partial s} g(s, a) \right|^2 ds \\ &+ \frac{1}{a} \int_a^b \left| \frac{\partial}{\partial t} g(a, t) \right|^2 dt \\ &+ \int_a^b \int_a^b \left| \frac{\partial}{\partial s} \frac{\partial}{\partial t} g(s, t) \right|^2 ds\, dt. \end{aligned} \tag{5.22}$$

Some indication of the method to be employed in solving the problem posed after (5.14) can be obtained by considering the properties which the sequence $\{\psi_\nu\}$ must have in order that the random variables $\{\xi_\alpha\}$, defined by

$$\xi_\alpha = (X, \psi_\alpha)_\mu , \tag{5.23}$$

be orthogonal. One may verify that

$$\begin{aligned} \mathrm{E}[\xi_\alpha \xi_\beta] &= \mathrm{E}\left[\int_T X(s)\psi_\alpha(s)\, \mu(ds) \int_T X(t)\psi_\beta(t)\, \mu(dt) \right] \\ &= \int_T \int_T \psi_\alpha(s)\psi_\beta(t)\, K(s, t)\, \mu(ds)\, \mu(dt). \end{aligned} \tag{5.24}$$

Now suppose the functions $\{\psi_\alpha\}$ not only have the property of being orthogonal as members of G, but also have the property that there exist non-zero constants λ_α such that, for $\alpha = 1, 2, \cdots$ and t in T,

$$\int_T \psi_\alpha(s) K(s, t)\, \mu(ds) = \lambda_\alpha \psi_\alpha(t). \tag{5.25}$$

Then

$$\mathrm{E}[\xi_\alpha \xi_\beta] = \lambda_\alpha \int_T \psi_\alpha(t)\psi_\beta(t)\, \mu(dt) = 0 \quad \text{if} \quad \alpha \neq \beta. \tag{5.26}$$

We now state conditions under which one can find functions $\{\psi_\alpha\}$ satisfying (5.25) as the eigenfunctions of a certain transformation.

A function belonging to the direct product Hilbert space $G \otimes G$ is said to have finite double-G norm. One can prove the following theorem.

THEOREM. *Let G be a Hilbert space of functions on T which is either an L_2-space or a reproducing kernel Hilbert space. Let $\{X(t), t \in T\}$ be a time series with covariance kernel K. Assume that K has finite double G-norm. Let* $\mathbf{K}$ *denote the transformation on G defined as follows: for g in G,* $\mathbf{K}g$ *is a function on T with value at t given by*

$$\mathbf{K}g(t) = (g, K(\cdot, t))_G . \tag{5.27}$$

Then (i) $\mathbf{K}$ *is well defined and, for all g in G,* $\mathbf{K}g \in G$; (ii) $\mathbf{K}$ *is a linear,*

self-adjoint, non-negative definite, and completely continuous transformation of G into itself.

A non-zero number λ is called an eigenvalue of $\mathbf{K}$ if there exists a non-zero function g in G such that

$$\mathbf{K}g = \lambda g; \tag{5.28}$$

the function g is called an eigenfunction corresponding to λ.

Since $\mathbf{K}$ is a self-adjoint, non-negative definite, completely continuous operator, it follows (see [14], p. 233) that the set $\{\lambda_\nu\}$ of non-zero eigenvalues of $\mathbf{K}$ is an infinite sequence of positive numbers converging to zero; arranging these eigenvalues in non-increasing order we write

$$\lambda_1 \geqq \lambda_2 \geqq \cdots \lambda_n \cdots \longrightarrow 0, \tag{5.29}$$

making the convention that each eigenvalue is written as many times as its multiplicity.

We let

$$\varphi_1 , \varphi_2 \cdots \tag{5.30}$$

denote the sequence of corresponding normalized eigenfunctions; that is,

$$\mathbf{K}\varphi_\nu = \lambda_\nu \varphi_\nu , \tag{5.31}$$

$$(\varphi_\alpha , \varphi_\beta)_G = \delta(\alpha, \beta). \tag{5.32}$$

It follows [13, p. 245] that the kernel K may be represented by the absolutely convergent series (for all s and t in T)

$$K(s, t) = \sum_{\nu=1}^{\infty} \lambda_\nu \varphi_\nu(s)\varphi_\nu(t). \tag{5.33}$$

Further,

$$\| K \|_{G\otimes G} = \sum_{\nu=1}^{\infty} \lambda_\nu^2. \tag{5.34}$$

The reproducing kernel Hilbert space $H(K)$ can be shown to consist of all functions h which belong to the Hilbert subspace $\mathbf{L}(\varphi_\nu ; \nu = 1, 2, \cdots)$ of G spanned by the eigenfunctions $\{\varphi_\nu\}$ for which

$$\sum_{\nu=1}^{\infty} \frac{1}{\lambda_\nu} | (h, \varphi_\nu)_G |^2 < \infty. \tag{5.35}$$

The reproducing kernel inner product of two functions h_1 and h_2 in $H(K)$ can be represented in terms of their inner products in G with the eigenfunctions $\{\varphi_\nu\}$:

$$(h_1 , h_2)_K = \sum_{\nu=1}^{\infty} \frac{1}{\lambda_\nu} (h_1 , \varphi_\nu)_G (h_2 , \varphi_\nu)_G . \tag{5.36}$$

The random variable $(X, h)_K$ is given by

$$(X, h)_K = \sum_{\nu=1}^{\infty} \frac{1}{\lambda_\nu} (X, \varphi_\nu)_G (h, \varphi_\nu)_G . \tag{5.37}$$

The expression $(X, \varphi_\nu)_G$ is well defined and satisfies

$$\mathrm{E}[|(X, \varphi_\nu)_G|^2] \leqq \| \varphi_\nu \|_G^2 \| K \|_{G\otimes G} = \| K \|_{G\otimes G} . \tag{5.38}$$

One can justify (5.38) in a number of ways.

Consider first the case that G is an L_2-space. Then $(X, \varphi_\nu)_G$ is a stochastic integral,

$$(X, \varphi_\nu)_G = \int_T X(t)\varphi_\nu(t)\, \mu(dt),$$

and

$$\begin{aligned} \mathrm{E}[|(X, \varphi_\nu)_G|^2] &= \int_T \int_T K(s, t)\varphi_\nu(s)\varphi_\nu(t)\, \mu(ds)\, \mu(dt) \\ &\leqq \left\{ \int_T \int_T \varphi_\nu^2(s)\varphi_\nu^2(t)\, \mu(ds)\mu(dt) \int_T \int_T K^2(s, t)\, \mu(ds)\, \mu(dt) \right\}^{\frac{1}{2}} \\ &= \| \varphi_\nu \|_G^2 \| K \|_{G\otimes G} . \end{aligned}$$

In the case that G is a reproducing kernel Hilbert space, we define $(X, \varphi_\nu)_G$ as follows. From (5.33) it follows that the time series $\{X(t), t \in T\}$ has the representation

$$X(t) = \sum_\nu \xi_\nu \varphi_\nu(t) \tag{5.39}$$

where $\{\xi_\nu\}$ is an uncorrelated sequence of random variables with mean squares

$$\mathrm{E}[\xi_\nu^2] = \lambda_\nu . \tag{5.40}$$

If we define

$$(X, \varphi_\nu)_G = \xi_\nu , \tag{5.41}$$

then the symbolism $(X, \varphi_\nu)_G$ has all the formal properties we desire it to have. For example, to prove that

$$\mathrm{E}[|(X, \varphi_\nu)_G|^2] = \lambda_\nu , \tag{5.42}$$

which follows by definition from (5.40) and (5.41), we write

$$\begin{aligned} \mathrm{E}[|(X, \varphi_\nu)_G|^2] &= \mathrm{E}[(X \otimes X, \varphi_\nu \otimes \varphi_\nu)_{G\otimes G} \\ &= (K, \varphi_\nu \otimes \varphi_\nu)_{G\otimes G} \\ &= ((K(\cdot, t), \varphi_\nu)_G , \varphi_\nu(t))_G = \lambda_\nu . \end{aligned}$$

In practice one will be able to evaluate $(X, \varphi_\nu)_G$ as a stochastic integral. Thus if $T = \{t: a \leqq t \leqq b\}$ and G is the reproducing kernel Hilbert space of L_2 differentiable functions with norm satisfying (5.21), then

$$(X, \varphi_\nu)_G = \frac{1}{a} X(a)\varphi_\nu(a) + \int_a^b X'(t)\varphi_\nu'(t)\, dt.$$

This expression is well defined as a stochastic integral if the covariance kernel $K(s, t) = \mathrm{E}[X(s)X(t)]$ is assumed to have continuous second derivatives.

6. Iterative evaluation of reproducing kernel inner products. The formulae (5.36) and (5.37) are not computationally convenient since they involve the calculation of eigenvalues and eigenfunctions. It is possible to give iterative methods (of steepest descent type) for evaluating the reproducing kernel inner product $(h, h)_K$ and the corresponding random variable $(X, h)_K$.

Let $\{X(t), t \in T\}$ be a time series with covariance kernel K and corresponding reproducing kernel Hilbert space $H(K)$. Let G be a Hilbert space such that K has finite double-G norm

$$N = \| K \|_{G\otimes G}. \tag{6.1}$$

Given a function h in $H(K)$ one may generate (in a multitude of ways) sequences $\{H_n\}$ of functions in G such that

$$\lim_{n\to\infty} \mathrm{E}[|(X, h)_K - (X, H_n)_G|^2] = 0, \tag{6.2}$$

$$(h, h)_K = \lim_{n\to\infty} (K, H_n \otimes H_n)_{G\otimes G} = \lim_{n\to\infty} (\mathbf{K}H_n, H_n)_G. \tag{6.3}$$

Using the method of proof in [8], one may prove that a sequence $\{H_n\}$ satisfying (6.2) and (6.3) is given by

$$H_{n+1} = H_n - \alpha_n(\mathbf{K}H_n - h), \qquad n = 1, 2, \cdots \tag{6.4}$$

where H_1 is chosen arbitrarily in G, and $\{\alpha_n\}$ is a sequence satisfying

$$0 < \alpha_n \leqq 2/N. \tag{6.5}$$

There exist a number of schemes for choosing the sequence $\{\alpha_n\}$ so as to achieve fastest convergence; these will be discussed in a book in preparation [12].

7. Linear and non-linear prediction. Reproducing kernel Hilbert spaces also provide a formal solution to the problems of minimum mean square error linear and non-linear prediction.

The Hilbert space spanned by a time series $\{X(t), t \in T\}$, denoted $\mathbf{L}(X(t), t \in T)$, is defined as the smallest Hilbert space which contains all

random variables U of the form

$$U = \sum_{j=1}^{n} c_j X(t_j)$$

for some points $t_1, \cdots, t_n$ in T and some real numbers $c_1, \cdots, c_n$. Intuitively, $\mathbf{L}(X(t), t \in T)$ is the space of all random variables which are linear functions, or functionals, of the time series $\{X(t), t \in T\}$.

The space of all *non-linear* functionals in a time series $\{X(t), t \in T\}$, denoted $\mathbf{N}(X(t), t \in T)$, is defined as the smallest Hilbert space containing all random variables U which have finite second moment and are of the form

$$U = g(X(t_1), \cdots, X(t_n))$$

for some $t_1, \cdots, t_n$ in T and some Borel function $g(x_1, \cdots, x_n)$.

Let U be a random variable (with finite second moment) whose value it is desired to predict, using the observations $\{X(t), t \in T\}$. The minimum mean square error non-linear prediction of U, given $\{X(t), t \in T\}$, is defined as that random variable U^* in $\mathbf{N}(X(t), t \in T)$ satisfying

$$\text{(7.1)} \qquad \mathrm{E}[|\, U^* - U \,|^2] = \min_{V \in N(X(t), t \in T)} \mathrm{E}[|\, V - U \,|^2].$$

The minimum mean square error (M.M.S.E.) linear predictor of U, given $\{X(t), t \in T\}$, is defined as that random variable U^* in $\mathbf{L}(X(t), t \in T)$ satisfying

$$\text{(7.2)} \qquad \mathrm{E}[|\, U^* - U \,|^2] = \min_{V \in \mathbf{L}(X(t), t \in T)} \mathrm{E}[|\, V - U \,|^2]$$

It is easily shown that the M.M.S.E. non-linear predictor, of U, given $\{X(t), t \in T\}$, is the *conditional mean* of U given $\{X(t), t \in T\}$, denoted $\mathrm{E}[U \mid X(t), t \in T]$, and defined as the unique random variable in $\mathbf{N}(X(t), t \in T)$ with the property that

$$\text{(7.3)} \qquad \mathrm{E}[VU] = \mathrm{E}[V\mathrm{E}[U \mid X(t), t \in T]]$$

for every V in $\mathbf{N}(X(t), t \in T)$. Similarly the M.M.S.E. linear predictor of U, given $\{X(t), t \in T\}$, is the *projection* of V onto $\{X(t), t \in T\}$, denoted $\mathrm{E}'[U \mid X(t), t \in T]$ and defined as the unique random variable in $\mathbf{L}(X(t), t \in T)$ with the property that

$$\text{(7.4)} \qquad \mathrm{E}[VU] = \mathrm{E}[V\mathrm{E}'[U \mid X(t), t \in T]]$$

for every V in $\mathbf{L}(X(t), t \in T)$.

Define

$$\text{(7.5)} \qquad \rho_U(t) = \mathrm{E}[UX(t)],$$

$$\text{(7.6)} \qquad \rho_U(t, v) = \mathrm{E}[Ue^{ivX(t)}].$$

It may be shown that $E[U \mid X(t), t \in T]$ is the random variable U^* which has minimum mean square $E[|\ U^*\ |^2]$ among all random variables satisfying the restraint, for all t in T and $-\infty < v < \infty$,

$$E[U^* e^{ivX(t)}] = \rho_U(t, v). \tag{7.7}$$

Similarly $E'[U \mid X(t), t \in T]$ is the random variable U^* which has minimum mean square among all random variables satisfying the restraint, for all t in T,

$$E[U^* X(t)] = \rho_U(t). \tag{7.8}$$

To find the solutions U^* of (7.7) and (7.8) one may proceed as follows. For s and t in T, and $-\infty < u, v < \infty$, define

$$K(s, t) = E[X(s)X(t)] \tag{7.9}$$

$$K(s, u; t, v) = E[e^{iuX(s)} e^{-ivX(t)}]. \tag{7.10}$$

Notice that $K(s, u; t, v)$ is the two-dimensional characteristic function of the time series (see [11] for a discussion of this terminology).

These functions are reproducing kernels. Further

$$\rho_U(t) \in H(K(s, t)), \qquad E'[U \mid X(t), t \in T] = (X, \rho_U)_K \tag{7.11}$$

$$\rho_U(t, v) \in H(K(s, u; t, v)), \qquad E[U \mid X(t), t \in T] = (e^{ivX(t)}, \rho_U)_K\ . \tag{7.12}$$

In words, if one can find a representation for $\rho_U(\cdot)$ in terms of linear operations on the functions $\{K(\cdot, t), t \in T\}$, then $E'[U \mid X(t), t \in T]$ can be written in terms of the corresponding linear operations on the time series $\{X(t), t \in T\}$. Similarly, if one can find a representation for $\rho_U(\cdot\,; \cdot)$ in terms of linear operations on the functions $\{K(\cdot, \cdot, t, v), t \in T, -\infty < v < \infty\}$, then $E[U \mid X(t), t \in T]$ can be written in terms of the corresponding linear operations on the family of random variables $\{e^{ivX(t)}, t \in T, -\infty < v < \infty\}$. *The expressions in* (7.11) *and* (7.12) *may be numerically evaluated using the iterative methods discussed in section* 6.

The solution to the non-linear prediction problem can be expressed in terms of polynomials in the observed random variables $\{X(t), t \in T\}$ if the following conditions are satisfied:

(i) for every t in T and integer k,

$$E[|\ X(t)|^k] < \infty,$$

(ii) for every finite subset $\{t_1, \cdots, t_n\}$ of T, and constants $c_1, \cdots, c_n$, the probability distribution of $\sum_{i=1}^{n} c_i X(t_i)$ is determined by its moments.

Assuming that these conditions are satisfied, define, for s, t in T and $j, k = 1, 2, \cdots$,

$$\rho_U(t, k) = E[UX^k(t)], \tag{7.13}$$

$$K(s, j; t, k) = \mathrm{E}[X^j(s)X^k(t)]. \tag{7.14}$$

Then

$$\mathrm{E}[U \mid X(t), t \in T] = (X^k(t), \rho_U)_K . \tag{7.15}$$

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] J. Hajek, *A property of J-divergences of marginal probability distributions*, Czechoslovak Math. J., 8 (83) (1958), pp. 460–463.

[3] J. Hajek, *On a property of normal distribution of any stochastic process* (in Russian), Czechoslovak Math. J., 8 (83) (1958), pp. 610–618. (A translation appears in American Mathematical Society Translations in Probability and Statistics, 1961.)

[4] P. Halmos, *Measure Theory*, Van Nostrand, New York, 1950.

[5] T. Kailath, *Correlation detection of signals perturbed by a random channel*, Trans. I.R.E., IT-6 (1960), pp. 361–366.

[6] S. Kullback, *Information Theory and Statistics*, John Wiley, New York, 1959.

[7] E. Parzen, *Statistical inference on time series by Hilbert space methods, I*, Department of Statistics, Stanford University, Technical Report No. 23, January 2, 1959.

[8] E. Parzen, *Regression analysis of continuous parameter time series*, Proceedings of the Fourth Berkeley Symposium on Probability and Mathematical Statistics, Vol. I, University of California Press, 1961.

[9] E. Parzen, *An approach to time series analysis*, Ann. Math. Statist., 32 (1961), pp. 951–989.

[10] E. Parzen, *Probability density functionals and reproducing kernel Hilbert spaces*, to be published in Proceedings of a Symposium on Time Series Analysis, John Wiley, New York, 1963.

[11] E. Parzen, *Stochastic Processes*, Holden Day, San Francisco, 1962.

[12] E. Parzen, *Foundations of Time Series Analysis and Statistical Communication Theory*, Holden Day, San Francisco, to be published in 1964.

[13] C. R. Rao, *Advanced Statistical Methods in Biometric Research*, John Wiley, New York, 1952.

[14] F. Riesz and B. Sz-Nagy, *Functional Analysis*, Blackie, London, 1956.

[15] G. L. Turin, *On optimal diversity reception*, Trans. I.R.E., IT-7 (1960), pp. 154–166.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 1
Printed in U.S.A., 1962

ON A NEW PARTIAL DIFFERENTIAL EQUATION FOR THE STABILITY ANALYSIS OF TIME INVARIANT CONTROL SYSTEMS*

G. P. SZEGÖ†

1. Introduction. The investigation of the stability properties of the equilibrium point of a control system poses various problems which, even if conceptually very similar, vary greatly in difficulty and in the methods appropriate for their solutions.

The first and easiest problem is what we may call the stability analysis of a *completely defined system*: given a particular control system, decide what stability properties this equilibrium point has.

The second problem deals with a system having a *fixed configuration*, but with parameters whose numerical values are to be determined. The problem is to find the boundaries in the parameter space where the stability properties of the system undergo a change.

The third problem is that of synthesizing stable systems and its solution implies knowledge of necessary and sufficient conditions that the equilibrium point of the system be stable. This latter problem is far from being solved and it is also doubtful if its practical solution will emerge from the classical theory of stability.

For the solution of the second problem, various precise techniques [1, 2, 3] have been proposed for the construction of suitable Lyapunov functions. These techniques have solved many stability problems, but there are systems for which they have failed. Each of these systems constitutes then a mathematical problem by itself and Lyapunov functions must be constructed more or less by inspection [4, 5, 7]. All these methods may be applied to the solution of the first problem.

This paper presents some new techniques leading to a new approach to the problem of stability analysis within the framework of Lyapunov's direct method. The potentiality of this new approach has not been completely exploited in this paper and further results will be presented later.

This study is limited to the investigation of completely defined systems. The new method for stability investigation which is the final outcome of this work, will in principle always yield some solution of the stability problem. The price we have to pay for assuring that the method always works is the restriction to a particular class of Lyapunov functions. The Lyapunov functions are solutions of a partial differential equation (19) which turns out to be the generalization of an analogous equation proposed by Zubov [8].

* Received by the editors August 3, 1962.

† Instituto di Meccanica Applicata del Politecnico, Milano. Visitor at the Research Institute of Advanced Studies (RIAS), Baltimore, Maryland.

2. Notations and terminology. We shall use standard vector notation with the following conventions: capital letters are matrices, small Latin letters are vectors, Greek letters and small Latin letters with subscripts are scalars, except that t, v are scalars.

We shall denote by (x, y) the inner product of the two vectors x and y.

We shall denote the transpose of a matrix A by A'.

We shall call an equilibrium state that is stable but not asymptotically stable *weakly stable*.

We say that a scalar function $\phi = \phi(x)$ is *positive (negative) definite on the trajectories of a system* in a region $S \subset E_n$ if $\phi(x) \geqq 0 (\phi(x) \leqq 0)$ in S and $\phi(x) \not\equiv 0$ on any nonsingular solution of the system. By a Lyapunov function we mean any scalar function which gives the answer to the stability properties of a solution of a system.

Unless otherwise stated, we shall assume throughout that all the scalar functions we use have continuous first partial derivatives.

3. Lyapunov's second method. Lyapunov's second method can be codified in a set of theorems [6, 9, 10, 11, 12] which prove that if for a given system there exists a scalar function with certain properties, called a Lyapunov function, then one can draw conclusions about the stability properties of the solutions of the system. There also exist a set of inverse theorems [6, 12] which guarantee the existence of such a function.

Given the nonlinear autonomous dynamic system [13]

$$\dot{x} = f(x), \qquad f(0) = 0, \tag{1}$$

where $f(x)$ is a vector valued function, the problem of the stability analysis of its equilibrium point is then formally reduced to the search for a positive definite [14] scalar function $\psi = \psi(x)$ and a scalar function $v = v(x)$, $v(0) = 0$, such that the partial differential equation

$$\psi(x) = -(\operatorname{grad} v(x), f(x)) = -\sum_{i=1}^{n} \frac{\partial v}{\partial x_i} f_i(x) \tag{2}$$

is satisfied.

From the form of the scalar function $v(x)$, one is able to draw conclusions about the stability properties of the solution $x = 0$ of (1) and about the range of these properties [15, 16].

The major problem is then to find a definite scalar function $\psi(x)$, such that the solution $v = v(x)$ of (2) satisfies the condition $v(0) = 0$. The inverse theorems [6, 12] guarantee that such scalar functions $\psi(x)$ and $v(x)$ exist. The stability problem is then reduced to the problem of finding necessary and sufficient conditions that a given scalar function $\psi(x)$ be positive definite. The existence and the possibility of using such conditions is however much in doubt.

A different and perhaps more sensible approach is that of finding a sufficient condition which guarantees that a scalar function $\psi(x)$ is at least definite on the trajectories of (1) and such that there always exists a function $\psi(x)$ satisfying this condition and a corresponding $v = v(x)$, $v(0) = 0$, which satisfies (2).

Now it is clearly possible to fix the form of $\psi(x)$ completely, and hence reduce the stability problem strictly to the integration of a well defined linear partial differential equation of the type (2). This approach however does not have much practical interest, because the integration of this equation usually requires the solution of (1).

The method we develop in this work overcomes these difficulties. This method is essentially based upon two major steps. The first is to change the state variable x of the system to z as defined in the relations (33)–(38). The second step is to study the solution of the nonlinear partial differential equation (41). The function $\theta(w_2) = \theta(z_i)$ which appears on the right-hand side of this equation is any scalar function which depends on only one particular component of the new state vector z. If the scalar function $\theta(w_2)$ is definite and no degeneracy occurs, then the solution of (41) will yield a Lyapunov function of (1).

In the following sections we assume that the problem of *analyzing* a scalar function, that is the problem of deciding if a given scalar function is positive or (negative) definite, positive (negative) semidefinite or indefinite is solved. In fact, various techniques are available for this analysis (reduction to a sum of squares, geometrical studies, etc.), and in each particular case this question can usually be answered without difficulty.

4. A generalization of Zubov's equation. In this section we shall make some remarks on the form of the scalar function $\psi(x)$ defined in (2). These remarks will constitute the basis of our method.

We pose the following question. If an arbitrary scalar function $v_1 = v_1(x)$ is given and

$$\psi_1(x) = (\text{grad } v_1(x), f(x)), \tag{3}$$

under what conditions on $\psi_1(x)$ is it possible to compute from the scalar function $v_1 = v_1(x)$ a new scalar function $v_2 = v_2(x)$, such that $\psi_2(x) = (\text{grad } v_2(x), f(x))$ has a certain required form?

Suppose that we are able to synthesize a scalar function having the property of being definite along the trajectories of (1). Then the problem of constructing a Lyapunov function from (2) can be thought of as that of integrating a Pfaffian differential equation. We shall see that this point of view is quite rewarding.

Consider the Pfaffian form

$$\omega(x) = (y(x), dx) \tag{4}$$

and the corresponding Pfaffian differential equation

$$(y(x), dx) = 0. \tag{5}$$

Let $\mu(x)$ be an integrating factor of this equation, that is

$$\mu(x)y(x) = \operatorname{grad} v(x). \tag{6}$$

The scalar function $v = v(x)$ is then a particular integral of (5), that is,

$$dv = (\operatorname{grad} v(x), dx). \tag{7}$$

Consider now an arbitrary function $\alpha = \alpha(v)$. From (5) it follows that

$$\mu(x)\frac{d\alpha}{dv}(y(x), dx) = 0. \tag{8}$$

By substituting (6) into (8) we have

$$\frac{d\alpha}{dv}(\operatorname{grad} v(x), dx) = 0 \tag{9}$$

which from (7) is identically equal to

$$\frac{d\alpha}{dv}dv = d\alpha = 0. \tag{10}$$

These results can be summarized in the following

THEOREM 4.1 *If $\mu(x)$ is an integrating factor of the Pfaffian differential equation* (4) *with solution $v = v(x)$ and $\alpha = \alpha(v)$ is an arbitrary scalar function, then $\mu(x)\, d\alpha/dv$ is also an integrating factor of* (4) *with solution $\alpha = \alpha(v(x)) = \alpha^*(x)$.*

The results of Theorem 4.1 are even more obvious if applied to (4).

In §3 we have shown that the stability problem is reduced to the search for scalar functions $v = v(x)$, $v(0) = 0$ and $\psi = \psi(x)$ positive definite on the trajectories of (1). Then the problem may be reduced to that of seeking a scalar function $\alpha(v)$ such that

$$\frac{d\alpha}{dv}\mu(x)\omega(x) = \psi(x) \tag{11}$$

or

$$\frac{d\alpha(v)}{dv(x)} = \frac{\psi(x)}{\mu(x)\omega(x)}. \tag{12}$$

The functional equation (12) can be readily solved if

$$\frac{\psi(x)}{\mu(x)\omega(x)} = \beta(v(x)). \tag{13}$$

On the basis of these considerations we can develop the following procedure.

Take a scalar function $v_1 = v_1(x)$, $v_1(0) = 0$ and compute its total time derivative

$$\frac{dv_1}{dt} = (\text{grad } v_1(x), f(x)) = \gamma(x), \qquad \gamma(0) = 0. \tag{14}$$

Next look for a scalar function $\psi(x)$ which is at least definite on the trajectories of (1) and a scalar function $\beta(v_1)$, $\int_0^{v_1} \beta(s)\, ds < \infty$, such that

$$\frac{\psi(x)}{\gamma(x)} = \beta(v_1). \tag{15}$$

Then the differential equation

$$\frac{d\alpha(v_1)}{dv_1} = \frac{\psi(x)}{\gamma(x)} = \beta(v_1) \tag{16}$$

may be integrated. Its solution $\alpha = \alpha(v_1) = \alpha^*(x)$ will be such that

$$\alpha^* = \text{grad } \alpha^*(x) f(x) = \psi(x) \tag{17}$$

which is at least definite on the trajectories of (1), and because of the assumptions made on $\alpha(v_1)$ and $v_1(x)$, it will solve our stability problem.

By substituting (15) into (14) we obtain

$$\frac{dv_1}{dt} = (\text{grad } v_1(x), f(x)) = \frac{\psi(x)}{\beta(v_1)} \tag{18}$$

which is a generalization of Zubov's equation [8]:

$$(\text{grad } v, f(x)) = \phi(x)(1 + v). \tag{19}$$

The stability theorem we deduce from (18) may be stated as

THEOREM 4.2. *The stability problem of the solution* $x = 0$ *of* (1) *is reduced to finding scalar functions* $v_1(x)$, $\psi(x)$, $\beta(v_1)$ *such that* $v_1(0) = 0$, $\int_0^{v_1} \beta(s)\, ds < \infty$ *and* $\psi(x)$ *is definite on the trajectories of* (1).

Let us now integrate equation (16),

$$\alpha(v_i) = \int_0^{v_1} \beta(s)\, ds \tag{20}$$

from which we deduce

THEOREM 4.3. *The solution* $x = 0$ *of* (1) *is asymptotically stable in a closed, bounded region* S: $\alpha^*(x) \leqq \delta$, $\delta > 0$, *if there exist scalar functions* $v_1(x)$, $\psi(x)$, $\beta(v_1)$ *satisfying the conditions*

(i) $v_1(0) = 0$,

(ii) $\psi(x)$ *is negative definite on the trajectories of* (1),

(21) (iii) $\int_0^{v_1} \beta(s)\, ds < \infty$,

(iv) $\alpha^*(x) = \int_0^{v_1} \beta(s)\, ds > 0$ in S, $x \neq 0$, $\alpha^*(0) = 0$,

and equation (18).

COROLLARY 4.1. *The solution* $x = 0$ *of* (1) *is asymptotically stable in the large if all the conditions of Theorem* 4.3 *are satisfied and*

$$\lim_{\|x\| \to \infty} \alpha^*(x) = \lim_{\|x\| \to \infty} \int_0^{v_1} \beta(s)\, ds = \infty. \tag{22}$$

COROLLARY 4.2. *If all the conditions of Theorem* 4.3 *are satisfied with the sign of* $\psi(x)$ *changed, then the solution* $x = 0$ *of* (1) *is completely unstable* [17].

REMARK 4.1. It is always possible to give sufficient conditions for complete instability, from any theorem on asymptotic stability. In the following sections we are going to present numerous theorems on asymptotic stability, and it is always implied, even if not explicitly stated, that a similar theorem for completely instability holds.

Since given any scalar function $v_2 = v_2(x)$ it is always possible to find a functional $\Omega = \Omega(v_2(x))$, $\Omega(0) = 0$, such that the scalar function $\Omega^* = (x) = \Omega(v_2(x))$ is semidefinite, we can develop the following simplified procedure for constructing Lyapunov's functions.

Let us, first of all, seek a scalar function $v_2 = v_2(x)$, $v(0) = 0$, such that

$$\frac{dv_2}{dt} = (\operatorname{grad} v_2(x), f(x)) = \theta(v_2) \tag{23}$$

where $\theta = \theta(v_2)$ is a bounded scalar function. Equation (20) is a special case of (17).

Then it is always possible to integrate the equation

$$\frac{d\alpha_2(v_2)}{dv_2} = \frac{\Omega(v_2)}{\theta(v_2)} = \beta_2(v_2) \tag{24}$$

and its solution $\alpha_2 = \alpha_2(v_2) = \alpha_2{}^*(x)$ will be such that

$$\frac{d\alpha_2}{dt} = (\operatorname{grad} \alpha_2{}^*(x), f(x)) = \Omega^*(x) \tag{25}$$

where $\Omega^*(x)$ is semidefinite. If no degeneracy occurs and if

$$\int_0^{v_2} \beta_2(x) < \infty \tag{26}$$

then the scalar function $x_2 = \alpha_2{}^*(x)$ is a Lyapunov function of (1).

Theorems (4.1) and (4.2) and corollaries apply with minor changes to this case.

The following theorem answers the question of the existence of solutions of (23).

THEOREM 4.4. *There always exists a scalar function* $\theta(v_2)$ *such that* (23) *has a solution which satisfies the condition* $v_2(0) = 0$. *In particular if* (1) *is asymptotically stable* $v_2(x)$ *is definite and* $\theta(v_2)$ *may be chosen so that* $\theta(v_2) = \lambda v_2$, $\operatorname{Re}\{\lambda\} < 0$.

Proof. This theorem may be proved directly following a method used in [18], or by using the inverse theorem [6, 12]. Let us sketch a proof using this latter way for the case of asymptotic stability.

Given any Lyapunov function $v = v(x)$, $v(0) = 0$, $\dot{v}(x) \not\equiv 0$ for $x \not\equiv 0$, $v(0) = 0$, this function represents a hypersurface in E_{n+1} with a strong minimum for $x = 0$.

Take any section of this hypersurface with the hyperplane $v = \text{const.} \not\equiv 0$. This section is a closed bounded hypersurface. Let us represent it in parametric form

$$x_i = x_i(t_1, \cdots, t_{n-1}). \tag{27}$$

Since $\dot{v}(x) > 0$ for $x \not\equiv 0$ it is possible to construct a unique integral surface $S = v^*(x)$ of the equation

$$\dot{v}^*(\operatorname{grad} v^*(x), f(x)) = -v^*(x) \tag{28}$$

going through the hypersurface (27).

Consider the characteristic system of (28)

$$\frac{dv^*}{-v^*} = dt, \qquad \frac{dx_i}{f_i(x)} = dt, \qquad i = 1, \cdots, n. \tag{29}$$

Since (1) is by assumption asymptotically stable, the solutions $x_i = x_i(t)$, $v^* = v^*(t)$ of (29) tend to zero as $t \to \infty$. Hence $v^*(x)|_{x=0} = 0$.

REMARK 4.2. An existence theorem analogous to Theorem 4.4 may be proved also for (18). We notice that the necessary and sufficient conditions that the scalar function $\psi(x)$ must satisfy for (18) to have a solution $v_1 = v_1(x)$ such that $v_1(0) = 0$ are not identical to the conditions on $\phi(x)$ in Zubov's equation (19). In this latter case,

$$\int_0^t \phi(x(\tau))\, d\tau < \infty, \tag{30}$$

a relation which in our case is only sufficient.

It is worthwhile to emphasize the major differences between (18) and (23). As we previously pointed out the v-functions $\alpha = \alpha^*(x)$ obtained from (18) are Lyapunov functions for (1). The only possible degenerate case in which v is semidefinite will never arise.

In fact if $\alpha = \alpha^*(x)$ is continuous and semidefinite then it has a strong minimum on the manifold M on which $\alpha^*(x) = 0$. Hence grad $\alpha^*(x) = 0$ on M and M is an integral manifold of (1). This contradicts the hypothesis that $\psi(x)$ is definite on the trajectories of (1). This reasoning does not apply to (23) where $\Omega^*(x)$ is only semidefinite. Then $\alpha^*(x) = \Omega^*(x) = 0$ on some integral manifold N. All the information we obtain in this case concerns the stability of the manifold N and not of the equilibrium point. In some cases then the procedure must be repeated in order to find a new $\alpha_1 = \alpha_1^*(x)$ satisfying an equation of the type (17) which may be semidefinite as long as there does not exist a point $x = x_e \neq 0$ on which $\alpha^*(x_e) = \alpha_1^*(x_e) = 0$. In this case one not only obtains information about the stability properties of the equilibrium point, but one is also able to find some integral of the system. The same situation may arise from (18) if $\psi(x)$ is a semidefinite function.

Equation (23) is very important in itself since its solutions are the so-called isochrones of (1). The knowledge of the isochrones gives important information about the qualitative behavior of the solutions of (1).

REMARK 4.3. For some systems it may happen that the scalar function $\psi(x)$ in (18) or $\Omega^*(x)$ in (20) is identically zero. The scalar function $\alpha = \alpha^*(x)$ is then a first integral of the system.

5. A useful change of variables. The method for constructing Lyapunov functions developed in the previous section contrasts strongly with the methods in use up to now. The essence of our method is the introduction of the functions $\beta(v_1)$ and $\theta(v_2)$ respectively in (18) and (23). The major step is now to find a scalar function $v = v(x)$, $v(0) = 0$, such that dv/dt has the form $\theta_2(v)$ or $\psi(x)/\beta(v)$, but is otherwise completely arbitrary.

If we think of the problem of constructing a Lyapunov function by solving a partial differential equation we see that the original linear partial differential equation (2) (with unknown right hand side) has become a quasi-linear partial differential equation (18) or (23) whose right hand side has a certain well defined form.

In this paragraph we shall perform a particular transformation of variables $x \Rightarrow z$. One of the components of the next state vector z is the scalar function v.

The aim of this transformation is to make full use of the very well defined form of the right-hand sides of (18) and (23). The stability problem will then be reduced to a search for a scalar function $\xi = \xi(z)$, satisfying a certain nonlinear partial differential equation whose right-hand side is any definite function which depends on only one particular component of the vector z.

Consider the scalar equation

$$w = v(x), \qquad v(0) = 0,$$

which is equivalent to the equation

$$V(x, w) = v(x) - w = 0. \tag{31}$$

Under certain very mild conditions which a v-function satisfies [6, 12], we may solve this equation with respect to an arbitrary component of the vector x, say x_i. We obtain

$$x_i = \xi_i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n, w). \tag{32}$$

Since we shall perform this operation only once in the whole procedure we may as well simplify the notations. Let us introduce the n-vector z defined as follows:

$$\begin{aligned} z_k &= x_k, && k \neq i, \\ z_i &= w. \end{aligned} \tag{33}$$

We may write (32) as

$$x_i = \xi_i(z) \tag{34}$$

so that

$$v(x_1, \cdots, x_{i-1}, \xi_i(z), x_{i+1}, \cdots, x_n) - w = 0 \tag{35}$$

from which

$$\begin{aligned} \frac{\partial v}{\partial x_j} + \frac{\partial v}{\partial x_i}\frac{\partial \xi_i}{\partial x_j} &= 0, && i \neq j, \\ \frac{\partial v}{\partial x_i}\frac{\partial \xi_i}{\partial w} &= 1, \end{aligned} \tag{36}$$

and finally

$$\begin{aligned} \frac{\partial v}{\partial x_i} &= \frac{1}{\partial \xi_i / \partial w}, \\ \frac{\partial v}{\partial x_j} &= -\frac{\partial \xi_i / \partial x_j}{\partial \xi_i / \partial w}, && i \neq j. \end{aligned} \tag{37}$$

Now we are ready to perform the transformation of coordinates

$$x \Rightarrow z : x_i = \xi_i(x) \tag{38}$$

on (18) and (23).

These equations take respectively the form

$$\left(f_i(x) - \sum_{j=1}^{n} \frac{\partial \xi_i}{\partial z_j} f_j(x)\right)\Bigg|_{x_i = \xi_i} = \frac{\psi(x)}{\beta(x_1)}\Bigg|_{x_i = \xi_i} \cdot \frac{\partial \xi_i}{\partial w_1}, \qquad i \neq j, \tag{39}$$

and

$$\left(f_i(x) - \sum_{j=1}^{n} \frac{\partial \xi_i^*}{\partial z_j} f_j(x)\right)\Big|_{x_i=\xi_i} = \theta(w_2)\,\frac{\partial \xi_i^*}{\partial w_2}, \qquad i \neq j. \tag{40}$$

The problem of finding $w = v(x)$ is now reduced to the integration of the nonlinear partial differential equations of the type (39) or (40), with the side condition $\xi_i(0) = 0$.

We are in particular interested in (40) which can be written as

$$\left(f_i(x) - \sum_{j=1}^{n} \frac{\partial \xi_i^*}{\partial z_j} f_j(x)\right)\Big|_{x_i=\xi_i} \cdot \frac{1}{\partial \xi_i^*/\partial w_2} = \theta(w_2), \qquad j \neq 1. \tag{41}$$

In this latter equation the usefulness of our method is much emphasized.

We can see that the only requirement we have is that the right-hand side of (41) depends only on w_2. If the function $\beta(w_2)$, defined by (24) satisfies (26), the problem of the stability of the solution $x = 0$ of the system under investigation is solved. If one cannot find a scalar function $\alpha(w_2)$ such that $\alpha(w_2)/\theta(w_2)$ is bounded, then it is still possible to study the stability of some first integral of the system, going through the origin.

In other words, whatever the function $\theta(w_2)$ is we shall always be able to have some answer about the stability properties of the system.

Unfortunately the solution of the partial differential equation is not a simple matter and it is possible to integrate it explicitly only if it is possible to integrate its characteristic system. A more reasonable approach is to choose a suitable form for the unknown scalar function $\xi_i = \xi_i(x)$, $\xi_i(0) = 0$, having a certain number of unknown coefficients, then compute the unknown coefficients in such a way that the right-hand side of (41) is a function which depends only on w_2.

The possibility of doing this depends of course on the right choice of the form ξ_i^*. Although numerous examples have been solved, no general information about suitable forms for ξ_i^* is available.

6. Examples. In the case of nonlinear systems, if one is not able to integrate the characteristic system of the equations (39) or (40), then an alternative procedure is to look for a suitable form of the unknowns $v = v(x)$ or $\xi_i(z, w)$ which allows a separation of variables.

Consider for example the system

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -ay - ax^3 - x^2 y \end{aligned} \tag{42}$$

for which

$$v_1 = ax + y \tag{43}$$

and

$$\dot{v}_1 = -x^2 v_1 . \tag{44}$$

We see that the solution $y = -ax$ is asymptotically stable. By integrating it one obtains $x = \exp(-ta)$.

We conclude that the equilibrium point $x = y = 0$ is asymptotically stable for $a > 0$ and unstable for $a < 0$. In this particular case we have solved the stability problem directly from (43) and (44). This is not always the case.

Consider for example

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= ax + ax^2 y - y^3 - y, \qquad a > 0. \end{aligned} \tag{45}$$

Assume $v = ax^2 - y^2$; then $\dot{v} = 2y^2(1 - v)$ from which no conclusion can be drawn. Choose

$$\psi(x) = 2y^2(1 - \dot{v})2;$$

then

$$\begin{aligned} \alpha(v) &= \int_0^v \beta(s)\, ds = \int_0^v (1 - s)\, ds = v - \tfrac{1}{2}v^2, \\ \alpha^*(x) &= (ax^2 - y^2) - \tfrac{1}{2}(ax^2 - y^2)^2, \\ \dot{\alpha}^*(x) &= 2y^2(1 - v)^2 \end{aligned}$$

from which we conclude that $x = y = 0$ is unstable.

In the next example we shall illustrate what is the advantage of (41) with respect to the other formulations. Consider the system

$$\begin{aligned} \dot{x} &= y^3 - x, \\ \dot{y} &= x - \tfrac{1}{2}y \end{aligned} \tag{46}$$

for which (41) takes the form

$$\left[y^3 - \xi - \frac{\partial \xi}{\partial y}(\xi - \tfrac{1}{2}y)\right] \frac{1}{\partial \xi / \partial w} = \theta(w). \tag{47}$$

Assume

$$\xi = a(w + f(y))^{\frac{1}{2}} \tag{48}$$

for which (47) becomes

$$\begin{aligned} \frac{2}{a} y^3 (w + f(y))^{\frac{1}{2}} - 2w - 2f(y) - a \frac{\partial f(y)}{\partial y} (w + f(y))^{\frac{1}{2}} \\ + \tfrac{1}{2} y \frac{\partial f(y)}{\partial y} = \theta(w) \end{aligned} \tag{49}$$

from which we may set

$$\theta(w) = -2w, \tag{50}$$

$$-2f(y) + \tfrac{1}{2}y\frac{\partial f}{\partial y} = 0. \tag{51}$$

$$\frac{2}{a}y^3 = a\frac{\partial f}{\partial y} = 0. \tag{52}$$

Intregating (51) we obtain

$$f(y) = \pm y^4, \tag{53}$$

and from (52)

$$f(y) = y^4 \quad \text{and} \quad a^2 = \tfrac{1}{2}. \tag{54}$$

By substituting these results into (40) we obtain

$$\xi = \frac{1}{\sqrt{2}}(w + y^4)^{\frac{1}{2}} \tag{55}$$

which may also be written in the usual form

$$v = 2x^2 - y^4, \tag{56}$$

and we may check that

$$\dot{v} = \theta(w)v = -2v.$$

We conclude that the solution $2x^2 = y^4$ is asymptotically stable.

Let us investigate this solution by integrating it. We obtain

$$x = \pm\frac{1}{2\sqrt{2}}t^2 + x_0. \tag{57}$$

We conclude that the solution $x = 0$ of the (46) is unstable.

7. Conclusion. In the present work we have solved the general problem of stability analysis of the equilibrium point of a control system represented in (1).

This general problem has been reduced to the integration of partial differential equations (23), (39), (18) or (41).

The closed-form integration of (23) or (30) presents, in the case of non-linear systems, the same difficulties as the equations which arise from otpimal control problems; hence in most of the cases numerical methods must be applied. On the other hand, for equations of the form (18) and (41) various examples with closed-form solutions have been found.

8. Acknowledgment. The author wishes to thank Dr. J. P. LaSalle and Dr. R. E. Kalman of the Research Institute of Advanced Studies of

Baltimore, Maryland for very helpful comments and suggestions. Part of this research was carried out while the author was at the Control and Information Systems Laboratory of the School of Electrical Engineering, Purdue University and supported by National Science Foundation Grant No. G-16460. The author wishes to thank Professor J. E. Gibson, Director of the Control and Information Systems Laboratory, who made possible the author's stay at Purdue, for his invaluable support.

REFERENCES

[1] G. P. Szegö, *A contribution to Lyapunov's second method: nonlinear autonomous systems*, Trans. ASME Ser. D. J. Basic Engrg. (To appear)

[2] D. R. Ingwerson, *A modified Lyapunov method for nonlinear stability analysis*, Trans. I.R.E., PGAC-6 (1961), pp. 199–210.

[3] G. P. Szegö, *On the application of the Zubov method for construction of Liapunov's functions for nonlinear control systems*, Trans. ASME Ser. D. J. Basic Engrg. (To appear)

[4] A. I. Lur'e, *Nekotorye nelineinye zadachi teorii abtomaticheskogo regulirovania*, Gostechteorechizdat, Moskow, 1951. (German Translation: Akademic Verlag, Berlin, 1957.)

[5] A. M. Letov, *Ustoichivoct' nelineinykh reguliruemykh sistem*, Gostechteorechizdat, Moscow, 1955. (English translation: Princeton University Press, Princeton, 1961.)

[6] N. N. Krasovskii, *Neketorye zadachi teorii ustoichivosti dvizheniia*, Fizmatigiz, Moscow, 1959.

[7] D. G. Schultz and J. E. Gibson, *The variable gradient method for generating Lyapunov functions*, Trans. A.I.E.E. (To appear)

[8] V. I. Zubov, *Voprocy teorii vtorogo metoda Liapunova*, Prikl. Mat. Meh., 19(1955), pp. 179–210.

[9] A. Liapunov, *Probléme Genéral de la Stabilitè du Mouvement*, Annals of Math. Studies, Vol. 17, Princeton,

[10] J. P. LaSalle and S. Lefschetz, *Stability of Liapunov's Direct Method*, Academic Press, New York, 1961.

[11] H. A. Antosiewicz, *Survey of Lyapunov's Second Method*, Annals of Math. Studies, No. 41, Princeton, 1958.

[12] J. L. Massera, *Contributions to stability theory*, Annals of Math., 64(1956), pp. 182–206.

[13] R. E. Kalman and J. E. Bertram, *Control system analysis and design via the second method of Lyapunov*, J. Basic Engr., Trans. A.S.M.E. Series D, 82(1960), pp. 371–393.

[14] F. R. Gantmacher, *Theory of Matrices*, Chelsea, New York, 1959.

[15] J. P. LaSalle, *Some extensions of Lyapunov's second method*, Trans. I.R.E., PGCT-7 (1960), pp. 520–527.

[16] E. A. Barbashin and N. N. Krasovskii, *O suschectvovanii funktsii Liapunova v sluchne asimptoticheskoi ustoichivosti v tselom*, Prikl. Mat. Meh., 18(1954), pp. 345–350.

[17] W. Hahn, *Theorie und Anwendung der Direkton Methode von Lyapunov*, Springer-Verlag, Berlin, 1959.

[18] I. Vrkoc, *Ob obrashchenii teopomy chetaeva*, Czech. Math. J., 5(1955), pp. 451–461.

[19] I. G. Malkin, *Teoriia Ustoichivosti Dvizheniia*, Gostechteorechizdat, Moskow, 1962. (English translation: AEC transl. no. 3352, 1958.)

J.SIAM CONTROL
Ser. A, Vol. 1, No. 1
Printed in U.S.A., 1962

ON CERTAIN QUESTIONS IN THE THEORY OF OPTIMAL CONTROL*

A. F. FILIPPOV†

I. The existence of a solution of the optimal control problem within the class of bounded, measurable functions. Given a system of n equations (in vector notation)

$$\frac{dx}{dt} = f(t, x, u), \tag{1}$$

where x and f are n-dimensional vectors, and $u = u(t)$ is the control parameter (an r-dimensional vector) which, for any given t and x, can take on values in a given set $Q(t, x)$ (in [1] the case where Q depends neither on t nor on x is considered). The optimal control problem consists of the following: for given $x^{(0)}$ and x^*, find a function $u(t)$ such that the solution $x(t)$ of (1), with u set equal to $u(t)$, and initial condition $x(0) = x^{(0)}$, attains the point x^* in the least possible time, where, in addition, $u(t) \in Q(t, x(t))$.

We shall assume that the vector function $f(t, x, u)$ is continuous in all of the variables t, x, and u, is continuously differentiable with respect to x, and that

$$x \cdot f(t, x, u) \leqq C(|x|^2 + 1), \tag{2}$$

for all t and x, and all $u \in Q(t, x)$, where the dot denotes the scalar product, and $|x|$ denotes the length of the vector x. Let $Q(t, x)$ be closed and bounded. When u describes $Q(t, x)$, $f(t, x, u)$ describes a set which we shall denote by $R(t, x)$. We shall assume that $Q(t, x)$ is upper semicontinuous (in t and x) with respect to inclusion; i.e., that for any t, x and $\epsilon > 0$ there exists a $\delta = \delta(\epsilon, t, x) > 0$ such that $Q(t', x')$ is contained in an ϵ-neighborhood of $Q(t, x)$ whenever $|t' - t| < \delta$ and $|x' - x| < \delta$. Then, $R(t, x)$ will have the same semicontinuity property (on account of the continuity of f).

THEOREM 1. *Suppose that the conditions stated above are satisfied, and that the set $R(t, x)$ is convex for every t and x. Also suppose that there exists at least one measurable function $\tilde{u}(t) \in Q(t, \tilde{x}(t))$ such that the solution*[1]

* Originally published in Vestnik Moskov. Univ. Ser. Mat. Mech. Astr., 2 (1959) pp. 25–32. Translated by L. W. Neustadt, Aerospace Corporation, El Segundo, California.

From time to time the editors will include translations of important foreign papers on control.

† Chair of Differential Equations, Moscow State University, U.S.S.R.

[1] An absolutely continuous vector function which satisfies (1) almost everywhere will be called a solution. Under the assumptions we have made, such a solution exists by a theorem of Carathéodory [2, page 140].

$\tilde{x}(t)$ of (1), with $u = \tilde{u}(t)$, and initial condition $\tilde{x}(0) = x^{(0)}$, attains x^* for some $t^* > 0$. Then there also exists an optimal control, i.e., a measurable function $u(t) \in Q(t, x(t))$ for which the solution $x(t)$ of (1), with initial condition $x(0) = x^{(0)}$, attains x^* in the least possible time.

Proof. It follows from (1) and (2), and the fact that $x(0) = x^{(0)}$, that

$$\frac{dX}{dt} \leqq 2CX, \qquad X(0) = |x^{(0)}|^2 + 1 = A$$

for almost all t, where $X(t) = |x(t)|^2 + 1$. Consequently, $X(t) \leqq Ae^{2Ct}$ for $t \geqq 0$. Hence, for every admissible $u(t)$, the solution of (1) with initial condition $x(0) = x^{(0)}$ satisfies the inequality $|x(t)| \leqq A^{\frac{1}{2}}e^{Ct^*}$ on the interval $0 \leqq t \leqq t^*$ (if we consider that $C \geqq 0$).

We shall show that the $Q(t, x)$ are uniformly bounded for $0 \leqq t \leqq t^*$ and for $|x| \leqq A^{\frac{1}{2}}e^{Ct^*}$; i.e., that there exists an N such that $|u| \leqq N$ for every $u \in Q(t, x)$ for the indicated values of t and x. Indeed, otherwise there would be sequences $t_n \to t$, $x_n \to x$, and $u_n \in Q(t_n, x_n)$ such that $|u_n| \to \infty$. But, by the assumptions we have made, $Q(t_n, x_n)$ is contained in an ϵ-neighborhood of the bounded set $Q(t, x)$, for sufficiently large n. This contradiction shows that the $Q(t, x)$ are uniformly bounded for the indicated values of t and x. Since f is continuous, $|f(t, x, u)| \leqq M$ if $0 \leqq t \leqq t^*$, $|x| \leqq A^{\frac{1}{2}}e^{Ct^*}$, $u \in Q(t, x)$.

Now consider the set of all the solutions $x(t)$ of (1) (for various $u(t) \in Q(t, x(t))$), for which $x(0) = x^{(0)}$ and $x(t') = x^*$, $0 < t' \leqq t^*$. The numbers t' can differ for different solutions. Since one such solution exists by hypothesis, this set is not empty. If this set is finite, the assertion of the theorem is obvious. If it is infinite, we shall select a sequence from this set such that the sequence of the corresponding t' converges to t_1—the greatest lower bound of all such t'. The solutions of this sequence are uniformly bounded for $0 \leqq t \leqq t^*$ since $|x(t)| \leqq A^{\frac{1}{2}}e^{Ct^*}$. These solutions are also equicontinuous since they are absolutely continuous and $|dx/dt| = |f(t, x, u)| \leqq M$ almost everywhere. Let us choose a uniformly convergent subsequence $x_1(t)$, $x_2(t)$, $\cdots$ from this sequence, and let us denote the limit of this subsequence by $x(t)$. We shall show that $x(t)$ is the solution which corresponds to an optimal control $u(t)$.

Making use of the equicontinuity of the solutions $x_1(t)$, $x_2(t)$, $\cdots$, we conclude that $x(0) = x^{(0)}$ and $x(t_1) = x^*$. In addition, there exists no solution for which $x(0) = x^{(0)}$ and $x(t) = x^*$, with $0 < t < t_1$. Further, since all the $x_i(t)$ satisfy a Lipschitz condition with the same constant M, their limit $x(t)$ satisfies the same condition. Hence, $x(t)$ is absolutely continuous, and $|dx(t)/dt| \leqq M$. Let

$$\frac{dx(t)}{dt} = y(t), \qquad \frac{dx_i(t)}{dt} = y_i(t), \qquad i = 1, 2, \cdots.$$

The functions $y(t)$ and $y_i(t)$ are defined almost everywhere on $[0, t_1]$, are measurable, and are bounded. Let t_0 be a point of $[0, t_1]$ for which $dx(t_0)/dt$ exists. We shall show that then $dx(t_0)/dt \in R(t_0, x(t_0))$.

The set $R(t, x)$ is upper semicontinuous (u.s.c.) with respect to inclusion. Consequently, for any $\epsilon > 0$, there is a $\delta > 0$ such that $R(t, x) \subseteq U_\epsilon$ whenever $|t - t_0| < \delta$ and $|x - x(t_0)| < 2M\delta$, where U_ϵ is a closed ϵ-neighborhood of $R(t_0, x(t_0))$. Making δ smaller, if necessary, we may suppose that

$$\left| \frac{x(t) - x(t_0)}{t - t_0} - \frac{dx(t_0)}{dt} \right| < \epsilon \tag{3}$$

whenever $|t - t_0| < \delta$. But

$$\begin{aligned} \frac{x(t) - x(t_0)}{t - t_0} &= \lim_{i\to\infty} \frac{x_i(t) - x_i(t_0)}{t - t_0} = \lim_{i\to\infty} \frac{1}{t - t_0} \int_{t_0}^{t} y_i(\tau)\, d\tau \\ &= \lim_{i\to\infty} \int_0^1 y_i(t_0 + (t - t_0)s)\, ds. \end{aligned} \tag{4}$$

Furthermore,

$$y_i(\tau) = dx_i(\tau)/dt = f(\tau, x_i(\tau), \quad u_i(\tau)) \in R(\tau, x_i(\tau))$$

for almost all τ. For all sufficiently large i and $|\tau - t_0| < \delta$, we have

$$|x_i(t_0) - x(t_0)| < M\delta, \qquad |x_i(\tau) - x_i(t_0)| < M\delta.$$

Consequently, $|x_i(\tau) - x(t_0)| < 2M\delta$. But then, from the previously made estimates, $R(\tau, x_i(\tau)) \subseteq U_\epsilon$. Thus, if $|t - t_0| < \delta$, and for all sufficiently large i, the integrand in (4) is contained (for almost all s) in U_ϵ. Hence, the whole of the integral (4) is also contained in U_ϵ (U_ϵ being a convex set). Consequently, the left-hand side in formula (4) is also contained in U_ϵ. Now it follows from (3) that $dx(t_0)/dt$ is contained in a 2ϵ-neighborhood of $R(t_0, x(t_0))$. Since ϵ is arbitrarily small and R is closed,

$$\frac{dx(t_0)}{dt} \in R(t_0, x(t_0)).$$

Hence, there exists a $u \in Q(t_0, x(t_0))$ such that $dx(t_0)/dt = f(t_0, x(t_0), u)$. Thus, for all t for which the vector function $x(t)$ has a derivative, i.e., for almost all $t \in [0, t_1]$, $x(t)$ satisfies (1) where $u(t) \in Q(t, x(t))$. It follows from the lemma proved below that $u(t)$ can be considered to be measurable.

LEMMA. *Let the vector function $f(t, u_1, \cdots, u_r)$, or, more concisely, $f(t, u)$, be continuous; let the set $Q(t)$ be closed, bounded, and u.s.c. with respect to inclusion (in t); let the vector $f(t, u)$ describe a set $R(t)$ when the*

vector $u = (u_1, \cdots, u_r)$ *describes the set* $Q(t)$*; and let* $y(t)$ *be a measurable vector function such that* $y(t) \in R(t)$*. Then there exist measurable functions* $u_1(t), \cdots, u_r(t)$ *such that* $f(t, u_1(t), \cdots, u_r(t)) \equiv y(t)$ *for almost all* t*.*

Proof. For a given value $y \in R(t)$, we shall always take that $u = (u_1, \cdots, u_r)$, from among all the values of the vector $u \in Q(t)$ which satisfy the equation $f(t, u) = y$, for which the coordinate u_1 has the smallest value. If there is more than one such u, we shall take that one for which the coordinate u_2 has the smallest value, etc. (the smallest value exists since, on account of the continuity of the function f, the set of values u which satisfy the equation $f(t, u) = y$ is closed). We shall prove by induction that the functions $u_1(t), \cdots, u_r(t)$ are measurable. Let us suppose that $u_1(t), \cdots, u_{s-1}(t)$ are measurable (if $s = 1$, nothing need be assumed), and let us prove that $u_s(t)$ is measurable. There exists a closed set $E \subseteq [0, t_1]$, of measure greater than $t_1 - \epsilon$, such that the functions $y(t), u_1(t), \cdots, u_{s-1}(t)$ are continuous on E (this is a property of measurable functions). Let us show that, for any number a, the set of $t \in E$ for which $u_s(t) \leqq a$ is closed. Suppose the contrary. Then there is a sequence $t_n \in E$, $n = 1, 2, \cdots$, such that

$$t_n \to \bar{t} \in E, \qquad u_s(t_n) \leqq u_s(\bar{t}) - \epsilon_1, \qquad \epsilon_1 > 0. \tag{5}$$

Since $|u_i(t)| \leqq$ const. for all i and t, a subsequence t_{n_m} can be chosen from the t_n such that $u_i(t_{n_m}) \xrightarrow[m\to\infty]{} \tilde{u}_i$, for $i = 1, 2, \cdots, r$. Since $u(t_n) \in Q(t_n)$, and $Q(t)$ is closed and u.s.c. with respect to inclusion, $(\tilde{u}_1, \cdots, \tilde{u}_r) = \tilde{u} \in Q(\bar{t})$. It follows from (5) and the continuity on E of the functions $u_i(t)$, $i = 1, 2, \cdots, s - 1$, that

$$\begin{aligned} &\tilde{u}_i = u_i(\bar{t}), && i = 1, \cdots, s - 1, \\ &\tilde{u}_s \leqq u_s(\bar{t}) - \epsilon_1. \end{aligned} \tag{6}$$

Passing to the limit in the identity $f(t, u_1(t), \cdots, u_r(t)) \equiv y(t)$ with respect to the above chosen subsequence, and making use of the continuity of the function f, we obtain

$$f(\bar{t}, u_1(\bar{t}), \cdots, u_{s-1}(\bar{t}), \tilde{u}_s, \cdots, \tilde{u}_r) = y(\bar{t}). \tag{7}$$

By (6) and (7), $u_s(\bar{t})$ is not the smallest value u_s which satisfies the equation

$$f(\bar{t}, u_1(\bar{t}), \cdots, u_{s-1}(\bar{t}), u_s, \cdots, u_r) = y(\bar{t}).$$

This contradicts the definition of $u_s(t)$. Thus, our assumption is false, and the set of $t \in E$ for which $u_s(t) \leqq a$ is closed.

Hence, $u_s(t)$ is measurable on E. Since $E \subseteq [0, t_1]$, and the measure of E is greater than $t_1 - \epsilon$, where ϵ is arbitrarily small, $u_s(t)$ is measurable

on $[0, t_1]$. Hence, by induction, we conclude that all the $u_i(t)$, $i = 1, \cdots, r$, are measurable.

Note. In actual mechanical systems, of course, it is impossible to realize a control described by a function $u(t)$ if the function is merely measurable without being piecewise continuous. In spite of this, Theorem 1 does not lose interest since any measurable function $u(t)$ can be approximated by a continuous function $v(t)$ such that

$$\int_{t_0}^{t_1} | u(t) - v(t) | \, dt < \delta, \tag{8}$$

where δ is arbitrarily small.[2] If, in (1), we replace the optimal control $u(t)$ by a continuous control $v(t)$ which satisfies (8), then, by virtue of the theorem [3] on the continuous dependence of a solution on a parameter, the solution $x(t)$, although it may not attain the given point x^* at $t = t_1$, will come arbitrarily close to x^* if the number δ in (8) is made sufficiently small. Thus, in many cases, one may even get along with continuous controls.

II. Piecewise smoothness condition on the control functions $u_s(t)$. Let system (1) have the form

$$\frac{dx_i}{dt} = g_i(t, x_1, \cdots, x_n) + \sum_{s=1}^{r} b_{is}(t)u_s(t), \quad i = 1, \cdots, n, \tag{9}$$

where the g_i, $\partial g_i/\partial x_k$, and b_{is} are continuous, and the $u_s(t)$ must satisfy the conditions

$$| u_s(t) | \leqq 1, \qquad s = 1, \cdots, r. \tag{10}$$

By Theorem 1, there exist measurable bounded functions $u_s(t)$ which satisfy (10) for which the solution of system (9), with initial conditions $x_i(0) = x_{i0}$, $i = 1, \cdots, n$, attains the point $(x_1^*, \cdots, x_n^*)$ the most quickly of all. By [4], the functions $u_s(t)$ satisfy the "maximum principle",

$$f(t, x, u) \cdot \psi(t) = \max., \tag{11}$$

where $f = (f_1, \cdots, f_n)$ is the complete right-hand side of (9) in vector notation, $\psi = (\psi_1, \cdots, \psi_n)$ is a solution of the system

$$\frac{\partial \psi_i}{\partial t} = -\sum_{k=1}^{n} \frac{\partial f_k}{\partial x_i} \psi_k, \qquad i = 1, \cdots, n, \tag{12}$$

and the maximum is taken with respect to all the u_s which satisfy (10).

For the system (9), (10), the expression in (11) attains its maximum

[2] If we assume that $u(t) \in Q$, where the set Q is connected, locally connected, and does not depend on t or x, the function $v(t)$ in formula (8) can be chosen in such a way that $v(t) \in Q$.

when

$$u_s(t) = \operatorname{sgn} b_s(t) \cdot \psi(t), \qquad s = 1, \cdots, r, \tag{13}$$

where $b = (b_{1s}, \cdots, b_{ns})$. It is assumed that $b_s(t) \cdot \psi(t) = 0$ only for values of t which form a set of measure zero. If the values of t for which $b_s(t) \cdot \psi(t) = 0$ have no point of accumulation, the function $u_s(t)$ in (13) is piecewise continuous. Points of accumulation can only be those t where the relations $b_s(t) \cdot \psi(t) = 0$ and $d(b_s(t) \cdot \psi(t))/dt = 0$ hold simultaneously (here we assume that db_s/dt is continuous). Hence, we can conclude that, generally speaking, for the majority of solutions $\psi(t)$ of (12) such points of accumulation will not exist, and, therefore, for the majority of optimal trajectories $x(t)$, the controls $u_s(t)$ will be piecewise continuous.

If the system is linear with analytic (in particular, constant, as in [5]) coefficients and b_{is}, and if $b_s(t) \cdot \psi(t) \not\equiv 0$, then $b_s(t) \cdot \psi(t)$ can vanish only at isolated points, and, consequently, an optimal control is piecewise continuous. The verification of the conditions $b_s(t) \cdot \psi(t) \not\equiv 0$, $s = 1, \cdots, r$, may be carried out relatively easily, since it suffices to verify if these equalities hold in a neighborhood of the point $t = 0$.

Analogous results for linear systems have been obtained previously by other methods by N. N. Krasovskii [6] and R. V. Gamkrelidze; the original proof of Gamkrelidze's [5, pp. 472–474] was replaced by a proof due to the author of this article.

III. The case where the optimal control is continuous. Suppose that the conditions of section I are satisfied, and, in addition, that the set $R(t, x)$ is strictly convex[3] and depends continuously on t and x; i.e., for every $\epsilon > 0$, there exists a $\delta > 0$ such that if $|t_1 - t_2| < \delta$ and $|x_1 - x_2| < \delta$ then the distance from any point of $R(t_i, x_i)$ to the set $R(t_j, x_j)$ is less than δ (for $i = 1, j = 2$ or $i = 2, j = 1$). Suppose that to each value $v \in R(t, x)$ there corresponds only one $u \in Q(t, x)$ for which $f(t, x, u) = v$. Then, the value of u at which the maximum in (11) is attained will depend continuously on t.

Indeed, the maximum is attained at that u for which the point $f(t, x, u) = v \in R(t, x)$ lies on the support plane to the set $R(t, x)$ which is perpendicular to the vector ψ. By virtue of the strict convexity of $R(t, x)$, this point v depends continuously on ψ. Under the assumptions we have made, u depends continuously on v. From this it is easy to show that the optimal control $u(t)$ is continuous.

IV. The general formulation of the optimal control problem. There is given the system (1) in the notation of section I; a closed set A in the

[3] That is, every support plane to this set has only one point in common with the set.

hyperplane $t = t_0$; closed sets B and D in the half-space $t \geqq t_0$; and a closed set $Q(t, x)$, in the space of the u, which depends on t and x. It is required to find a point $x_0 \in A$ and vector functions $x(t)$ and $u(t)$ which satisfy (1) and the conditions $x(t_0) = x_0$, $(t, x(t)) \in D$, $u(t) \in Q(t, x(t))$, $(t_0 + T, x(t_0 + T)) \in B$, such that the number T is minimal.

This formulation differs from the formulation of the problem in [1] in that the set of admissible values of u, i.e., $Q(t, x)$, may depend on t and x; that the solution $x(t)$ must not leave the given region D, and must get not from one point to another point in the shortest time, but must get from one given set A onto another given set B.

If A is bounded and the conditions of section I are satisfied, then we can make an assertion analogous to the assertion of Theorem 1 concerning the existence of an optimal control within the class of bounded and measurable functions. This can be proved in the same way as Theorem 1. It is only necessary to add that the limit of a minimizing sequence of curves $x_n(t)$ which belong to the closed region D is a curve which also belongs to this region.

The formulation of the problem given above contains, as a special case, the optimal control problems as formulated in [1], [5]–[7]. The formulation of the problem of [1] has been given in section I. In [7] the requirement that the solution $x(t)$ should not leave a given region D was added.

In one of the problems described in [5], it is in addition required that some of the control functions, e.g., $u_{p+1}(t), \cdots, u_r(t)$, have derivatives which are bounded by a given constant. Let us show that this case is also included in the general statement of the problem given above. Let us add the following $r - p$ equations to the given system (1):

$$\frac{du_s}{dt} = v_s, \qquad s = p + 1, \cdots, r.$$

Now there are $n + r - p$ functions $x_1, \cdots, x_n, u_{p+1}, \cdots, u_r$ in the system, which are being sought. The last $r - p$ of these are subject to definite constraints (e.g., constraints of the form $| u_s(t) | \leqq N$, if such were the constraints imposed on the u_s when they were considered to be control parameters). Further, there are r control parameters $u_1, \cdots, u_p, v_{p+1}, \cdots, v_r$. The same constraints as before are imposed on the first p of these. The last $(r - p)$ parameters must satisfy the constraints $| v_s(t) | \leqq L$, $s = p + 1, \cdots, r$. The initial conditions for the desired functions will be as follows: $x_i(t_0) = x_{i0}$, $i = 1, \cdots, n$; $u_s(t_0)$ arbitrary, $s = p + 1, \cdots, r$, so long as they satisfy the imposed constraints (e.g., $| u_s(t_0) | \leqq N$). Thus, the set A (see the beginning of Section IV) is no longer a point. The set B is similarly defined.

The problem in which there is imposed the additional requirement that

the derivatives (up to any given order) of the $u_s(t)$ must be bounded, can be reduced by the same method to the formulation given at the beginning of section IV.

In [6] constraints of the form

$$\int_{t_0}^{t_0+T} [u_s(t)]^2\, dt \leqq C$$

are imposed on the control parameters. In this case, the functions

$$x_{n+s}(t) = \int_{t_0}^{t} [u_s(\tau)]^2\, d\tau, \qquad s = 1, \cdots, r,$$

can be added to the functions of system (1) as functions to be solved for. Further, these additional functions are subject to the constraints $0 \leqq x_{n+s}(t) \leqq C$ and satisfy the equations

$$\frac{dx_{n+s}(t)}{dt} = [u_s(t)]^2, \qquad x_{n+s}(t_0) = 0, \qquad s = 1, \cdots, r.$$

The $u_s(t)$ remain as control parameters, but they are now subject to no constraints; i.e., $Q(t, x)$ is the entire space. The set $R(t, x)$ may also be unbounded, but if it is closed, convex, and u.s.c. with respect to inclusion (see section I), the proof given in section I on the existence of an optimal control remains in force (with small changes). However, in this way, only the square integrability of the functions $u_s(t)$ for which the control is optimal can be proved. In [6] their continuity is proved by other means.

V. On sliding regimes. Let us show that an optimal control may not exist if the set $R(t, x)$ in Theorem 1 is not convex. Consider the problem

$$\begin{gathered} \frac{dx}{dt} = -y^2 + u^2, \qquad \frac{dy}{dt} = u; \qquad |u(t)| \leqq 1; \\ x(0) = y(0) = 0; \qquad x(T) = 1, \qquad y(T) = 0, \qquad T > 0; \\ T = \min. \end{gathered} \tag{14}$$

Since $dx/dt \leqq 1$, $T \geqq 1$. It is easily seen that $T > 1$ for any solution of system (14), and that any sequence of solutions for which $|u_n(t)| = 1$ and $|y_n(t)| \leqq 1/n$ is a minimizing sequence since, for such a sequence, $x_n(t_n) = 1$ for $1 < t_n < 1 + 1/(n^2 - 1)$. Further, any minimizing sequence converges to the functions $x(t) = t$, $y(t) = 0$ which are not a solution of system (14) for any $u(t)$. Thus, an optimal control does not exist here.

In cases similar to the one considered, it is said that the minimizing sequence converges to a sliding regime, where the following meaning is assigned to this. Consider the problem (14). In order to obtain a solution for which the time T is arbitrarily near the minimum (but unattainable)

value $T = 1$, it is necessary that $u(t)$ pass sufficiently often from values near $+1$ to values near -1. More precisely, let $\epsilon_n \to 0$ and $\delta_n \to 0$, and, for an arbitrary time interval of duration greater than δ_n, let the fraction of the values of t for which $| u_n(t) - 1 | < \epsilon_n$ differ from $\frac{1}{2}$ by less than ϵ_n, and let the same situation hold for $| u_n(t) + 1 | < \epsilon_n$. Then, such a sequence $u_n(t)$ is minimizing.

Both in the example we have considered and in the case of more general equations of the form (1), sliding regimes arise from the nonconvexity of $R(t, x)$. It is clear from the proof of Theorem 1 that a sliding regime appears when the vector $dx(t)/dt$ does not belong to $R(t, x)$ on some interval $\alpha < t < \beta$. Then, the vector belongs to the complement of $R(t, x)$ with respect to a convex set. Here, $x(t)$ is the limit of the convergent subsequence of the minimizing sequence, as in section I.

We note that in [8, p. 103] results similar to Theorem 1 of this paper were obtained, but under somewhat different assumptions.

REFERENCES

[1] V. G. Boltyanskii, R. V. Gamkrelidze and L. S. Pontryagin, *On the theory of optimal processes*, Dokl. Akad. Nauk SSSR., 110, (1956) 7–10. (English translation in Report STL-T-Ru-22-60-5111-102, Space Technology Laboratories, Redondo Beach, California.)

[2] G. Sansone, *Equazioni Differenziali nel Campo Reale*, 2d ed., Vol. 2, Zanichelli, Bologna, 1949.

[3] J. Kurzweil and Z. Vorel, *Continuous dependence of solutions of differential equations on a parameter*, Czechoslovak Math. J., 7 (1957), pp. 568–583. (Russian, English summary.)

[4] R. V. Gamkrelidze, *On the general theory of optimal processes*, Dokl. Akad. Nauk SSSR., 123 (1958), pp. 223–226. (English translation in Automation Express, 1 (1959), pp. 37–39.)

[5] R. V. Gamkrelidze, *The theory of time-optimal processes for linear systems*, Izvestia Akad. Nauk SSSR., Ser. Mat., 22 (1958), pp. 449–474. (English translation in Report No. 61-7, University of California, Department of Engineering, Los Angeles, California.)

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

[7] A. Ya. Lerner, *On the limiting speed of response in automatic control systems*, Avtomat. i Telemeh., 15 (1954), pp. 461–477. (Russian.)

[8] R. Bellman, I. Glicksberg and O. Gross, *Some Aspects of the Mathematical Theory of Control Processes*, The Rand Corporation, Santa Monica, California, 1958.

[9]* E. Roxin, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[10]* E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[11]* J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

* References added by translator.

J.S.I.A.M. CONTROL
Ser. A, Vol. 1, No. 2
Printed in U.S.A., 1963

ON THE NONLINEAR CONTROL PROBLEM WITH CONTROL APPEARING LINEARLY*

H. HERMES† AND G. HAYNES†

1. Introduction. The problem of optimal control can be viewed as follows: given the system of ordinary differential equations

$$\dot{x}_i(t) = F_i(t, x(t), u(t)), \qquad x_i(0) = x_i^0, \qquad i = 1, 2, \cdots, m \tag{1}$$

where $\dot{x}_i = dx_i/dt$, choose the control vector $u = (u_1, \cdots, u_r)$ from a given class U to extremize a given functional, which may depend on the control u, and the corresponding solution φ^u of (1).

This paper is concerned with equations in which the control appears linearly, i.e.,

$$F_i(t, x, u) \equiv g_i(t, x) + h_{ij}(t, x)u_j\,, \tag{2}$$

a sum being taken over the repeated index j. This form is particularly fruitful in view of the existence results obtained by Markus and Lee in [10]. (See also [4] and [15].) For convenience the control set U is chosen to be the set of vector valued functions u, defined by

$$U = \{u : |\, u_j(t)\,| \leqq 1, \quad u_j \text{ measurable}, \quad t \in [0, \infty), \quad j = 1, 2, \cdots, r\}.$$

A control function is called a *bang-bang* control if it belongs to U and has the additional property $|\, u_j(t)\,| = 1$ for $j = 1, 2, \cdots, r$, $t \geqq 0$. Controls that are not bang-bang will be referred to as intermediate controls.

In [6] LaSalle has shown that if

$$F_i(t, x, u) \equiv a_{ij}(t)x_j + b_{iv}(t)u_v$$

then the set of points $\varphi^u(t)$, for some $t \geqq 0$ and $u \in U$, is equal to the set of points $\varphi^u(t)$ attainable by bang-bang control. Thus it can be concluded that for the time optimal problem, it is always possible to do as well with bang-bang control as with an arbitrary control, if the system is linear. It is shown in section 2 of this paper that if F is of the form given in (2), then the set of points $\varphi^u(t)$ with u a bang-bang control is dense (as a subset of Euclidean n space) in the set of points $\varphi^u(t)$ with u an arbitrary control. This result by no means implies that it suffices to study bang-bang controls only, since from a practical viewpoint, the number of switchings necessary to approach optimality with a bang-bang system may far exceed physical limitations.

* Received by the editors October 27, 1962.
† The Martin Company, Denver, Colorado.

The remainder of the paper considers the problem of synthesis of the optimal control for a system of the form

$$\begin{aligned} \dot{x}_1(t) &= A_1(x(t)) + B_1(x(t))u(t) \\ \dot{x}_2(t) &= A_2(x(t)) + B_2(x(t))u(t), \qquad x(0) = x^0, \end{aligned} \tag{3}$$

where $x = (x_1, x_2)$ and $U = \{u: | u(t) | \leqq 1, u \text{ measurable on } [0, \infty)\}$. The measure is always assumed to be Lebesgue measure. The problem is to determine that $u \in U$ which minimizes a functional of the form

$$\int_0^{t_f} L(x(\tau))\, d\tau$$

where t_f is the time a prescribed final state is attained. The results can immediately be generalized to an integrand function of the form $L_1(x) + L_2(x)u$. In particular if $L \equiv 1$, this is the time optimal problem.

In section 2, under the restriction that $\Delta(x) \equiv -B_2(x)A_1(x) + B_1(x)A_2(x) \neq 0$ in some set together with other suitable restrictions, it is possible to construct a set $S \subset E^2$ which is shown to be the region in which solutions could exist. Because of the form of the equations (3) and the functional to be minimized, this problem is particularly suited for solution by the Green's theorem approach; a method due to A. Miele [11]. By this method it is possible to obtain global conditions for optimality, and thus, in special cases, to resolve the singular control problem (section 3.2) that can arise when the coefficient of the control appearing in the Hamiltonian [8, Theorem 1] vanishes over some time interval of positive measure. In section 3, the optimal strategy deducible from the Green's theorem approach is discussed. In particular it is shown that if one can determine the optimal path in phase space, and this path can be realized as a trajectory of the system (3) with a control $u \in U$, then the condition $\Delta \neq 0$ along this path insures that u is unique to within a set of measure zero. An example of a time optimal problem is then given in which the optimal path requires an intermediate control, in fact $u \equiv 0$. The uniqueness of this control yields the result that although u enters the equations linearly, the optimal time cannot be attained via bang-bang control.

1. BANG-BANG CONTROL IN THE NON-LINEAR TIME OPTIMAL PROBLEM

In this section we consider the system of ordinary differential equations

$$\dot{x}(t) = g(t, x(t)) + H(t, x(t))u(t), \qquad x(0) = x^0 \tag{1.1}$$

where $g(t, x)$ is an m-vector with components $g_i(t, x)$, $u(t)$ is an r vector with components $u_i(t)$, while $H(t, x)$ is an $m \times r$ matrix with elements denoted by $h_{ij}(t, x)$.

The set of admissible control functions U is defined by

$$U \equiv \{u : |u_j(t)| \leqq 1, \quad u_j \text{ measurable on } [0,\infty),\ j=1, 2, \cdots, r\}$$

while we define

$$U^0 \equiv \{u \in U : |u_j(t)| = 1, \qquad j = 1, 2, \cdots, r\},$$

so that U^0 is the set of bang-bang control functions.

Let R^m denote real m-dimensional space, and for $y \in R^m$ define $\| y \| = \sum_{i=1}^m |y_i|$. The following assumptions are imposed on the functions g and H appearing in (1.1): for $i = 1, 2, \cdots, m; j = 1, 2, \cdots, r$;

i) $g_i(t, x)$, $h_{ij}(t, x)$ satisfy a Lipschitz condition in x uniformly with respect to t, of the form

$$\begin{aligned} |g_i(t, x) - g_i(t, \bar{x})| &\leqq K \| x - \bar{x} \|, \\ |h_{ij}(t, x) - h_{ij}(t, \bar{x})| &\leqq K \| x - \bar{x} \|, \end{aligned} \tag{1.2}$$

ii) $|h_{ij}(t, x)| \leqq M$, and

iii) $g_i(t, x)$, $h_{ij}(t, x)$ are measurable in t for each fixed x.

For simplicity it is assumed that these conditions hold for $(t, x) \in [0, \infty) \times R^m$. With these assumptions the system (1.1) has a unique solution in the extended sense (in the class of absolutely continuous functions) for any $u \in U$, which is denoted φ^u.

Define, for some $t \geqq 0$,

$$\begin{aligned} S(t) &\equiv \{\varphi^u(t) : u \in U\}, \\ S^0(t) &\equiv \{\varphi^u(t) : u \in U^0\}. \end{aligned}$$

Then $S(t)$ is the set of points in R^m attainable from x^0 at time t with an arbitrary control, while $S^0(t)$ is the set of points in R^m which can be attained with bang-bang control.

THEOREM 1.1. *If for the system* (1.1) *the conditions* (1.2) *are satisfied, then for every* $t \geqq 0$, $S^0(t)$ *is dense in* $S(t)$.

Proof. Let $u \in U$ and $\bar{u} \in U^0$. Then the corresponding solutions φ^u and $\varphi^{\bar{u}}$ of (1.1) satisfy

$$\begin{aligned} (\varphi^u(t) - \varphi^{\bar{u}}(t)) \equiv &\int_0^t [g(\tau, \varphi^u(\tau)) - g(\tau, \varphi^{\bar{u}}(\tau))]\, d\tau \\ &+ \int_0^t [H(\tau, \varphi^u(\tau)) - H(\tau, \varphi^{\bar{u}}(\tau))]\bar{u}(\tau)\, d\tau \\ &+ \int_0^t H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}(\tau)]\, d\tau. \end{aligned}$$

By using the Lipschitz continuity of g and H and the fact that $|\bar{u}_j(t)| = 1$,

one can readily obtain

$$\| \varphi^u(t) - \varphi^{\bar{u}}(t) \| \leqq mK(r+1) \int_0^t \| \varphi^u(\tau) - \varphi^{\bar{u}}(\tau) \| \, d\tau + \left\| \int_0^t H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}(\tau)] \, d\tau \right\|. \tag{1.3}$$

It now suffices to show that for $\epsilon > 0$ and any given t^* and $u \in U$, it is possible to find a $\bar{u} \in U^0$ such that $\|\varphi^u(t^*) - \varphi^{\bar{u}}(t^*)\| < \epsilon$.

The procedure will be to show that $\bar{u} \in U^0$ can be constructed so that

$$\left\| \int_0^t H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}(\tau)] \, d\tau \right\| < \epsilon. \tag{1.4}$$

With this shown, the required result follows from (1.3) by use of the Gronwall inequality.

For given t^*, break the interval $[0, t^*]$ in n equal subintervals, each of length δ. By Lemma 2, in the paper [6] of LaSalle, it follows that for any $1 \leqq j \leqq n$, there exists a function $\bar{u}^j$ such that

$$|\bar{u}_i{}^j(t)| = 1, \qquad i = 1, 2, \cdots, r, t \in [(j-1)\delta, j\delta]$$

and

$$\left\| \int_{(j-1)\delta}^{j\delta} H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}^j(\tau)] \, d\tau \right\| = 0.$$

We define $\bar{u}$ on the interval $(0, t^*]$ by

$$\bar{u}(t) = \bar{u}^j(t), \qquad (j-1)\delta < t \leqq j\delta, j = 1, 2, \cdots, n.$$

Then for any $t \in (0, t^*]$, $(j-1)\,\delta < t \leqq j\delta$ for some $1 \leqq j \leqq n$ and

$$\left\| \int_0^t H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}(\tau)] \, d\tau \right\| = \left\| \int_0^{(j-1)\delta} H(\tau, \varphi^u(\tau)) [u(\tau) - \bar{u}(\tau)] \, d\tau + \int_{(j-1)\delta}^t H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}(\tau)] \, d\tau \right\| = \left\| \int_{(j-1)\delta}^t H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}(\tau)] \, d\tau \right\|.$$

Now since the elements of H are bounded by M and the length of the interval $[(j-1)\delta, t]$ is less than δ,

$$\left\| \int_0^t H(\tau, \varphi^u(\tau))[u(\tau) - \bar{u}(\tau)] \, d\tau \right\| \leqq 2Mmr\delta.$$

By choosing $\delta = \epsilon/2Mmr$, or equivalently $n = 2t^*Mmr/\epsilon$, (1.3) yields

$$\| \varphi^u(t) - \varphi^{\tilde{u}}(t) \| \leqq \epsilon + mK(r+1) \int_0^t \| \varphi^u(\tau) - \varphi^{\tilde{u}}(\tau) \| \, d\tau.$$

The Gronwall inequality can now be applied, giving

$$\|\varphi^u(t) - \varphi^{\bar{u}}(t)\| \leqq \epsilon \exp \{mK(r+1)t\}$$

for any $t \in [0, t^*]$, which completes the proof.

From Theorem 1.1 it immediately follows that if $R^0(x^0)$ is the attainable set from x^0 with bang-bang controls, i.e. $R^0(x^0) \equiv \bigcup_{t \geqq 0} S^0(t)$, and $R(x^0) \equiv \bigcup_{t \geqq 0} S(t)$ is the attainable set from x^0, then $R^0(x^0)$ is dense in $R(x^0)$. It has been shown that for the problem (1.1) $S^0(t)$ is dense in $S(t)$ and it might be conjectured that $S^0(t)$ is equal to $S(t)$. In particular, the construction led to a control $\bar{u} \in U^0$ such that not only was $\| \varphi^u(t^*) - \varphi^{\bar{u}}(t^*) \|$ arbitrarily small for given $u \in U$ and t^*, but $\| \varphi^u(t) - \varphi^{\bar{u}}(t) \|$ was small for all $0 \leqq t \leqq t^*$, which certainly is not a necessary restriction. However, using the synthesis method described in sections 2 and 3 [in particular see Theorem 3.1 and Example 3-7] we show that in some cases $S^0(t) \neq S(t)$.

It will be useful in what follows, to be able to approximate measurable control functions with continuous controls. The required results are given in

THEOREM 1.2. *If for the system* (1.1), *the conditions* (1.2) *are satisfied let*

$$S(t^*) \equiv \{\varphi^u(t^*) : u \in U\}, t^* \geqq 0$$

$$S^c(t^*) \equiv \{\varphi^u(t^*) : u \in U, u \text{ continuous on } [0, t^*]\}.$$

Then $S^c(t^*)$ *is dense in* $S(t^*)$. *Further, if given any* $\epsilon > 0$ *and control* $u \in U$, *there exists a continuous control* $\tilde{u} \in U$ *such that*

$$\| \varphi^u(t) - \varphi^{\tilde{u}}(t) \| < \epsilon, \qquad 0 \leqq t \leqq t^*.$$

Proof. Since u_j is measurable on $[0, t^*]$ for each j and $| u_j(t)| \leqq 1$, there exists (see [12], p. 106), for any given $\delta > 0$, a continuous function $\tilde{u}$ such that $| \tilde{u}_j(t)| \leqq 1$ and the measure of the set $\{t : u_j(t) \neq \tilde{u}_j(t)\}$ is less than δ. It follows that given $u \in U$, $t^* \geqq 0$ and any $\epsilon > 0$, there exists a continuous control $\tilde{u}$, $| \tilde{u}_j(t)| \leqq 1$, such that

$$\left\| \int_0^t H(\tau, \varphi^u(\tau))[u(\tau) - \tilde{u}(\tau)] \, d\tau \right\| < \epsilon$$

for $0 \leqq t \leqq t^*$. Using this and the Gronwall inequality in (1.3), the required result is obtained.

2. THE TWO-DIMENSIONAL PROGRAM

2.1. Formulation of the problem. In the following sections the synthesis for a two-dimensional system with control function appearing linearly is

considered. The differential equations are taken to be of the form

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

with the condition $x(0) = x^0$, where

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad A(x) = \begin{pmatrix} A_1(x) \\ A_2(x) \end{pmatrix}, \qquad B(x) = \begin{pmatrix} B_1(x) \\ B_2(x) \end{pmatrix},$$

while u is a scalar valued control function. It is assumed that A_1, A_2, B_1 and B_2 are once continuously differentiable in an open, simply connected set $D \subset R^2$. The initial point x^0 and terminal point x^f will always be considered to be in D.

We define the set of admissible control functions U as

$$U \equiv \{u: \mid u(t)\mid \leqq 1, \quad u \text{ measurable}, \quad t \in [0, \infty)\},$$

and denote by $\varphi^u(\cdot, x^0)$ the unique solution of (2.1), in the class of absolutely continuous functions, for control $u \in U$, such that $\varphi^u(0, x^0) = x^0$. No distinction will be made between controls which differ only on a set of zero measure.

In the development to follow, it will often be of interest to consider constant control functions, in which case the system (2.1) is autonomous. To emphasize the constant control, we will use φ^α to designate the solution of (2.1) for $u(t) \equiv \alpha$, $-1 \leqq \alpha \leqq 1$.

Also, since for any given $u \in U$, our interest is only in values of t for which $\varphi^u(t, x^0) \in D$, the notation $T(u, x^0)$ will be used to denote the largest interval of non-negative real numbers containing zero such that if $t \in T(u, x^0)$, $\varphi^u(t, x^0) \in D$.

The problem considered is to find that $u \in U$, such that $\varphi^u(t_f, x^0) = x^f$, which minimizes the functional

$$C(u, x^0, x^f) \equiv \int_0^{t_f} L(\varphi^u(\tau, x^0))\, d\tau$$

where L is a given once continuously differentiable function in D, and t_f is the first time the given state x^f is attained. It has been shown by Markus and Lee [10] that if for some $u \in U$ and some $t \geqq 0$, $\varphi^u(t, x^0) = x^f$, then there exists an optimal control. This result is a special case of a more general existence theorem obtained independently in [13] and [4].

2.2. The attainable set for the problem (2.1). Before the synthesis of the optimal control can be discussed, a study of the attainable set for the problem (2.1) is necessary.

Define

$$R(x^0) \equiv \{x \in R^2: x = \varphi^u(t, x^0) \quad \text{for some} \quad t \in T(u, x^0), u \in U\},$$
$$R(x^f) \equiv \{x \in R^2: x = \varphi^u(-t, x^f) \quad \text{for some} \quad t \in T(u, x^f), u \in U\}$$

where $\varphi^u(-t, x^f)$ is the solution of (2.1) with $\dot{x}$ replaced by $-\dot{x}$. Thus $R(x^0)$ denotes the set of points which can be attained from x^0, while $R(x^f)$ denotes the set of points from which x^f can be attained.

Obviously, if a solution to the optimal control problem for (2.1) exists, the segment of the trajectory connecting x^0 to x^f must lie in $R(x^0) \cap R(x^f)$. Also if $R(x^0) \cap R(x^f) \neq \varnothing$, the empty set, there is a control $u \in U$ such that an arc of the trajectory φ^u joins x^0 and x^f (i.e. existence). The remainder of this section will, therefore, be a discussion of this set.

Our goal will be to obtain sufficient conditions that the trajectories $\varphi^1(\cdot, x^0)$, $\varphi^{-1}(\cdot, x^0)$, $\varphi^1(\cdot, x^f)$ and $\varphi^{-1}(\cdot, x^f)$ determine $R(x^0) \cap R(x^f)$. To show the sense in which these trajectories may bound $R(x^0) \cap R(x^f)$ we define for any $y \in D$, $|\alpha| \leqq 1$, the vector $\xi(\alpha, y) \equiv (A_1(y) + \alpha B_1(y), A_2(y) + \alpha B_2(y))$. Thus the possible directions which a solution trajectory to (2.1) can assume at the point y are given by $\{\xi(\alpha, y): |\alpha| \leqq 1\}$.

Let

$$\Delta(y) \equiv -B_2(y)A_1(y) + A_2(y)B_1(y). \tag{2.2}$$

Note that $\Delta(y) \neq 0$ implies y is *not* a critical point of (2.1) for any $|u| \leqq 1$.

Let $\theta(\alpha, y)$ be the angle traced out by the ray $\xi(\sigma, y)$ as σ varies continuously from -1 to α. The angle will be called positive if it is traced out in a counterclockwise direction, and negative if in a clockwise direction.

LEMMA 2.1 *If* $\Delta(y) \neq 0$, *the set* $\{\xi(\alpha, y): |\alpha| \leqq 1\}$ *of possible directions is bounded by* $\xi(-1, y)$ *and* $\xi(1, y)$ *with* $0 < |\theta(1, y)| < \pi$.

Proof. For any $-1 \leqq \alpha \leqq 1$, $\xi(\alpha, y)$ lies on the line segment joining $\xi(-1, y)$ and $\xi(1, y)$, since we can write $\xi(\alpha, y) = ((\alpha + 1)/2)\xi(1, y) + ((1 - \alpha)/2)\xi(-1, y)$. Thus $\xi(-1, y), \xi(1, y)$ bound $\{\xi(\alpha, y): |\alpha| \leqq 1\}$.

Letting $|\xi(\alpha, y)|$ denote the length of the vector $\xi(\alpha, y)$, the condition $\Delta(y) \neq 0$ implies $|\xi(\alpha, y)| \neq 0$ and that $\xi(-1, y)$ and $\xi(\alpha, y)$ cannot be parallel for any $-1 < \alpha \leqq 1$, thus $0 < |\theta(1, y)| < \pi$.

In view of this lemma the directions $\xi(1, x^0)$ and $\xi(-1, x^0)$ bound the set of possible directions at x^0, and the angle $\theta(1, x^0)$, which we may assume for the sake of this discussion to be positive, is such that $0 < \theta(1, x^0) < \pi$. The next lemma will show that if we were to observe the angle $\theta(1, \varphi^1(t, x^0))$ as t increases from zero, the condition $\Delta(\varphi^1(t, x_0)) \neq 0$ will insure that the sign of θ will not change. Also $\Delta(\varphi^{-1}(t, x^0)) \neq 0$ implies the invariance of the sign of $\theta(1, \varphi^{-1}(t, x^0))$. Intuitively one would expect that all possible trajectories are restricted to a wedge-shaped region bounded by $\varphi^1(\cdot, x^0)$ and $\varphi^{-1}(\cdot, x^0)$. We will now proceed to show this.

LEMMA 2.2 *Let* $\gamma(\sigma)$, $\sigma_0 \leqq \sigma \leqq \sigma_f$, *be a continuous curve in* D *along which* $\Delta(\gamma(\sigma)) \neq 0$; *then* sgn $\theta(1, \gamma(\sigma))$ *is invariant along the curve.*

Proof. Since $\theta(\gamma(\sigma))$ is a continuous function of σ, if it changes sign there would be a value $\sigma_1 \in [\sigma_0, \sigma_f]$ such that $\theta(\gamma(\sigma_1)) = 0$. The assumption $\Delta(\gamma(\sigma)) \neq 0$ and Lemma 2.1 show that this cannot happen.

DEFINITION. A curve Γ, homeomorphic to (0, 1), will be said to properly separate D if $D - \Gamma$ is the union of two non-empty, open (in D), disjoint, sets H_1 and H_2.

The following are some consequences of this definition. Since H_1 and H_2 are both open and closed relative to $D - \Gamma$, the set $D - \Gamma$ admits the partition $H_1 \mid H_2$ and is disconnected. Now *arcwise connected* implies *connected*, hence *not connected* implies *not arcwise connected*. Thus *if* Γ *properly separates* D, $D - \Gamma$ *is not arcwise connected.*

We next show that Γ *has no limit points in* D. Indeed, suppose p is a limit point of Γ and $p \in D$. Then $p \in H_1$, or $p \in H_2$; assume $p \in H_1$. Since H_1 is open there is an ϵ-neighborhood of p contained in H_1. Then this neighborhood contains points of Γ, a contradiction.

An immediate result of H_1 and H_2 being closed in $D - \Gamma$ is that $\bar{H}_1 \subset H_1 \cup \Gamma$, $\bar{H}_2 \subset H_2 \cup \Gamma$, where $\bar{H}_i$ denotes the closure of H_i in D. This means that *the frontier of* H_1 *and* H_2 *is contained in* Γ.

One last result, needed in Lemma 2.3, is that *every point of* Γ *is a frontier point of both* H_1 *and* H_2. This will allow us to speak of the side of Γ in the direction of H_1. The proof of this is given in the appendix.

Define

$$\Gamma(x^0) \equiv \{\varphi^1(t, x^0) : t \in T(1, x^0)\} \cup \{\varphi^{-1}(t, x^0) : t \in T(-1, x^0), t > 0\},$$

$$\Gamma(x^f) \equiv \{\varphi^1(-t, x^f) : t \in T(1, x^f)\} \cup \{\varphi^{-1}(-t, x^f) : t \in T(-1, x^f), t > 0\}.$$

LEMMA 2.3. *Assume* $\Gamma(x^0)$ *properly separates* D, *forming the partition* $H_1 \mid H_2$. *Let* H_1 *be the side in the direction of* $\xi(0, x^0)$. *Then if* $\Delta \neq 0$ *along* $\Gamma(x^0)$ *and* $y \in H_2$, $y \notin R(x^0)$.

The statement also holds if x^0 *is replaced throughout by* x^f *while* $\xi(0, x^0)$ *is replaced by* $-\xi(0, x^f)$.

Proof. The proof will be given for the statement concerning $\Gamma(x^0)$, $\xi(0, x^0)$, that for $\Gamma(x^f)$, $-\xi(0, x^f)$ following similarly.

Assume $y \in R(x^0)$ and $y \in H_2$. Then there exists a $u \in U$ such that $\varphi^u(0, x^0) = x^0$, $\varphi^u(t_1, x^0) = y$ and $\varphi^u(t, x) \in D$ for $0 \leqq t \leqq t_1$. Since H_2 is open let $N(y)$ be a neighborhood of y contained in H_2. Noting that D is open, it follows from Theorem 1.2 that there exists a continuous control function which will have an associated trajectory arc which initiates at x^0 for $t = 0$, terminates within $N(y)$ for $t = t_1$, while the points of the trajectory arc all lie in D for values of t between 0 and t_1. This shows that there is no loss in generality in assuming that u is continuous, which we now do.

Since $\Gamma(x^0)$ properly separates D, $\Gamma(x^0)$ has no limit points in D and there must exist a point $p \in \Gamma(x^0)$ such that for some $t_2 \geqq 0$, $\varphi^u(t_2, x^0) = p$ while $\varphi^u(t, x^0) \in H_2$ for $t_2 < t \leqq t_1$. We first consider $t_2 > 0$. Assume without

loss of generality, that $p = \varphi^1(t_3, x^0)$. Since $\Delta(p) \neq 0$, there exists a neighborhood $N(p)$ of p, contained in D, such that

i) $\Delta(y) \neq 0$ for $y \in N(p)$.

ii) $|\theta(1, p) - \theta(1, y)| < \pi/4$ for $y \in N(p)$. (This is assured by the continuity of the right sides of (2.1).)

Let $t_4 > t_2$ and such that $\varphi^u(t_4, x^0) = p_4 \in N(p)$. Then $p_4 \in H_2$.

Since the right sides of (2.1) are continuously differentiable with respect to x, solutions are continuously differentiable with respect to initial conditions. Thus there exists an orientation preserving continuously differentiable homeomorphism which maps $N(p)$ onto a neighborhood of the origin of a rectangular coordinate system, such that $\varphi^1(\cdot, x^0)$ maps into the $\bar{x}_1$ axis of this system; trajectories of the form $\varphi^1(\cdot, y)$, $y \in N(p)$, map into lines parallel to the $\bar{x}_1$ axis; while the line perpendicular to $\varphi^1(\cdot, x^0)$ at p maps into the $\bar{x}_2$ axis. Let increasing t correspond to increasing $\bar{x}_1$, while points which were in $H_2 \cap N(p)$ map into points with positive $\bar{x}_2$ coordinate. (See Figure 1.)

Since it was assumed that $\xi(0, x^0)$ is in the direction of H_1 and $\Delta \neq 0$ along Γ, $\xi(0, p)$ is in the direction of H_1. Now any point $y \in N(p)$ can be joined to p by a continuous curve along which $\Delta \neq 0$. Thus the homeomorphism has been constructed so that the image of any point $\epsilon\xi(\alpha, y)$, $y \in N(p)$, $\epsilon > 0$, $|\alpha| \leqq 1$, will be a point of the form $(\bar{x}_1, \bar{x}_2)$ with $\bar{x}_2 \leqq 0$. This property characterizes allowable directions.

The image of the point q under the homeormophism will be denoted by $\bar{q}$. Thus p_4 maps into a point $\bar{p}_4$ having coordinates $(\bar{a}, \bar{b})$ with $\bar{b} > 0$, while $\{\varphi^u(t, x^0) : t_1 \leqq t \leqq t_4\}$ maps into an arc with a tangent defined at each point, joining the origin of the $\bar{x}_1$, $\bar{x}_2$ system to $\bar{p}_4$.

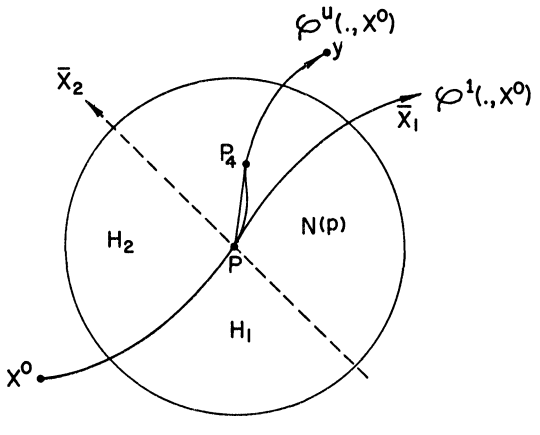

FIGURE. 1

Join the origin to $\bar{p}_4$ with the line segment $\ell(t) = t\bar{p}_4$ $0 \leqq t \leqq 1$. Then $\dot{\ell}(t)$ can be represented as the couple $(\bar{a}, \bar{b})$. By the mean value theorem, there exists a $t_5 \in (t_1, t_4)$ such that $\xi(u(t_5), \varphi^u(t_5, x^0)) = \varphi^u(t_5, x^0) = k(\bar{a}, \bar{b})$ where $k > 0$. But this is not an allowable direction, hence a contradiction which completes the proof for $t_2 > 0$.

If $t_2 = 0$, we consider a neighborhood $N(x^0)$ of x^0, and extend $\varphi^1(\cdot, x^0)$, $\varphi^{-1}(\cdot, x^0)$ for small negative values of t. This can be done since $\Delta(x^0) \neq 0$. Let φ^1 separate $N(x^0)$ into $H_3 \mid H_4$, while φ^{-1} gives rise to $H_5 \mid H_6$. Assume points which were in H_1 are in $H_4 \cap H_6$.

Again let $t_4 > 0$ be such that $p_4 = \varphi^u(t_4, x^0) \in N(x^0)$. If $p_4 \in H_3$ we can use the previous argument with H_4 replacing H_1 and H_3 replacing H_2 throughout. If $p_4 \in H_5 - (H_3 \cap H_5)$, we consider φ^{-1} as the separating trajectory and proceed as before.

In order to assure the existence of solutions of (2.1) joining x^0 to x^f, the following condition is imposed.

CONDITION 2.1. Either $\Gamma(x^0)$ or $\Gamma(x^f)$ properly separates D, and there exist $t_1, t_2, t_3, t_4 > 0$ such that

i) $\varphi^1(t_1, x^0) = \varphi^{-1}(-t_2, x^f)$.

ii) $\varphi^{-1}(t_1, x^0) = \varphi^1(-t_4, x^f)$.

iii) The trajectory arcs $\varphi^1(t, x^0)$, $0 \leqq t \leqq t_1$; $\varphi^{-1}(t, x^0)$, $0 \leqq t \leqq t_3$; $\varphi^1(-t, x^f)$, $0 \leqq t \leqq t_4$; $\varphi^{-1}(-t, x^f)$, $0 \leqq t \leqq t_2$, all lie in D.

iv) $\Delta(x) \neq 0$ for x in the set $\Gamma(x^0)$ or $\Gamma(x^f)$ which properly separates D.

A problem will be said to satisfy Condition 2.1 if its associated $\varphi^1(\cdot, x^0)$, $\varphi^{-1}(\cdot, x^0)$, $\varphi^1(\cdot, x^f)$ and $\varphi^{-1}(\cdot, x^f)$ satisfy the condition.

REMARKS.

1. If the values t_1, t_2, t_3 and t_4 exist, they are unique.

2. It is possible that $\varphi^1(\bar{t}, x^0) = x^f$, $[\varphi^{-1}(\bar{t}, x^0) = x^f]$ for some $\bar{t} \geqq 0$, and that the problem satisfies Condition 2.1 with $t_1, t_2, t_3, t_4 > 0$ replaced by $t_1, t_2, t_3, t_4 \geqq 0$. Then since $\Delta \neq 0$ along the arc of $\varphi^1(\cdot, x^0)$, $[\varphi^{-1}(\cdot, x^0)]$, connecting x^0 to x^f, it is the only allowable trajectory arc connecting these points. In this event the problem is trivial, so it will be omitted from further consideration.

3. This condition is sufficient for existence, but certainly not necessary.

DEFINITION. Assuming Condition 2.1, S will denote the compact simply connected subset of D bounded by the arcs of $\varphi^1(\cdot, x^0)$, $\varphi^{-1}(\cdot, x^0)$, $\varphi^1(\cdot, x^f)$ and $\varphi^{-1}(\cdot, x^f)$.

The following theorem now gives the desired characterization of $R(x^0) \cap R(x^f)$, in terms of the trajectories thru x^0 and x^f with controls 1 and -1.

THEOREM 2.1. *If a problem satisfies Condition* 2.1, *and* $\Delta(y) \neq 0$ *for* $y \in S$, *then* $S = R(x^0) \cap R(x^f)$.

Proof. a) Assume $y \in S$. It will be shown that $y \in R(x^0) \cap R(x^f)$. It suffices to consider $y \in S$ interior, for if y belongs to the boundary of S, the

result is immediate. It must be shown that there exists a control $u \in U$ and $t_1 \geqq 0$ such that $\varphi^u(t_1, x^0) = y$; and a control $\tilde{u} \in U$ and $t_2 \geqq 0$ such that $\varphi^{\tilde{u}}(-t_2, x^f) = y$.

Now S is a simply connected, compact subset of D with no critical points. It follows that every semiorbit which initiates within S must leave this set in finite time. Thus the semiorbit $\varphi^1(-t, y)$, $t \geqq 0$, must intersect the boundary of S, and do so along the arc of $\varphi^{-1}(t, x^0)$ which contributes to the boundary. (This is easily verified by observing that $\varphi^1(-t, y)$ cannot intersect $\varphi^1(t, x^0)$ or $\varphi^1(-t, x^f)$ by the uniqueness property, while if it intersects $\varphi^{-1}(-t, x^f)$ a contradiction to Lemma 2.2 occurs.) Assume that the intersection occurs at the point $\bar{x}$ of the boundary of S, and that $\varphi^1(-t_3, y) = \bar{x}$ while $\varphi^{-1}(t_4, x^0) = \bar{x}$. Then

$$u(t) = \begin{cases} -1 \text{ for } 0 \leqq t \leqq t_4, \\ 1 \text{ for } t_4 < t \leqq t_3 + t_4 \end{cases}$$

is an allowable control such that $\varphi^u(t_3 + t_4, x^0) = y$.

Similarly construct $\varphi^1(t, y)$, $t \geqq 0$. This must intersect $\varphi^{-1}(-t, x^f)$ and the control $\tilde{u}$ is easily constructed.

Thus $y \in S \rightarrow y \in R(x^0) \cap R(x^f)$.

b) Assume $y \in R(x^0) \cap R(x^f)$ but $y \notin S$. A contradiction will be shown. We consider the case where $\Gamma(x^0)$ properly separates D, the proof being similar with $\Gamma(x^f)$.

Let $\Gamma(x^0)$ give rise to the partition $H_1 \mid H_2$, and assume, without loss of generality, that the interior of S is in H_1. Let t_1, t_2, t_3 and t_4 be as in Condition 2.1. Then either

i) $y \in H_2$,

ii) $y \in H_1 - S$,

iii) $y = \varphi^1(t, x^0)$ for some $t \in T_D(1, x^0)$, $t > t_1$, or
$y = \varphi^{-1}(t, x^0)$ for some $t \in T_D(-1, x^0)$, $t \geqq t_3$.

We consider these cases in order.

By Lemma 2.3, $y \in R(x^0) \cap R(x^f) \Rightarrow y \notin H_2$.

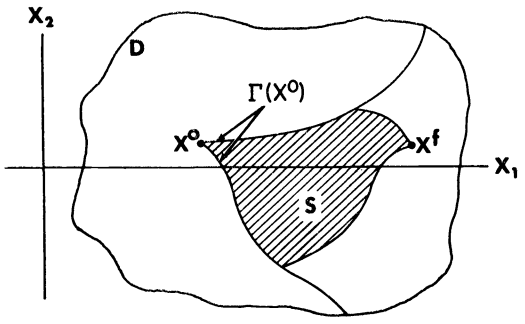

FIGURE 2.

Next consider H_1 as our new domain. Then

$$\Gamma(x^f) \equiv \{\varphi^1(-1, x^f) : 0 \leqq t < t_4\} \cup \{\varphi^{-1}(-t, x^f) : 0 < t < t_2\}$$

properly separates H_1, forming a partition $H_3 \mid H_4$. Since $\Delta(y) \neq 0$ for $y \in S$, $\Delta(x) \neq 0$ for $x \in \Gamma(x^f)$. Assume, without loss of generality that the interior of S is in H_3. Then by Lemma 2.3, $y \in R(x^0) \cap R(x^f) \Rightarrow y \notin H_4 = H_1 - S$.

This leaves only iii) above, which is quickly ruled out since the condition $\Delta(x) \neq 0$ for $x \in \Gamma(x^0)$ implies that the intersections of $\varphi^1(\cdot, x^0)$, $\varphi^{-1}(\cdot, x^0)$, $\varphi^1(\cdot, x^f)$ and $\varphi^{-1}(\cdot, x^f)$ are unique, and cannot reoccur for values of $t > t_1$ or $t > t_3$.

Thus $y \in R(x^0) \cap R(x^f) \Rightarrow y \in S$, which completes the proof.

3. SYNTHESIS OF THE OPTIMAL CONTROL FOR THE TWO-DIMENSIONAL PROBLEM

3.1. Synthesis by Green's theorem. Let $\Delta(x) \neq 0$ for $x \in R(x^0) \cap R(x^f) = S$, (assuming S is defined). Then for a given trajectory arc φ^u connecting two points P_0 and $P_1 \in S$, the functional C (see Section 2.1) can be expressed as a line integral along this arc.

$$(3.1) \qquad C(u, P_0, P_1) = \int_{t_{P_0}}^{t_{P_1}} L\, dt = \int_{P_0}^{P_1} -\frac{LB_2}{\Delta}\, dx_1 + \frac{LB_1}{\Delta}\, dx_2 .$$

Suppose φ^{u_1} and φ^{u_2} are two different solutions to (2.1), each joining P_0 to P_1 in S, and having no points other than P_0 and P_1 in common. Let Γ be the closed curve formed by these trajectory arcs. If we traverse Γ in a counterclockwise fashion by following first the arc of φ^{u_1} from P_0 to P_1 and next the arc of φ^{u_2} from P_1 to P_0, then

$$(3.2) \qquad C(u_1, P_0, P_1) - C(u_2, P_0, P_1) = \oint_\Gamma -\frac{LB_2\, dx_1}{\Delta} + \frac{LB_1\, dx_2}{\Delta} .$$

Since the bounding curve Γ is a Jordan Curve, applying Green's theorem to (3.2), which is permissible by virtue of $A_1, B_1, A_2, B_2, L \in C^1(S)$, the class of once continuously differentiable functions, the following result is obtained:

$$(3.3) \qquad C(u_1, P_0, P_1) - C(u_2, P_0, P_1) = \iint_{\mathfrak{R}} w(x)\, dS$$

where

$$(3.4) \qquad w(x) = \frac{\partial}{\partial x_1}\left(\frac{LB_1}{\Delta}\right) + \frac{\partial}{\partial x_2}\left(\frac{LB_2}{\Delta}\right)$$

and $\mathfrak{R}$ is the region enclosed by Γ.

Since $w(x)$ is uniquely determined for all $x \in S$, (3.3) provides a direct means for determining the optimal strategy.

3.2. Relationship between $\omega = 0$ and the singular problem. The control problem (2.1) is said to be singular when the coefficient of u (see (3.5)) appearing in the Hamiltonian function vanishes over some time interval of positive measure. In this case the principle of the maximum, or its classical counterpart condition II of Weierstrass, fails to yield any information regarding the optimal control. The vanishing of the coefficient of u can be used to determine the intermediate nature of the singular control [7]; however, there are no criteria available that determine when intermediate control should be abandoned in favor of a bang-bang control and conversely. The control problem (2.1) is singular when

$$p_1(t)B_1(\varphi^u(t)) + p_2(t)B_2(\varphi^u(t)) = 0 \tag{3.5}$$

for some subset of positive measure of the interval $t_0 \leqq t \leqq t_f$, where p is the costate vector whose elements are determined by the differential equations

$$\begin{aligned} \dot{p}_1 &= \frac{\partial L}{\partial x_1} - p_1\left(\frac{\partial A_1}{\partial x_1} + \frac{\partial B_1}{\partial x_1}\,u\right) - p_2\left(\frac{\partial A_2}{\partial x_1} + \frac{\partial B_2}{\partial x_1}\,u\right) \\ \dot{p}_2 &= \frac{\partial L}{\partial x_2} - p_1\left(\frac{\partial A_1}{\partial x_2} + \frac{\partial B_1}{\partial x_2}\,u\right) - p_2\left(\frac{\partial A_2}{\partial x_2} + \frac{\partial B_2}{\partial x_2}\,u\right). \end{aligned} \tag{3.6}$$

(For convenience $\varphi^u(t, x^0)$ is denoted $\varphi^u(t)$.)

From [2] the Hamiltonian is a constant, and the constant is zero so that

$$\begin{aligned} L(\varphi(t)) - p_1(t)A_1(\varphi^u(t)) - p_2(t)A_2(\varphi^u(t)) \\ - [p_1(t)B_1(\varphi^u(t)) + p_2(t)B_2(\varphi^u(t))]u(t) \equiv 0. \end{aligned} \tag{3.7}$$

This result is equivalent to the first integral of the calculus of variations [1], with the constant of integration zero, by virtue of the transverality condition and the boundary conditions.

Then from (3.5)

$$\begin{aligned} \frac{d}{dt}[p_1(t)B_1(\phi^u(t)) + p_2(t)B_2(\phi^u(t))] &= \dot{p}_1(t)B_1(\phi^u(t)) + \dot{p}_2(t)B_2(\varphi^u(t)) \\ &+ \left[p_1(t)\,\frac{\partial B_1}{\partial x_1}\,(\varphi^u(t)) + p_2(t)\,\frac{\partial B_2}{\partial x_1}\,(\varphi^u(t))\right]\dot{\varphi}_1{}^u(t) \\ &+ \left[p_1(t)\,\frac{\partial B_1}{\partial x_2}\,(\varphi^u(t)) + p_2(t)\,\frac{\partial B_2}{\partial x_2}\,(\varphi^u(t))\right]\dot{\varphi}_2{}^u \equiv 0 \end{aligned}$$

using (2.1), (3.5), (3.6), (3.7) and recalling the definitions of Δ, (2.2),

and ω, (3.4), this can be reduced to

$$(3.8)\quad \frac{d}{dt}\{p_1(t)B_1(\varphi^u(t)) + p_2(t)B_2(\varphi^u(t)\} \equiv \Delta(\varphi^u(t))\omega(\varphi^u(t)) \equiv 0.$$

Since $\Delta(x) \neq 0$ for all $x \in S$, (3.8) establishes a relationship between the singular control problem and $\omega(\varphi^u(t)) \equiv 0$. Thus, on arcs for which the optimal control need not be bang-bang, $\omega \equiv 0$.

Several comments can be made concerning the reverse implication, i.e., does $\omega(\varphi^u(t)) \equiv 0$ imply $p_1(t)B_1(\varphi^u(t)) + p_2(t)B_2(\varphi^u(t)) \equiv 0$? Again we assume $\Delta \neq 0$.

a) If $\omega \equiv 0$ along an optimal path obtained with an intermediate control ($|u(t)| < 1$) then $p_1B_1 + p_2B_2 \equiv 0$ along this path, since the necessary condition (the maximum principle) would otherwise require that $|u(t)| = 1$.

b) The condition $\omega(\varphi^u(t)) \equiv 0$ implies

$$\frac{d}{dt}(p_1 B_1 + p_2 B_2) + f(t)(p_1 B_1 + p_2 B_2) \equiv 0,$$

where

$$f(t) = -\frac{1}{\Delta}\left\{\dot{\Delta} - \Delta\left[\frac{\partial}{\partial x_2} A_2 + \left(\frac{\partial}{\partial x_2} B_2\right) u + \frac{\partial}{\partial x_1} A_1 + \left(\frac{\partial}{\partial x_1} B_1\right) u\right]\right\}.$$

Thus if $p_1B_1 + p_2B_2$ is zero at any point of the arc, it remains zero.

c) If $\omega = 0$ is an arc which intersects an optimal trajectory and is allowable (i.e., there exists a $u \in U$ such that the solution of (2.1) follows this arc) in some neighborhood of the point of intersection, then the problem is singular, i.e., p_1 and p_2 can be chosen initially so as to assure $p_1B_1 + p_2B_2 = 0$ at the point of intersection. In view of comment b), $p_1B_1 + p_2B_2$ remains zero.

3.3. Uniqueness of the intermediate control. The synthesis procedure will be to determine the "optimal path" in the phase space, and then determine, if possible, a control from U, which leads to a solution of (2.1) with a trajectory arc which coincides with the "optimal path".

Theorem 3.1, which follows, will show that if one can determine a *piecewise continuous* control from U, such that the corresponding solution arc of (2.1) coincides with the "optimal path", then this control is unique (up to a set of measure zero). It will be found that to obtain a constructive method to determine this control, one needs L, A_1, B_1, A_2, B_2 to be at least twice continuously differentiable.

An example (3-7) will be given of a time optimal problem of the form (2.1) which leads to an "optimal path" which is obtained as a trajectory arc of a solution of (2.1) with a continuous control u such that $|u(t)| < 1$

for all t. Theorem 3.1 then shows that there can be no bang-bang control which leads to a trajectory coinciding with the optimal path, thus the remark following Theorem 1.1, that $S^0(t)$ need not equal $S(t)$, beomes evident.

THEOREM 3.1. *If $\varphi^u(t)$ and $\varphi^{\tilde{u}}(s)$ are two solutions of the system* (2.1) *joining two points a and b of a simply connected subset $D \subset E^2$, such that*

i) $u \in C[0, t_1]$[1], $\quad t_1 > 0$,
ii) $\varphi^u(0) = a, \quad \varphi^{\tilde{u}}(0) = a$,
iii) $\varphi^u(t_1) = b, \quad \varphi^u(s_1) = b$,
iv) $\{\varphi^u(t) : 0 \leqq t \leqq t_1\} = \{\varphi^{\tilde{u}}(s) : 0 \leqq s \leqq s_1\} \equiv \Lambda \subset D$
v) $\Delta(y) \not\equiv 0$ for $y \in D$,

then $t_1 = s_1$ and $\tilde{u} \doteq u$, where we shall use $\doteq$ to denote equal except on a set of measure zero.

Proof. The restriction v), implies that the trajectory φ^u does not cross itself, i.e., there are no values $\bar{t}$ and $\hat{t}$ such that $0 \leqq \bar{t}, \hat{t} \leqq t_1$, $\bar{t} \neq \hat{t}$, and $\varphi^u(\bar{t}) = \varphi^u(\hat{t})$. Indeed, if this were the case, $\Delta \not\equiv 0$ in the loop created, hence constructing $\varphi^1(t, y)$ for y interior to this loop leads to a contradiction of Lemma 2.2.

Then in view of condition iv), there exists a continuous, monotone function f defined on the interval $[0, s_1]$ by

$$\varphi^u(f(s)) \equiv \varphi^{\tilde{u}}(s) \tag{3.9}$$

such that $f(0) = 0$ and $f(s_1) = t_1$.

It will next be shown that f is absolutely continuous. To show this, it should be noted that there exist $m, \delta > 0$ such that

$$m \mid t - \bar{t} \mid \leqq \mid \varphi_1{}^u(t) - \varphi_1{}^u(\bar{t}) \mid + \mid \varphi_2{}^u(t) - \varphi_2{}^u(\bar{t}) \mid \tag{3.10}$$

for $[t - \bar{t}] < \delta$; $t, \bar{t} \in [0, t_1]$.

Indeed, assume that this statement is false. Let $m = \frac{1}{2} \min_{0 \leqq t \leqq t_1} [\mid \dot{\varphi}_1{}^u(t) \mid + \mid \dot{\varphi}_2{}^u(t) \mid]$, and let $\delta_\nu \to 0$, $\delta_\nu > 0$. Then for each ν there exists $t_\nu, \bar{t}_\nu \in [0, t_1]$ such that $\mid \bar{t}_\nu - t_\nu \mid < \delta_\nu$ and

$$m \mid t_\nu - \bar{t}_\nu \mid > \mid \varphi_1{}^u(t_\nu) - \varphi_1{}^u(\bar{t}_\nu) \mid + \mid \varphi_2{}^u(t_\nu) - \varphi_2{}^u(\bar{t}_\nu) \mid.$$

Because of the compactness of $[0, t_1]$, there exist convergent subsequences $\{t_{\nu_k}\}$ and $\{\bar{t}_{\nu_k}\}$ of $\{t_\nu\}, \{\bar{t}_\nu\}$; and they must have the same limit, say t_0. But then

$$\mid \dot{\varphi}_1{}^u(t_0) \mid + \mid \dot{\varphi}_2{}^u(t_0) \mid$$

$$= \lim_{k \to \infty} \frac{\mid \varphi_1{}^u(t_{\nu_k}) - \varphi_1{}^u(\bar{t}_{\nu_k}) \mid + \mid \varphi_2{}^u(t_{\nu_k}) - \varphi_2{}^u(\bar{t}_{\nu_k}) \mid}{\mid t_{\nu_k} - \bar{t}_{\nu_k} \mid} \leqq m$$

and this contradicts the definition of m.

[1] If the optimal control is defined over $[0, t_f]$, the interval $[0, t_1]$ can be assumed to to represent an arbitrary subinterval over which u is continuous.

Assume now that $f(s)$ is not absolutely continuous on $[0, s_1]$. A contradiction will be shown.

f not absolutely continuous implies there exists an $\epsilon > 0$ and a sequence of mutually separate intervals $[\alpha_k(\nu), \beta_k(\nu)]$, $k = 1, 2, \cdots, n(\nu) < \infty$, such that

$$\text{(3.11)} \quad \begin{aligned} &\text{i) each interval is contained in } [0, s_1], \\ &\text{ii) } \sum_{k=1}^{n(\nu)} |\alpha_k(\nu) - \beta_k(\nu)| \to 0 \text{ as } \nu \to \infty, \\ &\text{iii) } \sum_{k=1}^{n(\nu)} |f(\alpha_k(\nu)) - f(\beta_k(\nu))| > \epsilon \text{ for all } \nu. \end{aligned}$$

Choose $\bar{\nu}$ large enough so that for each k and $\nu \geqq \bar{\nu}$,

$$|\alpha_k(\nu) - \beta_k(\nu)| < \delta.$$

From (3.10) one obtains for every k and $\nu \geqq \bar{\nu}$

$$\begin{aligned} m\,|f(\alpha_k(\nu)) - f(\beta_k(\nu))| &\leqq |\varphi_1^u(f(\alpha_k(\nu)) - \varphi_1^u(f(\beta_k(\nu)))| \\ &\quad + |\varphi_2^u(f(\alpha_k(\nu))) - \varphi_2^u(f(\beta_k(\nu)))| \equiv |\varphi_1^{\bar{u}}(\alpha_k(\nu)) - \varphi_1^{\bar{u}}(\beta_k(\nu))| \\ &\quad + |\varphi_2^{\bar{u}}(\alpha_k(\nu)) - \varphi_2^{\bar{u}}(\beta_k(\nu))|. \end{aligned}$$

Then from (3.11)

$$\begin{aligned} \epsilon &< \sum_{k=1}^{n(\nu)} |f(\alpha_k(\nu)) - f(\beta_k(\nu))| \\ &\leqq \frac{1}{m} \sum_{k=1}^{n(\nu)} \{|\varphi_1^{\bar{u}}(\alpha_k(\nu)) - \varphi_1^{\bar{u}}(\beta_k(\nu))| + |\varphi_2^{\bar{u}}(\alpha_k(\nu)) - \varphi^{\bar{u}}(\beta_k(\nu))|\} \end{aligned}$$

for all $\nu \geqq \bar{\nu}$. In view of (3.11), this is a contradiction to the absolute continuity of $\varphi^{\bar{u}}$ on the interval $[0, s_1]$. This contradiction shows that f is absolutely continuous.

By assumption

$$\text{(3.12)} \qquad \dot{\varphi}_i^u(t) = A_i(\varphi^u(t)) + B_i(\varphi^u(t))u(t), \qquad i = 1, 2.$$

Also, since $\varphi^u \in C^1[0, t_1]$ and f is absolutely continuous,

$$\text{(3.13)} \qquad \frac{d}{ds}\varphi^u(f(s)) \doteq \dot{\varphi}^u(f(s))f'(s).$$

Now, in (3.12), substituting $f(s)$ for t and multiplying both sides of the identities by $f'(s)$ yields

$$\dot{\varphi}_i^u(f(s)f'(s) \doteq [A_i(\varphi^u(f(s))) + B_i(\varphi^u(f(s)))u(f(s))]f'(s), \quad i = 1, 2,$$

or in view of (3.9) and (3.13)

$$\dot{\varphi}_i{}^{\tilde{u}}(s) \doteq [A_i(\varphi^{\tilde{u}}(s)) + B_i(\varphi^{\tilde{u}}(s))u(f(s))]f'(s), \qquad i = 1, 2, \tag{3.14}$$

But

$$\dot{\varphi}_i{}^{\tilde{u}}(s) \doteq A_i(\varphi^{\tilde{u}}(s)) + B_i(\varphi^{\tilde{u}}(s))\tilde{u}(s), \qquad i = 1, 2. \tag{3.15}$$

(It is only required that $\tilde{u} \in U$.)

Equating the right sides of (3.14), (3.15) yields

$$\begin{pmatrix} A_1(\varphi^{\tilde{u}}(s)) & B_1(\varphi^{\tilde{u}}(s)) \\ A_2(\varphi^{\tilde{u}}(s)) & B_2(\varphi^{\tilde{u}}(s)) \end{pmatrix} \begin{pmatrix} f'(s) \\ f'(s)u(f(s)) \end{pmatrix} \doteq \begin{pmatrix} A_1(\varphi^{\tilde{u}}(s)) & B_1(\varphi^{u}(s)) \\ A_2(\varphi^{\tilde{u}}(s)) & B_2(\varphi^{\tilde{u}}(s)) \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{u}(s) \end{pmatrix}$$

which implies $f'(s) \doteq 1$, $u(f(s)) \doteq \tilde{u}(s)$. But f is absolutely continuous. $f(0) = 0$, hence $s_1 = t_1$ and $u(s) \doteq \tilde{u}(s)$.

3.4. Optimal strategy. The advantage of the Green's theorem approach is that the relative optimality of two distinct paths can be obtained directly. It may be observed from (3.3) that if

$$\omega \gtrless 0 \text{ in } \mathfrak{R}, \text{ then } C(u_1, P_0, P_1) - C(u_2, P_0, P_1) \gtrless 0.$$

The optimal strategy for the control problem considered is in some cases quite obvious, and be simply determined by inspection of the behavior of $\omega(x)$ in $R(x^0) \cap R(x^f)$. However, there are instances where this procedure fails, and presumably a definite integration of (3.3) is required for the determination of the optimal strategy. The purpose of this section is to list only those cases where the optimal strategy is deducible directly from the behavior of $\omega(x)$ for all $x \in R(x^0) \cap R(x^f)$.With regard to the behavior of $\omega(x)$ the following subsets are defined.

$$\begin{aligned} \omega_+ &= \{x \in R(x^0) \cap R(x^f) : \omega(x) > 0\} \\ \omega_- &= \{x \in R(x^0) \cap R(x^f) : w(x) < 0\} \\ \omega_0 &= \{x \in R(x^0) \cap R(x^f) : \omega(x) = 0\} \end{aligned} \tag{3.17}$$

CASE 1. $\omega_0 \equiv R(x^0) \cap R(x^f)$.

If $\omega_0 = R(x^0) \cap R(x^f)$, and the interior of $R(x^0) \cap R(x^f)$ is assumed to be non-empty, then the control problem is degenerate. The value of the cost functional is by (3.3) independent of the path taken in $R(x^0) \cap R(x^f)$, and hence independent of control.

EXAMPLE 3-1. Given

$$\begin{aligned} \dot{x}_1 &= x_1 + x_2 u, \\ \dot{x}_2 &= x_2 + x_1 u \end{aligned} \tag{3.18}$$

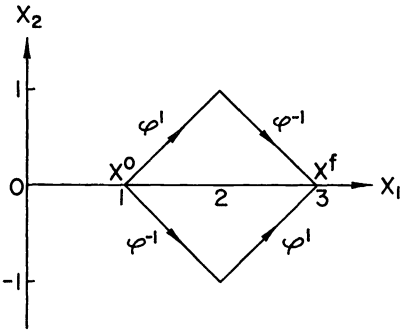

FIGURE 3.

with $| u | \leqq 1$. Let $x^0 = (1, 0)$, $x^f = (3, 0)$ so that S is bounded, and the boundary of S is illustrated on Figure 3. Consider a time optimal problem so that $L \equiv 1$; then $\omega(x) \equiv 0$ for all $x \in S$. An integral of (3.18) with initial condition x^0 when $t = 0$ is

$$x_1^2 - x_2^2 = e^{2t}. \tag{3.19}$$

so that the transit time is independent of control. Since $\Delta(x) = 0$ on $x_2 = \pm x_1$, which do not intersect S, and all motions from x^0 lie in the quarter plane of points with positive first coordinate and bounded by the lines $x_1 = \pm x_2$, Theorem 2.1 applies with the trajectories properly separating the quarter plane. Thus $R(x^0) \cap R(x^f) \equiv S$, so that there exists more than one control $u \in U$ for which the solution to (3.18) satisfies the boundary conditions. Hence, in view of this fact and (3.19) there is no unique time optimum to this problem.

CASE 2. Either $\omega_+ = \varnothing$ or $\omega_- = \varnothing$ and ω_0 is the union of a finite number of arcs and points in $R(x^0) \cap R(x^f)$.

If $\omega(x)$ has a constant sign except on a finite number of arcs and points were $\omega(x) = 0$, in $R(x^0) \cap R(x^f)$, then the optimum control path lies in the boundary of $R(x^0) \cap R(x^f)$. This follows directly from (3.3) when comparing any interior path to a boundary path of $R(x^0) \cap R(x^f)$. The appropriate branch of the boundary that contains the optimum control path is determined by the sign of $\omega(x)$.

EXAMPLE 3-2. As an example where $\omega(x)$ was a constant sign everywhere in $R(x^0) \cap R(x^f)$, consider (3.18) and the boundary conditions of Example 3-1, with $L = x_1 + x_2$. Then $\omega = 1/(x_2 - x_1)$ and $\omega_- \equiv R(x^0) \cap R(x^f)$, so that the optimal control path is the lower boundary of $R(x^0) \cap R(x^f)$ shown in Figure 3.

EXAMPLE 3-3. For the case where ω_0 has a finite number of elements, let $L = -4x_2 + x_1x_2 + 4 \log (x_1 + x_2)$ together with (3.18) and the

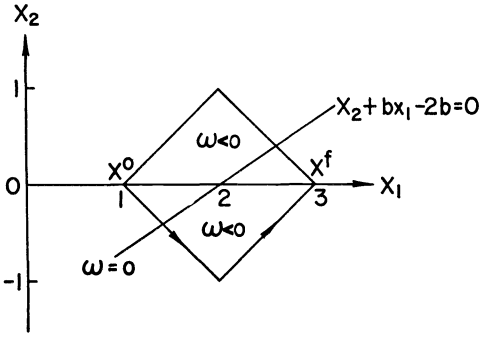

FIGURE 4.

boundary conditions of Example 3-1. Then

$$\omega = \frac{-x_2^2 - (x_1 - 2)^2}{x_1^2 - x_2^2}$$

and $\omega_+ = \varnothing$, and ω_0 has a single element (2, 0). As in Example 3-2 the optimum control path is the lower boundary of $R(x^0) \cap R(x^f)$.

EXAMPLE 3-4. For the case where ω_0 is no more than a finite number of curves in $R(x^0) \cap R(x^f)$, let

$$L = 4bx_1 + 4b^2 x_2 - bx_1^2 - \frac{(1 + b^2)}{2} x_1 x_2 - 4b^2 \log (x_1 + x_2) - \frac{(b^2 - 1)}{2} (x_1^2 - x_2^2) \log (x_1 + x_2)$$

where b is some constant, while (3.18) and the boundary conditions of Example 3-1 apply. Then

$$\omega = \frac{-\{x_2 + bx_1 - 2b\}^2}{x_1^2 - x_2^2}$$

so that $\omega_+ = \varnothing$, and ω_0 is the intersection of $R(x^0) \cap R(x^f)$ as shown in Figure 4 and is identical with the optimum control paths for Example 3-2 and 3-3. At this stage there is an important point to be stressed. As shown in section 3.2, if there is an allowable arc along which $\omega \equiv 0$, the problem may be singular. If in Example 3-4 the constant b were chosen in the range

$$-1 \leqq b \leqq 1$$

then there is a control $u \in U$ for which the solution to (3.18) satisfies (3.20), so that an intermediate control is not ruled out by minimizing the

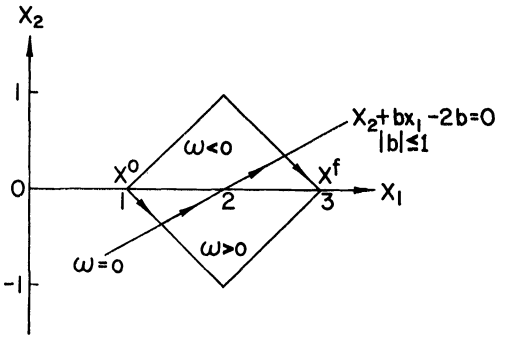

FIGURE 5.

Hamiltonian.[2] However, by using the Green's theorem approach this possibility has been ruled out, and it has been shown that the optimum control is bang-bang.

CASE 3. ω_+ and ω_- are both non-empty and ω_0 is at most a finite number of curves in $R(x^0) \cap R(x^f)$.

In this case $\omega(x)$ is not of constant sign in $R(x^0) \cap R(x^f)$. It is only in some instances that it is possible to find the optimal strategy directly. No purpose is served by enumerating the conditions when this is possible, since given a specific example it becomes a relatively simple task to see if this is the case.

EXAMPLE 3-5. Consider (3.18) and the boundary conditions of Example 3-1. Let

$$L = x_1 + bx_2 - 2b \log (x_1 + x_2)$$

where b is some constant, then

$$\omega = \frac{-x_2 - bx_1 + 2b}{x_1^2 - x_2^2} .$$

If $| b | \leqq 1$ then there is a solution to (3.18) with $u \in U$ that satisfies $\omega = 0$, i.e.,

$$x_2 + bx_1 - 2b = 0$$

over some interval of positive length. In this problem the intermediate control is part of the optimal strategy, as illustrated on Figure 5. If $b = 1$, then $u = 1$ for the arc of the solution trajectory coinciding with $\omega = 0$, which illustrates the point that situations can arise where the control problem is singular but is still bang-bang. (See comment c given in section 3.2.) If $| b | > 1$, there is no $u \in U$ which yields an arc of a solution tra-

[2] See comments given in section 3.2.

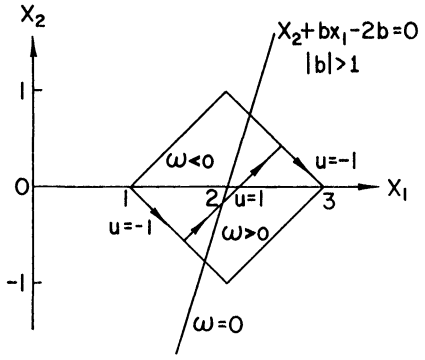

FIGURE 6.

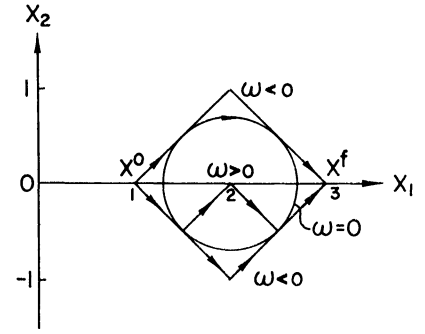

FIGURE 7.

jectory coinciding with $\omega = 0$, and the optimum control is bang-bang. The optimal control path has three sub-arcs, but to obtain the exact location of the second sub-arc (Figure 6) presumably requires integration of (3.3).

EXAMPLE 3-6. Consider (3.18) and the boundary conditions of Example 3-1 with

$$L = 4x_2 + x_1x_2 + \tfrac{7}{2}\log(x_1 + x_2)$$

so that

$$\omega = \frac{\frac{1}{2} - x_2^2 - (x_1 - 2)^2}{x_1^2 - x_2^2}.$$

On the upper and lower quadrants of the circle shown on Figure 7, there is some $u \in U$ for which an arc of the solution trajectory of (3.18) satisfies $\omega = 0$, that is

$$x_2^2 + (x_1 - 2)^2 = \tfrac{1}{2}.$$

On the other two quadrants there is no $u \in U$ for which the solution to (3.18)

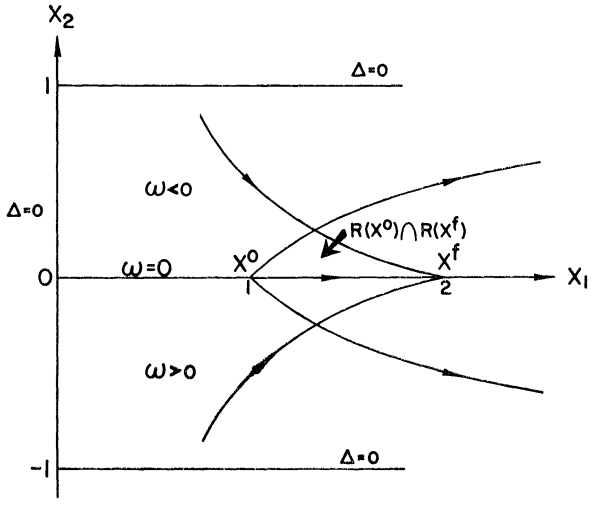

FIGURE 8.

satisfies $\omega = 0$. In this case the optimal strategy cannot be resolved directly. Paths that give possible relative minima are sketched on Figure 7.

EXAMPLE 3-7. This will be an example of a time optimal problem that is singular, and which illustrates, in view of Theorem 3.1, that it is not possible to do as well with a bang-bang control as with an intermediate control.

Consider

$$\begin{aligned} \dot{x}_1 &= x_1^2 - x_1^2 x_2 u, \\ \dot{x}_2 &= -x_2 + u, \quad |u| \leqq 1 \end{aligned} \tag{3.21}$$

with boundary conditions $x^0 = [1, 0]$, $x^f = [2, 0]$. The solutions to (3.20) with control $u \equiv 1$ and $u \equiv -1$ are

$$\varphi^1(t, x^0) = \begin{cases} e^t \\ 1 - e^{-t}, \end{cases} \quad \varphi^{-1}(t, x^0) = \begin{cases} e^t \\ e^{-t} - 1 \end{cases}$$

$$\varphi^1(-t, x^f) = \begin{cases} \dfrac{1}{-\frac{1}{2} + e^t}, \\ -e^t + 1 \end{cases} \quad \varphi^{-1}(-t, x^f) = \begin{cases} \dfrac{1}{-\frac{1}{2} + e^t} \\ e^t - 1 \end{cases}$$

and t_1, t_2, t_3 and t_4 exist. $R(x^0) \cap R(x^f)$ is illustrated on Figure 8, and $\Delta = x_1^2[x_2^2 - 1] \neq 0$ for $x \in R(x^0) \cap R(x^f)$. For this problem

$$\omega(x) = \frac{2x_2}{x_1^2 [1 - x_2^2]^2}.$$

Since $\omega(x) = 0$ when $x_2 = 0$, the optimal control is $u \equiv 0$.

REFERENCES

[1] G. A. Bliss, *Lectures on the Calculus of Variations*, The University of Chicago Press, Chicago, 1945.

[2] V. G. Boltyanskii, R. V. Gamkrelidze and L. S. Pontryagin, *Optimal control processes*, Uspelhi Mat. Nauk., 14 (1959), No. 1 (85), pp. 3–20.

[3] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill Book Co., Inc., New York, 1955.

[4] A. F. Fillipov, *On certain questions in the theory of optimal control*, J. Soc. Indust. Appl. Math. Ser. A: On Control, 1 (1962), no. 1, pp. 76–84. (English Translation.)

[5] R. E. Kalman, *The theory of optimal control and the calculus of variations*, RIAS Technical Report 6-13, Research Institute for Advanced Studies (RIAS), Baltimore, 1961.

[6] 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.

[7] D. F. Lawden, *Optimal powered arcs in an inverse square law field*, J. Amer. Rocket. Soc., 31 (1961), pp. 566–568.

[8] E. B. Lee, *Time optimal control of nonlinear processes*, Trans. A.S.M.E., Paper No. 61–JAC-3, Presented to J.A.C.C., Boulder, Colo. 1961.

[9] D. L. Lukes, *Steepest Descent and Maximum Principle Techniques of System Optimization*, Symposium on Vehicle Systems Optimization, Garden City, L. I., New York, 1961, pp. 36–41.

[10] L. Markus and E. B. Lee, *Optimal control for nonlinear processes*, Arch. Rational. Mech. Anal., 8 (1961), pp. 36–58.

[11] A. Miele, *Application of Green's Theorem to the Extremization of Linear Integrals*, Symposium on Vehicle Systems Optimization, Garden City, L. I., New York, 1961, pp. 26–35.

[12] I. P. Natanson *Theory of Functions of a Real Variable*, Vol. I, Fred. Ungar Pub. Co., New York, 1955.

[13] M. H. A. Newman, *Elements of the Topology of Plane Sets of Points*, Cambridge University Press, Princeton, 1954.

[14] E. Roxin, *A geometric interpretation of Pontryagin's maximum principle*, RIAS Technical Report 61-15, Research Institute for Advanced Studies (RIAS), Baltimore, 1961.

[15] E. Roxin, *On the existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

APPENDIX

We wish to show that if Γ properly separates D, every point of Γ is a frontier point of H_1 and H_2.

Every point of Γ is a frontier point of either H_1 or H_2, so assume Γ to be parametrized as $\Gamma(t)$, $t \in (0, 1)$ and $\Gamma(t_0) = p_0$ is not a frontier point of H_1. Thus there is a neighborhood $n(p_0)$ such that $n(p_0) - \Gamma \subset H_2$. We look for a contradiction.

Let y be any point in H_1. Join p_0 to y with a simple arc C in D, see Figure 9. Since points in $n(p_0) - \Gamma$ belong to H_2 while $y \in H_1$, there is a frontier

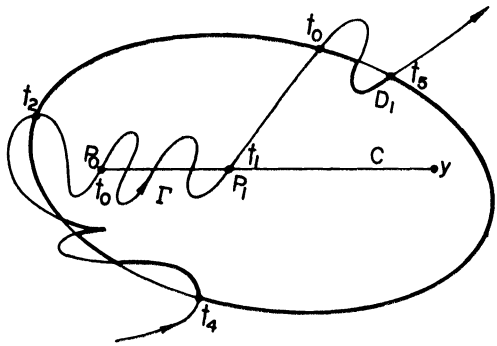

FIGURE 9.

point of H_1 and H_2 along C. Then this point is on Γ, since all frontier points of H_1 and H_2 are on Γ. Also, since Γ has no limit points in D, there must be a point $p_1 \in C$ such that $\Gamma(t_1) = p_1$ (assume $t_1 > t_0$) while $\Gamma(t) \notin C$ for $t > t_1$. Then every neighborhood of p_1 contains points of H_1, or else there would again be a frontier point of H_1 and H_2 on C between p_1 and y, which would be a point of Γ.

Let $\Gamma_1 \equiv \{\Gamma(t) : t_0 \leqq t \leqq t_1\}$. Take any simply connected open set D_1 in D, which contains Γ_1 in its interior, and has a simple closed curve in D as a frontier. Let $\bar{D}_1$ denote the closure of D_1 and $\mathfrak{F}D_1$ denote its frontier.

Define t_2 as the largest value of $t < t_0$ such that $\Gamma(t_2) \in \mathfrak{F}D_1$; t_3 as the smallest value of $t \leqq t_0$ such that $\Gamma(t_3) \in \mathfrak{F}D_1$; t_4 such that $\Gamma(t) \notin \bar{D}_1$ for $t < t_4$; while t_5 is such that $\Gamma(t) \notin \bar{D}_1$ for $t > t_5$. (See Figure 9.)

Let E_2 be the set whose frontier is made up of arcs of $\mathfrak{F}D_1$ and arcs of $\{\Gamma(t) : t \leqq t_2 \text{ or } t \geqq t_3\}$ by using the subarcs of the latter whenever they lie within $\bar{D}_1$. It is easily verified that D_2 is a simply connected domain, and $\Gamma_2 \equiv \{\Gamma(t) : t_2 \leqq t \leqq t_3\}$ is a cross cut. Thus (Chapter V, Theorem 11.7 [13]) $D_2 - \Gamma_2$ has two components, and Γ_2 is contained in the frontier of both. Now since neighborhoods of p_1 contain points of H_1, while there is a neighborhood $n(p_0)$ such that $n(p_0) - \Gamma$ contains only points of H_2, it follows that there must be a point of H_1 and a point of H_2 in the same component of $D_2 - \Gamma_2$. Connect these with a simple arc in that component. This arc must contain a frontier point of H_1, hence a point of Γ, which is a contradiction.

J.S.I.A.M. CONTROL
Ser. A, Vol. 1, No. 2
Printed in U.S.A., 1963

AN OPERATOR THEORETIC FORMULATION OF A CLASS OF CONTROL PROBLEMS AND A STEEPEST DESCENT METHOD OF SOLUTION*

A. V. BALAKRISHNAN†

1. Introduction. In this paper we examine a class of control problems in the context of abstract functional analysis. The formulation and method of solution are succinct in form and concept and very general in scope. The class of problems includes the linear final value control problems [1, 2] as well as a more general class. In the former we seek to minimize the Euclidean distance

$$\| x_0 - x(T) \|,$$

where x_0 is the desired (finite-dimensional) position vector (or point in phase space) and $x(T)$ is the actual response upon using a control $u(t)$:

$$x(T) = \int_0^T W(T, s)u(s)\, ds,$$

and the optimization problem is to choose the appropriate $u(t)$ from a given class. A more general class of problems is that of minimizing the integral square error

$$\int_0^T \| g(t) - \int_0^T W(t, s)u(s)\, ds \|^2\, dt$$

where $g(t)$ is the desired response and

$$x(t) = \int_0^T W(t, s)u(s)\, ds$$

is the actual response using the control $u(t)$ and $u(t)$ is again subject to constraint. For the case of an energy constraint, we also present a computational algorithm which in sufficiently many steps will approximate as close to the optimum as required. This algorithm is based on the method of steepest descent in a Hilbert space. We shall present some numerical computer results employing the methods advocated in this paper in the near future.

2. Abstract formulation of the optimization problem. We begin by formulating the final value control problem in the context of linear operators over a Hilbert space. In a large number of cases the control system can be

* Received by the editors March 9, 1962 and in revised form November 1, 1962.
† Department of Engineering, University of California, Los Angeles, California.

characterized by the dynamical equations

$$\frac{d}{dt}x(t) = A(t)x(t) + B(t)u(t), \tag{2.1}$$

where for each t,

$x(t)$ is a p-dimensional (column) vector or $p \times 1$ matrix,
$A(t)$ is a $p \times p$ matrix,
$B(t)$ is a $p \times q$ matrix, and
$u(t)$ is a q-dimensional (column) vector (or $q \times 1$ matrix) specified usually to be in some closed convex subset of the Euclidean q-space.

For any rectangular matrix M we shall define the norm, denoted $\| M \|$, by

$$\| M \|^2 = \text{trace } M^*M,$$

the star denoting the conjugate transpose. Given a preassigned vector x in the Euclidean p-space, the control problem that concerns us is that of choosing a Lebesgue measurable control function $u(t)$, $0 \leqq t \leqq T$, so as to minimize

$$\| x(T) - x \|^2,$$

assuming $A(t)$, $B(t)$ continuous and

$$\int_0^T \| u(t) \|^2 \, dt \qquad < \infty;$$

we may without loss of generality assume $x(0) = 0$. The point of departure for us is that from (2.1) we know first that there is a $p \times q$ matrix $W(t, s)$, $0 \leqq t, s \leqq T$ such that

$$x(T) = \int_\Lambda W(T, s)u(s) \, ds, \tag{2.2}$$

where

$$\int_\Lambda \| W(T, s) \|^2 \, ds \qquad < \infty, \tag{2.3}$$

Λ being the interval $[0, T]$.

If we denote by H the Hilbert (L_2) function space of Lebesgue measurable q-dimensional complex-valued functions square integrable over Λ, with inner product of any two functions $f(t)$, $g(t)$ in L_2 defined by

$$\left[f, g \right] = \int_\Lambda g(t)^* f(t) \, dt,$$

then we can define a linear operator on H by

$$Lu = \int_\Lambda W(T, s)u(s)\, ds \tag{2.3a}$$

which maps H into the unitary (Hilbert) space E_p of dimension p. By virtue of (2.3), L is a linear bounded transformation. We can then phrase the control problem at hand as follows:

Given any x is E_p, minimize the distance

$$\| Lu - x \| \tag{2.4}$$

where u is restricted to be in a closed convex set C in H. Let it be noted that this is more general than (2.1) since we can consider any linear system, not merely those described by differential equations, and all we require is the function $W(t, s)$ which need not even be continuous. Also, letting the control functions u be in C is in a sense more general than requiring that the function be in a convex set at each point, and in particular it is not required that the control functions be piecewise continuous.

For (2.4) we can establish the following existence theorems based on standard results in Hilbert space theory.

THEOREM 2.1. *For each x in E_p there is a unique element z in the closure of $L(C)$ such that*

$$\inf_{u \in C} \| Lu - x \| = \| z - x \|. \tag{2.5}$$

Proof. We give a proof which is independent of the finite dimensionality of E_p and hence includes the general case where L maps H into another Hilbert space. The proof is based on the following well-known lemma:

LEMMA 2.1. *Every closed convex set in a Hilbert space has a unique element of minimal norm.*

Proof. For a proof see [3].

To prove Theorem 2.1 we have only to note that the set, denoted $L(C)$, consisting of all elements of the form Lu, $u \in C$, is convex, and hence so is the set of all elements of the form

$$Lu - x, \qquad u \in C.$$

Since the strong closure of this set is still convex, Lemma 2.1 applies and, in particular, shows that there is a sequence $\{u_n\}$ of elements in C such that the sequence Lu_n converges to a unique element z in the closure of $L(C)$ and

$$\inf_{u \in C} \| Lu - x \| = \lim_n \| Lu_n - x \| = \| z - x \|.$$

We note that L takes bounded sets in H into bounded sets in E_p, and since the closure of a bounded set in E_p is compact, L has the property

that it maps bounded sets in H into conditionally compact sets in E_p. A linear bounded transformation of one Hilbert space into another which has this property is called "compact" or "completely continuous". Using this we have

THEOREM 2.2. *Let C be any closed bounded convex set in H. Then there is a unique element u_0 of minimal norm in C such that*

$$\inf_u \| Lu - x \| = \| Lu_0 - x \|. \tag{2.6}$$

Proof. Let $\{u_n\}$ be a sequence of elements in C such that

$$\lim_{n\to\infty} \| Lu_n - x \| = \inf_{u\in C} \| Lu - x \|.$$

We recall [3] that the unit sphere in H is weakly compact, so that every bounded sequence contains a weakly convergent subsequence. Hence there is an element v in H such that a subsequence $\{u_{n_k}\}$ converges weakly to v. But since L is compact, we know [3] also that Lu_{n_k} converges strongly to Lv. Then

$$\inf_{u\in C} \| Lu - x \| = \| Lv - x \|,$$

and since C is convex, Lv by Theorem 2.1 is unique. Since C is convex and closed, we also have the deeper result [3] that v itself must belong to C. Next consider the class of elements u in C such that

$$Lu = Lv.$$

This is a closed convex set, and by Lemma 2.1 must contain a unique element of minimal norm. This is the element u_0 sought in the theorem.

The main problem is of course that of finding the element u_0 or an approximating sequence. We shall describe such a method for the case where C is actually a sphere. In order to do so, however, we shall need a few more general results.

Let L^* be the adjoint of L, so that L^* maps E_p into H:

$$L^*x = u,$$

where the function u is defined by

$$u(t) = W(T, t)^*x, \qquad 0 \leqq t \leqq T.$$

Then L^*L is a linear bounded transformation defined on H, mapping H into H and similarly, LL^* is a linear bounded transformation mapping E_p into E_p. Also, both LL^* and L^*L are compact and have the same non-zero

eigenvalues. Moreover, for each x in E_p,

$$LL^*x = \int_\Lambda W(T, t)W(T, t)^*x\, dt$$

$$= Ax,$$

where A is the $p \times p$ matrix

$$A = \int_\Lambda dtW(T, t)W(T, t)^*$$

which is self-adjoint ("Hermitian") and non-negative definite.

Let $\{\lambda_i\}$, $i = 1, \cdots, m \leqq p$; $\lambda_1 \geqq \lambda_2 \cdots \geqq \lambda_m$, be the non-zero (and hence positive) eigenvalues of A and let the e_i be the corresponding orthonormalized eigenvectors. Then denoting by $[\ ,\ ]$ the inner product and by $\|\ \|$ the norm in both H and E_p, we note that

$$\phi_i = \frac{L^*e_i}{\lambda_i^{1/2}}$$

are orthonormalized eigenvectors for L^*L, since

$$L^*L\,\phi_i = \frac{L^*LL^*e_i}{\lambda_i^{1/2}} = \lambda_i \frac{L^*e_i}{\lambda_i^{1/2}} = \lambda_i\phi_i.$$

Also for each u in H, we have

$$Lu = \sum_{i=1}^{m} [u, \phi_i]L\phi_i \tag{2.8}$$

and

$$L^*x = \sum_{i=1}^{m} [x, e_i]L^*e_i\,.$$

It can also be shown that

$$\sup_{\|u\| \leqq 1} \| Lu \|^2 = \lambda_1\,.$$

Again since H is infinite dimensional and E_p is finite dimensional, zero is always an eigenvalue of L. On the other hand, zero is not necessarily an eigenvalue of L^*, since for this we must have

$$W(T, t)^*x = 0 \tag{2.9}$$

for almost every t in Λ and some $x \neq 0$.
Again,

$$\inf_{u \in H} \| Lu - x \| > 0$$

for some x in E_p if and only if zero is an eigenvalue of L^*. For, if

$$L^*x = 0,$$

then for every u in H

$$[u, L^*x] = [Lu, x] = 0,$$

so that the range of L is properly contained in E_p. And conversely, if the range of L is properly contained in E_p, then there is a non-zero x such that

$$[Lu, x] = 0,$$

for every u in H, so that

$$[u, L^*x] = 0,$$

implying that

$$L^*x = 0. \tag{2.10}$$

It should perhaps be noted that (2.10) implies that

$$\inf_u \| Lu - kx \| > 0$$

for every nonzero scalar k; so that a "pure gain" constant does not result in zero error.

In terms of the eigenvector expansion (2.8), we can readily determine an optimal u in H which minimizes

$$\| Lu - x \|$$

for a given x in E_p. For this we note first that

$$\| Lu - x \|^2 = [L^*Lu, u] - [Lu, x] - [x, Lu] + [x, x], \tag{2.11}$$

so that minimizing (2.4) is equivalent to minimizing

$$[L^*Lu, u] - [u, g] - [g, u] \tag{2.11a}$$

for given $g = L^*x$ in H. Suppose there is an h in H such that

$$L^*Lh = L^*x. \tag{2.12}$$

Then

$$\| Lu - x \|^2 = [L^*L(u - h), (u - h)] + [x, x] - [Lh, Lh] \tag{2.13}$$

$$\geqq [x, x] - [Lh, Lh], \tag{2.14}$$

and equality holds if and only if

$$u = h. \tag{2.15}$$

Note that we have established that the minimum of (2.13) is given by (2.14) without resort to any variational arguments. With variational arguments we only obtain a local extremum. We have shown here that the minimum is actually an "absolute" minimum. Using (2.8) we obtain further that the element in (2.12) actually exists and can be computed from

$$L^*Lh = \sum_1^m \lambda_i[h, \phi_i]\phi_1 = L^*x = \sum_1^m [x, e_i]L^*e_i,$$

giving

$$h = \sum_1^m \frac{[x, e_i]}{\lambda_i} L^* e_i = \sum_1^m \frac{[L^*x, \phi_i]}{\lambda_i} \phi_i . \tag{2.16}$$

It is perhaps necessary to add that (2.12) does not necessarily imply that

$$Lh = x.$$

It should also be noted that h is in the range of L^*.

Denoting L^*L by R, we can define a self-adjoint positive square root for R by various means [3], for instance using (2.8):

$$R^{1/2}u = \sum_1^m [u, \phi_i]\lambda_i^{1/2}\phi_i . \tag{2.17}$$

In terms of $R^{1/2}$, if we set

$$v = \sum_1^m \frac{[L^*x, \phi_i]}{\lambda_i^{1/2}} \phi_i , \tag{2.18}$$

then using (2.8) and (2.17),

$$R^{1/2}v = \sum_1^m [L^*x, \phi_i]\phi_1 = \sum_1^m [x, e_i]L^*e_i = L^*x$$

and we obtain

$$\begin{aligned} \| Lu - x \|^2 &= [R^{1/2}u, R^{1/2}u] - [u, L^*x] + [x, x] - [L^*x, u] \\ &= \| R^{1/2}u - v \|^2 + [x, x] - [v, v] \end{aligned} \tag{2.19}$$

This shows in particular that we are seeking the minimum of the first term, and reduces the problem, of course, to one completely in H. It is of interest to examine whether the problem can also be stated as one entirely in E_p, at least when we wish to obtain the unique element of minimal norm whose existence has been proved in Theorem 2.2. For this, let $\bar{R}$ be the orthogonal complement of $\bar{Z}$ with $\bar{Z}$ the null space of L. Then $\bar{R}$ is precisely the range of L^*, since

$$[L^*x, u] = [x, Lu] = 0$$

for u in $\bar{Z}$. Let C_1 be the intersection of C and $\bar{R}$. This intersection cannot be empty, unless C is contained in $\bar{Z}$, in which case the minimization problem reduces to finding simply the element of minimal norm in C. Hence we may assume that C is not contained in $\bar{Z}$. If C contains the origin (that is, the zero element) as an interior point, for instance, C cannot obviously be contained in $\bar{Z}$ since

$$Lu = 0$$

for u in the sphere about the origin contained in C implies that the same is true for every element in H. In what follows we shall assume that C is closed and bounded, and contains the origin as an interior point. For a given x in E_p, let z be the minimizing element such that

$$\min_{u \in C} \| Lu - x \|^2 = \| z - x \|^2,$$

and consider the set of u in C such that

$$Lu = z.$$

Every such u can be written as

$$u_z + u_R$$

where u_R is the projection of u in $\bar{R}$ and is uniquely determined by z, and u_z is the projection of u in $\bar{Z}$, and

$$\| u \|^2 = \| u_z \|^2 + \| u_R \|^2.$$

In some cases, C may be such that this will imply that u_R is the element of minimal norm sought—for instance, if C is a sphere, or more generally, is specified by a functional inequality: C is the class of all elements u in H such that

$$f(u) \leqq M$$

where $f(u)$ is positive and

$$f(u_1) \geqq f(u_2) \quad \text{if} \quad \| u_1 \| \geqq \| u_2 \|,$$
$$f(u_1 + u_2) \leqq f(u_1) + f(u_2),$$
$$f(\alpha u) = \alpha f(u) \quad \text{for} \quad \alpha \quad \text{positive.}$$

In this case, then, we can write

$$\begin{aligned} \| Lu - x \|^2 &= \| LL^*y - x \|^2 \\ &= [LL^*y, LL^*y] - [LL^*y \ x] + [x \ x] - [x, LL^*y], \end{aligned}$$

which is stated entirely in terms of elements y in E_p, and y is such that

$$L^*y \in C_1.$$

The case where C is a sphere of radius M about the origin can be solved more explicitly using the eigenvectors $\{e_i\}$ or the eigenfunctions $\{\phi_i\}$. Thus since we know that the optimal element of minimal norm can be written

$$u = \sum_{1=1}^{m} a_i \phi_i ,$$

where

$$\sum_{1}^{m} a_i^2 \leqq M^2, \tag{2.20}$$

we are seeking the minimum of (using (2.13) and (2.16))

$$[L^*L(u - h), u - h] = \sum_{1}^{k} \lambda_i [a_i - c_i]^2$$

where

$$c_i = \frac{[L^*x, \phi_i]}{\lambda_i}$$

subject to (2.20). This is a finite dimensional problem, and we can therefore get a local extremum by using Lagrange multipliers. Thus we minimize

$$\sum_{1}^{m} \lambda_i (a_i - c_i)^2 + \sum_{1=1}^{m} \lambda a_i^2,$$

and obtain the solution as

$$a_i = \frac{\lambda_i c_i}{\lambda_i + \mu}, \qquad \mu \geqq 0 \tag{2.21}$$

where μ is adjusted so that

$$\sum_{1}^{m} a_i^2 = M^2.$$

We note that, for this choice of $\{a_i\}$,

$$[R + \mu I] \sum_{1}^{m} a_i \phi_i = \sum_{1}^{m} \lambda_i c_i \phi_i = L^*x,$$

where I is the identity operator on H. In other words, since $[R + \mu I]$ has a bounded inverse for every $\mu > 0$, the optimal element is of the form

$$[R + \mu I]^{-1} L^*x.$$

We have used the Lagrange multiplier method to obtain this result. We shall now give a general proof that it is indeed the optimal element sought, and, moreover, we shall not make use of the finite dimensionality of E_p

in the proof. For this let, for each u in H,

$$Q(u) = [Ru, u] - 2\mathrm{Re}\,[u, g] \tag{2.22}$$

where Re stands for "real part," and where

$$R = L^*L$$

and

$$g = L^*x.$$

We note that minimizing $\| Lu - x \|^2$ is the same as minimizing $Q(u)$. The general problem is then to minimize the form $Q(u)$ over H, subject to $\| u \|^2 \leqq M^2$, the operator R being compact, self-adjoint and non-negative. Since R is non-negative, for each positive number k, $(R + kI)$ has a linear bounded inverse.

Let

$$u_k = [R + kI]^{-1}g.$$

Then $\| u_k \|^2$ is a monotone decreasing function of k. Indeed,

$$\| u_{k_2} \|^2 - \| u_{k_1} \|^2$$
$$= [k_1 - k_2][(R + k_1I)^{-2}(R + k_2I)^{-2}(2R + (k_1 + k_2)I)g, g],$$

and the second factor is positive. Moreover $\| u_k \|^2$ goes to zero as k goes to infinity. Next, let us consider the problem of minimizing, for fixed k,

$$Q_k(u) = [(R + kI)u, u] - 2\,\mathrm{Re}\,[u, g].$$

We have

$$Q_k(u) = [(R + kI)u, u] - 2\,\mathrm{Re}\,[u, (R + kI)u_k]$$
$$= [(R + kI)(u - u_k), (u - u_k)] - [(R + kI)u_k, u_k]$$

where the second term is positive and fixed, while the first term is non-negative. Hence the minimum is attained at $u = u_k$:

$$\inf_{u \in H} Q_k(u) = Q_k(u_k) = -[(R + kI)u_k, u_k] = -[g, u_k]. \tag{2.23}$$

We shall need to distinguish between two cases.

Case i. Suppose

$$\sup_{k>0} \| u_k \|^2 \leqq M^2.$$

Since R is compact and non-negative, let us use the fact that it has at most a countable number of positive eigenvalues $\{\lambda_i\}$, and let $\{\phi_i\}$ be a cor-

responding set of orthonormalized eigenfunctions. Since g is in $\bar{R}$, we know that

$$g = \sum_1^\infty [g, \phi_i]\phi_i .$$

Also

$$u_k = \sum_1^\infty \frac{[g, \phi_i]}{\lambda_i + k} \phi_i ,$$

so that

$$\| u_k \|^2 = \sum_1^\infty \frac{|[g, \phi_i]|^2}{(\lambda_i + k)^2} \leqq M^2.$$

Hence

$$\sum_1^\infty \frac{|[g, \phi_i]|^2}{\lambda_i^2} \leqq M^2,$$

and we can define

$$u_0 = \sum_1^\infty \frac{[g, \phi_i]}{\lambda_i} \phi_i .$$

Then u_k converges to u_0 as $k \to 0$, and $Ru_0 = g$. As we have seen, this is enough to prove that u_0 minimizes $Q(u)$ and, u_0 being in $\bar{R}$, is also the unique element of minimal norm which minimizes $Q(u)$.

Case ii. Suppose next that

$$\sup_{k>0} \| u_k \|^2 > M^2.$$

Then there is obviously a positive number, call it k_0, such that

$$\| u_{k_0} \|^2 = M^2.$$

Let

$$Q_0 = \inf_{u \in C} Q(u)$$

Now we know from Theorem 2.2 that there is an element u_0 in C such that

$$Q_0 = Q(u_0)$$

It should be noted that only the compactness of L was used in the argument, and not the finite dimensionality of E_p. Let

$$\| u_0 \|^2 = m^2 \leqq M^2.$$

We shall first show that actually $m^2 = M^2$. For,

$$\begin{aligned} Q(u_0) = \inf_{\|u\|^2=m^2} Q(u) &= \inf_{\|u\|^2=m^2} [(R + kI)u, u] - 2\,\mathrm{Re}\,[u, g] - km^2 \\ &\geqq \inf_{u \in H} [(R + kI)u, u] - 2\,\mathrm{Re}\,[u, g] - km^2 \\ &= Q_k(u_k) - km^2 \\ &= Q(u_k) + k[\|u_k\|^2 - m^2]. \end{aligned}$$

Hence, in particular,

$$Q(u_0) \geqq Q[u_{k_0}] + k_0[\|u_{k_0}\|^2 - m^2].$$

But u_{k_0} is in C, and k_0 is positive. Hence we must have

$$M^2 = \|u_{k_0}\|^2 = m^2 = \|u_0\|^2. \tag{2.24}$$

Also, by Theorem 2.1,

$$\begin{aligned} L[u_0] &= L[u_{k_0}] \\ L[u_{k_0} - u_0] &= 0, \end{aligned}$$

so that

$$u_0 = u_{k_0} + u_z,$$

where u_z is an element in $\bar{Z}$ the null space of L (or R). But

$$[u_z, u_{k_0}] = 0,$$

since

$$\begin{aligned} [u_z, u_{k_0}] &= [u_z, (R + k_0 I)^{-1} g] \\ &= [(R + k_0 I)^{-1} u_z, g] \\ &= [u_z/k_0, g] = [Lu_z/k_0, x] = 0, \end{aligned}$$

and hence

$$\|u_0\|^2 = \|u_{k_0}\|^2 + \|u_z\|^2$$

so that by (2.24), u_z must be zero.

We may now state our result in its full generality.

THEOREM 2.3. *Let L be a compact linear bounded transformation mapping a Hilbert space H_1 into another Hilbert space H_2. Suppose for a given x in H_2 it is required to minimize*

$$\|Lu - x\|^2,$$

subject to u being in the sphere C in H_1.

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

Then either

$$\sup_{k>0} \|[L^*L + kI]^{-1}L^*x \| \leqq M \tag{2.25}$$

in which case the sequence

$$u_k = [L^*L + kI]^{-1}L^*x$$

is such that u_k converges to the optimal element u_0 of minimal norm

$$\lim_{k\to 0} \| Lu_k - x \|^2 = \inf_{u\in C} \| Lu - x \|^2 = \| Lu_0 - x \|^2,$$

or

$$\sup_{k>0} \|(L^*L + kI)^{-1}L^*x \| > M$$

in which case

$$u_0 = [L^*L + k_0I]^{-1}L^*x \tag{2.26}$$

where k_0 is adjusted so that

$$\| u_0 \| = M$$

yields the unique solution to the minimization problem.

As a corollary to this theorem we may note that the optimal solution is always in the range of L^*, or its closure. This is immediate if condition (2.26) holds, since if

$$[R + kI]^{-1}L^*x = u,$$

then

$$L^*x = [R + kI]u = L^*Lu + ku$$

or

$$ku = L^*[x - Lu].$$

If condition (2.25) holds, and

$$[R + k_nI]^{-1}L^*x = u_n,$$

then similarly

$$k_nu_n = L^*[x - Lu_n],$$

so that u_n is in the range of L^* for every n. It may be also noted that in this case

$$x = Lu_0 + z$$

where $L^*z = 0$.

This generalization for instance includes the problem of minimizing

$$\int_0^T \left\| \int_0^T W(t, s)u(s)\, ds - x(t) \right\|^2 dt,$$

where it is assumed that

$$\int_0^T \| x(t) \|^2 dt < \infty, \qquad \int_0^T \int_0^T \|W(t, s) \|^2 ds\, dt < \infty,$$

and it is required that

$$\int_0^T \| u(t) \|^2 dt < M^2.$$

3. Method of steepest descent. We next consider the problem of actually computing the optimal control function. If we know what the (non-zero) eigen values and eigen vectors of LL^* or L^*L are, we can, as we have seen, display an explicit solution in terms of these. On the other hand, in many cases, such a determination can be difficult; and in any event, since what is really needed is an approximation to the optimum for each given x, a method of iteration which is proven to converge to the optimum is of value. We shall now indicate an iterative "steepest descent" method in Hilbert space for the case where the convex set C is a sphere with center at the origin.

It has been already noted that the optimal solution is in the range of L^* or its closure. We may therefore work either with LL^* or L^*L. We shall describe the iteration method in the generality of Theorem 2.3, since the fact that H_2 (or E_p) is finite dimensional will not be used. Reference may be made to the work of Kantarovich [4] who appears to have been among the first to describe a steepest descent method in Hilbert space, which in our context is applicable to the case where C is the whole space H, or the sphere is of infinite radius. As before, (2.22), we may state the problem in H as that of minimizing

$$Q(u) = [Ru, u] - 2 \operatorname{Re} [u, g],$$

where

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

Let u_n be the n-th iteration. We define, as a function of k, $0 < k < \infty$.

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

where

$$\epsilon_n = [z_n , z_n]/[(R + kI)z_n , z_n]$$
$$z_n = (R + kI)u_n - g.$$

Now we note that

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

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

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

then we define the $(n + 1)$-th iteration as

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

Otherwise, we choose the positive number k_n such that

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

and define the $(n + 1)$-th iteration as

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

We shall first show that thus defined,

$$\lim_n Q(u_n) = \inf Q(u), \qquad \| u \|^2 \leqq M^2. \tag{3.1}$$

For this, let

$$Q_{n+1}(u) = [(R + k_n I)u, u] - 2 \operatorname{Re} [u, g].$$

Then we have:

$$Q_{n+1}(u_{n+1}) = Q_n(u_n) + [k_n - k_{n-1}][u_n , u_n] - r_n^2 \tag{3.2}$$

where

$$r_n^2 = \frac{[z_n , z_n]^2}{[(R + k_n I)z_n , z_n]}.$$

It is convenient to distinguish between several cases. Suppose first that $k_n = 0$ for $n \geqq$ some N. Then for $n \geqq (N + 1)$, we note from (3.2) that

$$Q_{n+1}(u_{n+1}) = Q_n(u_n) - r_n^2.$$

Hence also

$$Q(u_{n+1}) = Q(u_n) - r_n^2,$$

so that $Q(u_n)$ is a monotone non-increasing sequence and

$$Q(u_{n+p}) = Q(u_n) - \sum_{n}^{n+p-1} r_k^2.$$

Since from (2.11),

$$Q(u_n) \geqq -[x, x],$$

it follows that the infinite series $\sum_N^\infty r_k^2$ converges, and hence in particular r_n goes to zero. But

$$r_n^2 = \frac{[z_n, z_n]^2}{[Rz_n, z_n]} \geqq \frac{[z_n, z_n]^2}{\| R \| \| z_n \|^2} = \frac{\| z_n \|^2}{\| R \|}$$

where $\| R \|$ denotes the norm of the operator R and is positive. Hence

$$\| z_n \| = \| Ru_n - g \| \to 0.$$

Making use of the fact that the unit sphere in a Hilbert space is weakly compact, we note that any subsequence of $\{u_n\}$ contains a further subsequence, call it $\{u_{n_k}\}$, such that u_{n_k} converges weakly to an element u_0 in the space. Since R is compact, Ru_{n_k} converges to Ru_0 and hence

$$Ru_0 = g. \tag{3.3}$$

Suppose we choose the first iteration u_1 in the range of L^*. Then u_n will be in the range of L^* for every n, since

$$u_n = u_{n-1} - \epsilon_{n-1} L^*[Lu_n - g].$$

Now the range of L^* is contained in $\bar{R}$ which is a subspace, and hence u_0 also belongs to it. Hence u_0 is unique; that is, it is independent of the particular subsequence chosen, because of (3.3). Hence the sequence u_n itself converges weakly to u_0. Now

$$Q(u_n) = [Ru_n - g, u_n] - [u_n, g]$$

and hence converges to

$$Q(u_0) = -[u_0, g],$$

and since $Ru_0 = g$, we know that u_0 minimizes $Q(u)$ and provides the element of minimal norm which does so.

Next suppose $k_n \neq 0$ for every $n \geqq (N + 1)$. Using (3.2) again, we can write

$$\begin{aligned} Q_{n+1}(u_{n+1}) = Q_m(u_m) &+ \sum_{m}^{n-1} k_i\{[u_i, u_i] - [u_{i+1}, u_{i+1}]\} \\ &+ k_n[u_n, u_n] - k_{m-1}[u_m, u_m] - \sum_{m}^{n} r_i^2. \end{aligned} \tag{3.4}$$

Using the fact that now for n larger than $(N + 1)$, $\| u_n \|^2 = M^2$, we can reduce (3.4) to

$$Q(u_{n+1}) = Q(u_m) - \sum_m^n r_i^2, \qquad n > m$$

for m larger than $(N + 1)$. But then, as before, this implies that $Q(u_n)$ converges and

$$r_n^2 \geqq \frac{\| z_n \|^2}{\| R + k_n I \|} \geqq \frac{[k_n M - \| Ru_n - g \|]^2}{\| R \| + k_n}.$$

Since r_n^2 goes to zero it follows from this that k_n must be bounded. But

$$r_n^2 \geqq \frac{\| z_n \|^2}{\| R \| + \sup k_n},$$

so that

$$z_n = [R + k_n I]u_n - g \to 0. \tag{3.5}$$

Now $\| u_n \|^2 = M^2$ and, since R is compact, any subsequence contains a subsequence, call it $\{u_{n_i}\}$, such that Ru_{n_i} converges and hence by (3.5), $k_{n_i}u_{n_i}$ converges, and hence also

$$k_{n_i} \| u_{n_i} \| = k_{n_i} M,$$

so that $\{k_{n_i}\}$ converges to k_0, say. Suppose first that $k_0 \neq 0$. Then dividing through by k_{n_i} in (3.5),

$$Ru_{n_i}/k_{n_i} + u_{n_i} \to g/k_0 .$$

Hence the sequence $\{u_{n_i}\}$ itself converges. Let u_0 be the limit. Then

$$u_0 = [R + k_0 I]^{-1} g.$$

Also every subsequence $\{k_{n_i}\}$ must have the same limit since

$$\| u_{n_i} \|^2 = M^2 = \| u_0 \|^2,$$

and there is only one value of k_0 such that

$$\| [R + k_0 I]^{-1} g \| = M. \tag{3.6}$$

This shows that $\{u_n\}$ itself must converge to u_0. Next suppose that

$$k_0 = 0.$$

Then again, every subsequence k_{n_i} must have limit zero, for if

$$\text{limit } Ru_{n_i} = g, \qquad \| u_{n_i} \| = M$$

and

$$[R + k_0 I]^{-1} g = u_0 , \qquad \| u_0 \| = M, \qquad k_0 > 0,$$

we have a contradiction which follows from the fact that

$$\text{limit } \|(R + k_0 I)^{-1} R u_{n_i}\| = \|u_0\| = M.$$

On the other hand, the norm of the operator $(R + k_0 I)^{-1} R$ is actually less than one (R being compact and non-negative) so that

$$\|(R + k_0 I)^{-1} R\| \, \|u_{n_i}\| > \|u_{n_i}\| = M.$$

Hence it follows that

$$\lim_n R u_n = g$$

and since $Q(u_n)$ converges, this is again enough to prove (3.1).

Hence the only remaining case we have to consider is one in which there is a subsequence $k_{n_i} = 0$ and a subsequence $k_{m_n} \neq 0$. Suppose then that

$$\begin{aligned} k_m &\neq 0 \\ k_{m+1} &= 0 \\ &\vdots \quad \vdots \\ k_{n-1} &= 0 \\ k_n &\neq 0. \end{aligned}$$

We have using (3.2),

$$Q(u_{n+1}) = Q(u_{m+1}) - \sum_{m+2}^{n-1} r_i^2,$$

implying again that r_n goes to zero and hence that

$$\lim_n (R + k_n I) u_n = g.$$

Arguing, as before, it follows that the sequence k_n itself converges, and so does $\{u_n\}$, establishing (3.1) again. It must be noted that the method of iteration does not require any a priori estimates concerning x.

4. Relation to prediction theory. It is well-known that the optimal mean square prediction (or estimation) problem for stochastic processes can be stated as a minimization of a quadratic form over Hilbert space, the original treatment of the problem by Kolmogorov in 1941 [5] being already in Hilbert space language. The explicit connection is simply given by the step from (2.11) to (2.11a) and has been enunciated as a "duality principle" by R. E. Kalman [6]. In (2.11) the operator $R = L^*L$ is a non-negative "integral" operator with the corresponding kernel function being non-negative definite. Let $x(t)$ be a Gaussian process having this function for its covariance and suppose

$$x(t) = s(t) + N(t)$$

where $s(t)$ and $N(t)$ are Gaussian processes (not necessarily independent) and let it be required to predict (estimate) $s(T + a)$ from $x(t), 0 \leqq t \leqq T$ with minimum mean square error. Then if we denote the optimal operation by

$$\int_0^T x(t)u(t)\, dt,$$

the function $u(t)$ minimizes the form

$$[Ru, u] - 2[c, u] \tag{4.1}$$

where u is now being considered in the L_2 space with respect to Lebesgue measure over $[0, T]$ and

$$c(t) = E[s(T + a)x(t)].$$

Thus in both the "control" problem and the "prediction" problem we are minimizing the same form, but there is the important difference that in the "prediction" problem u is free to be any element of the space, not necessarily restricted to a convex set. There is also no unique way of going back from (4.1) to the control problem, since R can be factored as L^*L in many ways, unless we require "physical realizability", that is, we require in

$$Lu = v$$

that

$$v(t) = \int_0^t W(t, s)u(s)\, ds.$$

The conditions under which, and the means by which, this can be done are still largely unsolved. Also there are many prediction problems involving non-linear operations, the control interpretation of which has dubious meaning.

REFERENCES

[1] A. Rosenbloom, *Final value systems with total effort constraints*, Proceedings of the First International Congress of the International Federation of Automatic Control, Butterworths, London, 1961.

[2] R. C. Booton, *Optimum design of final value control systems*, Proceedings of the Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, New York, 1956, pp. 233–242.

[3] F. Riesz and Sz. Nagy, *Functional Analysis*, Frederick Ungar Publishing Co., New York, 1955.

[4] L. V. Kantarovich, *Functional Analysis and Applied Mathematics*, Usp. Mat. Nauk., 3 (1948) pp. 89–185; English translation, National Bureau of Standards, 1953.

[5] A. Kolmogorov, *Interpolation und extrapolation von stationaren zufälligen folgen*, Bulletin de l'Académie des Sciences des U.R.S.S. Ser. Math., 5 (1941), pp. 3–14.

[6] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. A.S.M.E. Ser. D: J. Basic Engrg., 82 (1962), pp. 35–45.

J.S.I.A.M. CONTROL
Ser. A, Vol. 2, No. 1
Printed in U.S.A., 1963

CONTROLLABILITY AND OBSERVABILITY IN MULTIVARIABLE CONTROL SYSTEMS*

ELMER G. GILBERT†

1. Introduction. The importance of linear multivariable control systems is evidenced by the large number of papers [1–12] published in recent years. Despite the extensive literature certain fundamental matters are not well understood. This is confirmed by numerous inaccurate stability analyses, erroneous statements about the existence of stable control, and overly severe constraints on compensator characteristics. The basic difficulty has been a failure to account properly for *all* dynamic modes of system response. This failure is attributable to a limitation of the transfer-function matrix—it fully describes a linear system if and only if the system is controllable and observable.

The concepts of controllability and observability were introduced by Kalman [13] and have been employed primarily in the study of optimal control.[1] In this paper, the primary objective is to determine the controllability and observability of composite systems which are formed by the interconnection of several multivariable subsystems. To avoid the limitations of the transfer-function matrix, the beginning sections deal with multivariable systems as described by a set of n first order, constant-coefficient differential equations. Later, the extension to systems described by transfer-function matrices is made. Throughout, emphasis is on the fundamental aspects of describing multivariable control systems. Detail design procedures are not treated.

2. Definitions and notation. Let a multivariable system S be represented by

$$\begin{aligned} \dot{x} &= Ax + Bu \\ v &= Cx + Du \end{aligned} \tag{1}$$

where:

$u = u(t)$, p-dimensional input vector.
$v = v(t)$, q-dimensional output vector.
$x = x(t)$, n-dimensional state vector, n is the order of S.
$\dot{x} = \dot{x}(t)$, time derivative of state
A, constant nth order differential transition matrix.

* Received by the editors July 5, 1962 and in revised form November 1, 1962. Presented at the Symposium on Multivariable System Theory, SIAM, November 1, 1962 at Cambridge, Massachusetts.

† Instrumentation Engineering, University of Michigan, Ann Arbor, Michigan.

[1] Reference [14] gives a historical account of controllability and lists other references.

B, constant, n row, p column, input matrix.

C, constant, q row, n column, output matrix.

D, constant, q row, p column, transmission matrix.

If $n = 0$ the system is said to be *static.*

The characteristic roots λ_i, $i = 1, \cdots n$, of A are assumed to be distinct. This greatly simplifies the proof of theorems and prevents the main course of the paper from becoming obscured. Besides, there are few practical systems which cannot be satisfactorily approximated with an A which has distinct roots.[2]

Let ρ be an n-th order nonsingular matrix which diagonalizes A:[3]

$$\rho^{-1}A\rho = \Lambda = \begin{bmatrix} \lambda_1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \lambda_2 & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \lambda_n \end{bmatrix}, \tag{2}$$

Define *normal coordinates* as the components of the n-dimensional state vector y,

$$x = \rho y. \tag{3}$$

Then the *normal form* representation of S is given by

$$\begin{aligned} \dot{y} &= \Lambda y + \beta u \\ v &= \gamma y + Du, \end{aligned} \tag{4}$$

where

$$\beta = \rho^{-1}B, \quad \text{the normal form input matrix,} \tag{5}$$

$$\gamma = C\rho, \quad \text{the normal form output matrix.} \tag{6}$$

The normal coordinates are not unique. If desired, they may be made so by arranging the λ_i in order of increasing magnitude (roots with identical magnitudes may be taken in order of increasing angle) and choosing the column vectors of ρ, ρ_i, $i = 1, \cdots, n$, to have unit Euclidean length.

For the purpose considered here, the system S is *stable* if Re $\lambda_i < 0$ for all i.

The rank of the input r_u is defined as the rank of the matrix B (or equivalently, the rank of β). It is the "effective" number of inputs which can

[2] See Bellman [15, p. 198.]

[3] Familiar results of matrix theory will be used without comment. These results can be found in Bellman [15] or other standard texts.

influence the state vector. The integer $(p - r_u) \geqq 0$ is therefore the number of ineffectual inputs. It is possible with no loss of generality to reduce the number of components of u by $(p - r_u)$.

The rank of the output r_v is defined as the rank of the matrix C (or γ). It is the effective number of outputs available for observing the state of the system. The integer $(q - r_v) \geqq 0$ gives the number of outputs (components of v) which are linearly dependent if $D = 0$. It is possible without loss of generality to reduce the number of columns of C by $(q - r_v)$.[4]

3. Observability and controllability. A system S is *controllable* if β has no rows which are zero. Coordinates y_i corresponding to non-zero rows of β are called controllable; coordinates corresponding to zero rows of β are called uncontrollable. Uncontrollable coordinates can in no way be influenced by the input u. Thus a system which is not controllable has dynamic modes of behavior which depend solely on initial conditions or disturbance inputs. Disturbance inputs are not indicated in (1) and will not be treated in this study. Sometimes, they may be satisfactorily handled by means of appropriately introduced initial conditions.

A system S is *observable* if γ has no columns which are zero. Coordinates y_i corresponding to non-zero columns of γ are called observable; coordinates y_i corresponding to zero columns of γ are called unobservable. Unobservable coordinates are not detectible in the output v. Thus a system which is not observable has dynamic modes of behavior which cannot be ascertained from measurement of the available outputs.

A few general remarks are in order. First, the definition of controllability is different from Kalman's [14]: "A system is controllable if any initial state can be transferred to any desired state in a finite length of time by some control action." However, under the restrictions of the previous section the two definitions are equivalent. More recently, Kalman [16] has taken the same point of view given in this paper. For some additional remarks see the note by Ho [17].

Second, there is a striking similarity in the definitions of controllability and observability, the rows of β playing the same role as the columns of γ. This is also true of Kalman's definitions, and means that remarks similar to those of the previous paragraph can be made about observability. More importantly, for every conclusion concerning controllability, there is a corresponding one concerning observability. This will be evident in the statement and proof of theorems which follow.

Finally, the definitions become more involved when the characteristic

[4] Usually it is desirable to eliminate ineffectual inputs and superfluous columns of C. Exceptions occur when amplitude constraints are imposed on the u_i (such as $|u_i| < k_i$, $i = 1, \cdots p$) or noise is present in the measurement of the v_i.

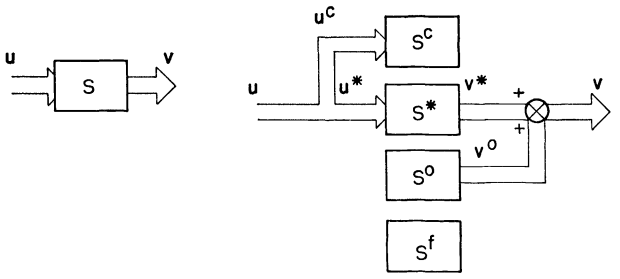

FIG. 1. System S and its partitioned representation

roots are not distinct. The diagonal matrix is replaced by a Jordan normal form and the conditions on β and γ are not so simply stated.

In order to deal more concisely with the above concepts consider:

THEOREM 1. *A system S may always be partitioned into four possible subsystems (shown in Figure 1):*

1) *a system S^* which is controllable and observable and has a transmission matrix D,*
2) *a system S^o each of whose normal coordinates are observable and uncontrollable,*
3) *a system S^c each of whose normal coordinates are controllable and unobservable,*
4) *a system S^f each of whose normal coordinates are uncontrollable and unobservable.*

All subsystems have zero transmission matrices except S^*. Also, $u^* = u^c = u$, $v = v^* + v^o$, and $n = n^* + n^o + n^c + n^f$.

The proof of Theorem 1 follows directly from equations (4) by partitioning y according to the restrictions 1) through 4). A somewhat more involved partitioning may result when the characteristic roots are not distinct.

Thus the only subsystem which has to do with the relationship of v to u is S^*. The observable system S^o only adds a disturbance v^o to the controlled part of the output v^*. Although S^o, S^c and S^f appear to have little importance in system analysis this is not necessarily so. If state variables appropriate to the description of S^o, S^c, and S^f get large, neglected nonlinear couplings may become important or physical damage of the system may result. This certainly will be the case if S^o, S^c or S^f are unstable, i.e. there are hidden instabilities.

From Theorem 1 it is clear that a necessary and sufficient condition for the absence of S^o, S^c, and S^f is that S be controllable and observable. It is possible to determine if S is controllable and observable without recourse to the normal form representation by means of the following theorem.

THEOREM 2. *Let b_i, $i = 1, \cdots, p$, be the columns of B and c_i^T, $i = 1, \cdots$*

q, be the rows of C.[5] *A system S is controllable (observable) if and only if the vectors $e_{ki} = A^k b_i$, $i = 1, \cdots, p$, $k = 0, \cdots, n-1 (e_{ki} = (A^T)^k (c_i)$, $i = 1, \cdots, q$, $k = 0, \cdots, n-1)$ span the n-dimensional coordinate space.*

The controllability part of this theorem has been proved using the previously mentioned alternative definition of controllability [14]. By duality [13, 16] the observability part may be obtained for an alternative definition of observability [13]. The fact that the same results are obtained for the different definitions proves their equivalence.

Proof. First consider the controllability part of the theorem.

To prove necessity assume S is controllable and write

$$e_{ki} = A^k b_i = (\rho\Lambda\rho^{-1})^k b_i = \rho\Lambda^k\rho^{-1} b_i = \rho\Lambda^k\beta_i . \tag{7}$$

Since $\beta = [\beta_1 \cdots \beta_p]$ has no zero row it is possible to form a vector $\beta^+ = k_1\beta_1 + \cdots + k_p\beta_p$ none of whose components is zero. Clearly, the vectors $e_k{}^+ = \rho\Lambda^k\beta^+$, $k = 0, \cdots, n-1$ form a subspace of the space defined by the e_{ki}. But

$$\begin{aligned} &\det [e_0{}^+ \cdots e_{n-1}{}^+] \\ &= \det \left\{ \rho \begin{bmatrix} \beta_1{}^+ & 0 & 0 & \cdots & 0 \\ 0 & \beta_2{}^+ & 0 & & 0 \\ 0 & 0 & & & \vdots \\ \vdots & & & & \\ 0 & \cdots & \cdots & & \beta_n{}^+ \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{bmatrix} \right\} \\ &= (\det \rho)(\det V)(\beta_1{}^+\beta_2{}^+ \cdots \beta_n{}^+) \neq 0 \end{aligned} \tag{8}$$

because the Vandermonde determinant V is nonzero for distinct λ_i, ρ is nonsingular, and the $\beta_i{}^+$ are all nonzero. Thus the subspace is n-dimensional. Therefore the e_{ki} must span the n-dimensional space.

To prove sufficiency assume the e_{ki} span the n-dimensional space. Then for any $r \neq 0$, say

$$r = (\rho^T)^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

the inner product (r, e_{ki}) cannot be zero for all k and i. But

$$(r, e_{ki}) = (r, \rho\Lambda^k\beta_i) = (\Lambda^k\rho^T r, \beta_i) = \lambda_1{}^k\beta_{1i} . \tag{9}$$

Assume all $\beta_{1i} = 0$ and a contradiction is obtained. Thus not all $\beta_{1i} = 0$. By

[5] The superscript T indicates the transpose of a matrix or vector.

changing r the argument also works on all other rows of β. Hence S is controllable.

To prove the observability part of the Theorem note (γ_i^T is i-th row of γ)

$$
\begin{aligned}
e_{ki} = (A^T)^k c_i &= \{(\rho\Lambda\rho^{-1})^T\}^k(\gamma_i^T\rho^{-1})^T \\
&= (\rho^{-1})^T\Lambda^k\rho^T(\rho^{-1})^T\gamma_i \\
&= (\rho^{-1})^T\Lambda^k\gamma_i\,.
\end{aligned} \tag{10}
$$

Since (10) is similar to (7) the remaining steps are the same as those in the controllability part.

4. Observability and controllability of composite systems. In this section the controllability and observability of composite systems are related to the controllability and observability of their subsystems. Theorems 3 and 4 treat respectively the parallel and cascade connection of two subsystems. Successive application of these theorems extends the result to composite systems which consist of many subsystems connected in parallel and cascade. Theorem 5 is the central theorem of the paper. It states conditions for the controllability and observability of a general feedback system.

THEOREM 3. *Let the parallel connection of systems S_a and S_b form a composite system S (see Figure 2). Then:*

i) $n = n_a + n_b$;

ii) $\lambda_1, \cdots \lambda_n = \lambda_{1a}, \cdots, \lambda_{n_a a}, \lambda_{1b}, \cdots, \lambda_{n_b b}$;

iii) *a necessary and sufficient condition that S be controllable (observable) is that both S_a and S_b be controllable (observable).*

To prove Theorem 3 let S_a and S_b be represented in normal form. Then

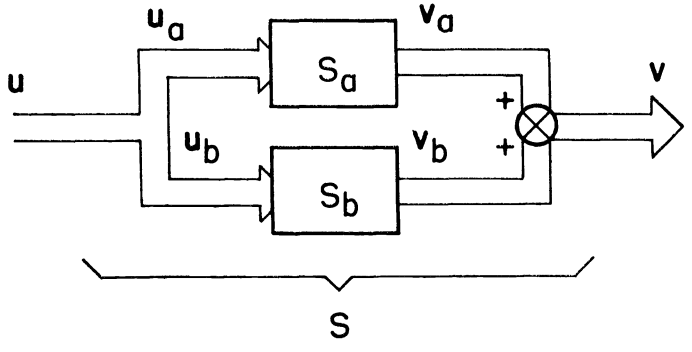

FIG. 2. Parallel connection of S_a and S_b

from the notation in Figure 2 the normal form of S can be chosen so that

$$(11)\qquad y = \begin{bmatrix} y_a \\ y_b \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \Lambda_a & 0 \\ 0 & \Lambda_b \end{bmatrix},$$
$$\beta = \begin{bmatrix} \beta_a \\ \beta_b \end{bmatrix}, \qquad \gamma = [\gamma_a \gamma_b], \qquad D = D_a + D_b .$$

Simple inspection of (11) yields all parts of the theorem.

THEOREM 4. *Let the cascade connection of system S_a followed by S_b form a composite system S (see Figure 3). Then:*

i) $n = n_a + n_b$;

ii) $\lambda_1 , \cdots , \lambda_n = \lambda_{1a} , \cdots , \lambda_{n_a a} , \lambda_{1b} , \cdots \lambda_{n_b b}$;

iii) *a necessary (but insufficient) condition for the controllability (observability) of S is that both S_a and S_b be controllable (observable);*

iv) *if S_a and S_b are both controllable (observable) any uncontrollable (unobservable) coordinates of S must originate, when designated according to characteristic root, in $S_b(S_a)$.*

Using the normal form representations of S_a and S_b yields

$$(12)\qquad \dot{x} = \begin{bmatrix} \Lambda_a & 0 \\ \beta_b \gamma_a & \Lambda_b \end{bmatrix} x + \begin{bmatrix} \beta_a \\ \beta_b D_a \end{bmatrix} u, \quad \text{where } x = \begin{bmatrix} y_a \\ y_b \end{bmatrix},$$
$$v = [D_b \gamma_a \ \gamma_b] x + D_b D_a u.$$

as the set of equations representing S.

To put these equations in normal form define

$$(13)\qquad x = \begin{bmatrix} I & 0 \\ \phi & I \end{bmatrix} y, \qquad y = \begin{bmatrix} I & 0 \\ -\phi & I \end{bmatrix} x,$$

where $-\phi\Lambda_a + \Lambda_b\phi = -\beta_b\gamma_a$, i.e.,

$$(14)\qquad [\phi_{ij}] = \frac{(\beta_b\gamma_a)_{ij}}{\lambda_{ia} - \lambda_{jb}}.$$

$(\beta_b\gamma_a)_{ij}$ denotes the ij element of $\beta_b\gamma_a$. The assumption of distinct roots

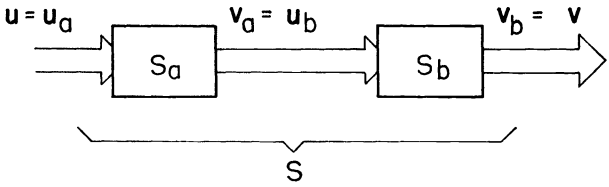

FIG. 3. Cascade connection of S_a followed by S_b

requires $\lambda_{ja} - \lambda_{ib} \neq 0$ all i and j. It is easily shown that

$$\begin{aligned} \dot{y} &= \Lambda y + \beta u \\ v &= \gamma y + Du\,, \end{aligned} \tag{15}$$

where

$$\Lambda = \begin{bmatrix} \Lambda_a & 0 \\ 0 & \Lambda_b \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_a \\ (-\phi\beta_a + \beta_b D_a) \end{bmatrix}$$
$$\gamma = [(D_b\gamma_a + \gamma_b\phi)\gamma_b], \qquad D = D_a D_b\,. \tag{16}$$

Results i) and ii) follow immediately from inspection of (15). Consider the controllability parts of results iii) and iv). From (14) and (16) it is obvious that a null row of β_a or β_b will result in a null row of β. Thus the necessity of iii) follows. It is also clear that $-\phi\beta_a + \beta_b D_a$ may have a null row even if β_a and β_b do not. Thus iv) and the remainder of iii) hold. Corresponding reasoning applied to the columns of γ yields the observability results.

Formulas (16) can be used to determine if S is controllable or observable. Unfortunately, a fair amount of work is involved and there appears to be no way of getting simpler sufficient conditions for the controllability or observability of S.

It is helpful to consider a few simple examples where S is uncontrollable or unobservable even though S_a and S_b are controllable and observable. Let S_a and S_b be given by:

$$\begin{aligned} &\dot{y}_{1a} = -y_{1a} + u_1 \qquad \dot{y}_{1b} = -2y_{1b} + u_{1b} - u_{2b} = -2y_{1b} + v_{1a} - v_{2a} \\ &v_{1a} = y_{1a}\,, \qquad\qquad v_1 = y_{1b} \\ &v_{2a} = y_{1a}\,. \end{aligned} \tag{17}$$

Then if $x_1 = y_{1a}$ and $x_2 = y_{1b}$ define the state vector of S,

$$\begin{aligned} \dot{x} &= \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_1 \\ v_1 &= [0 \quad 1]x. \end{aligned} \tag{18}$$

In this example S is uncontrollable and unobservable because the matrices, $D_a\,, D_b\,, \gamma_a\,, \beta_b\,$, which "couple" S_a and S_b are such ($D_a = D_b = 0, \beta_b\gamma_a = 0$) that the input u_1 never reaches the normal coordinate of S_b and the normal coordinate of S_a is not passed on to the output v_1. This particular situation cannot happen in single-input, single-output systems, since it would imply either $\gamma_a = 0$ or $\beta_b = 0$.

For the second example let

$$\begin{aligned} &\dot{y}_{1a} = -y_{1a} + u_1 \qquad \dot{y}_{1b} = -2y_{1b} + u_{1b} = -2y_{1b} + v_{1a} \\ &v_{1a} = y_{1a} + u_1 , \qquad v_1 = y_{1b} - u_{1b} . \end{aligned} \tag{19}$$

Taking the state vector of S as $x_1 = y_{1a}$, $x_2 = y_{1b}$ gives

$$\begin{aligned} \dot{x} &= \begin{bmatrix} -1 & 0 \\ 1 & -2 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u_1 , \\ v_1 &= [1 \quad -1]x - u_1 , \end{aligned} \tag{20}$$

and for

$$\rho = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \tag{21}$$

the normal form representation is

$$\begin{aligned} y &= \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} y + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_1 \\ v_1 &= [0 \quad -1]y - u_1 . \end{aligned} \tag{22}$$

Equation (20) shows that $x_1 = y_{1a}$ and $x_2 = y_{1b}$ can individually be controlled and observed. Yet from equation (22), S is clearly uncontrollable and unobservable. This apparent paradox is resolved by observing that the uncontrolled (and therefore unalterable) coordinate $y_2 = -x_1 + x_2 = -y_{1a} + y_{1b}$. Therefore y_{1a} and y_{1b} cannot *independently* be controlled or observed.

A third example arises, applicable to the *parallel* connection of S_a and S_b, if the assumption in section 2 of distinct characteristic roots is waived. Then iii) of Theorem 3 becomes analogous to iii) of Theorem 4, in that the stated condition is necessary but not sufficient. Let S_a and S_b be identical first order systems

$$\dot{y}_{1c} = -y_{1c} + u_{1c} , \qquad c = a, b. \tag{23}$$

Then S is given by

$$\dot{x} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u. \tag{24}$$

While $y_{1a} = x_1$ and $y_{1b} = x_2$ are controllable, they are not independently controllable, since their difference is given by the solution of

$$(\dot{x}_1 - \dot{x}_2) = -(x_1 - x_2). \tag{25}$$

THEOREM 5. *Systems S_a and S_b form respectively the forward and return paths of a feedback system S (see Figure 4). Let the cascade connection of S_a*

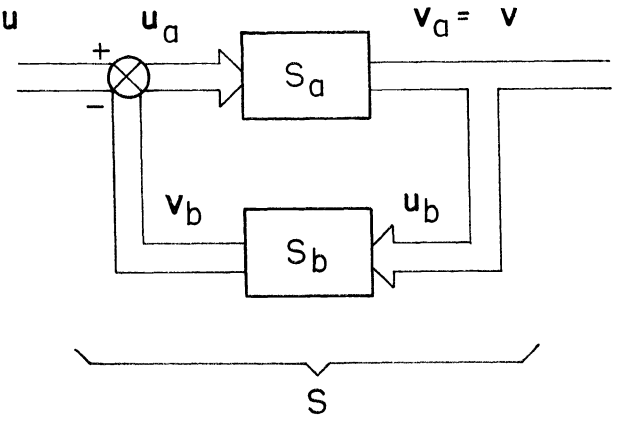

FIG. 4. Feedback system S with S_a in the forward path and S_b in the return path

followed by S_b *be* S_c *and of* S_b *followed by* S_a *be* S_o. *Assume that* $(I + D_aD_b)$ *is nonsingular. Then*:

i) $n = n_a + n_b$,

ii) *a necessary and sufficient condition that* S *be controllable* (*observable*) *is that* $S_c(S_o)$ *be controllable* (*observable*),

iii) *a necessary but not sufficient condition that* S *be controllable* (*observable*) *is that both* S_a *and* S_b *be controllable* (*observable*),

iv) *if* S_a *and* S_b *are both controllable* (*observable*) *any uncontrollable* (*unobservable*) *coordinates of* S *are uncontrollable* (*unobservable*) *coordinates of* $S_c(S_o)$ *and originate in* S_b.

Before going on with the proof, a few general observations are made. The nonsingularity of $(I + D_aD_b)$, which is equivalent to the nonsingularity of $(I + D_bD_a)$, is physically reasonable, for if it is broken the static gain $D = (I + D_aD_b)^{-1}D_a = D_a(I + D_bD_a)^{-1}$ of the closed-loop system S is undefined. Introduction of systems S_c and S_o is a natural consequence of proving separately the controllability and observability parts of the theorem. Since controllability involves only the influence of the input u on S, the system shown in Figure 5a suffices. Similarly, determination of observability leads to the system of Figure 5b. Statements analogous to ii) of Theorems 3 and 4 are not possible, since feedback alters characteristic roots.

By employing

$$\begin{aligned} u_a &= u - v_b \\ v &= v_a = u_b \end{aligned} \tag{26}$$

and the equations describing S_a and S_b, the equations describing S are obtained. Inspection of these equations shows i) is true; however, they are too complex to yield a simple proof of ii).

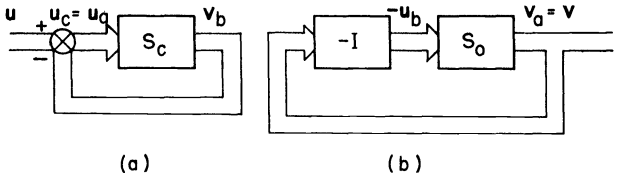

FIG. 5. Systems for determining the controllability (a) and observability (b) of a feedback system.

Consider first the controllability part of ii). From Figure 5a

$$\begin{aligned} v_b &= v_c \\ u_c &= u - v_c . \end{aligned} \tag{27}$$

Using these equations and the normal form equations for S_c gives ($x = y_c$)

$$\dot{x} = Ax + Bu \tag{28}$$

for S, where

$$A = \Lambda_c - B\gamma_c \tag{29}$$

$$B = \beta_c(I + D_bD_a)^{-1}. \tag{30}$$

It is easily shown from the nonsingularity of $(I + D_bD_a)^{-1}$ that a row of B will be zero if and only if the corresponding row of β_c is zero. Thus B has non-zero rows if and only if S_c is controllable.

The sufficiency part of ii) is proved by contradiction. Let S_c be controllable and assume that S is uncontrollable. Then from Theorem 2 the vectors e_{ki}, $k = 0, \cdots, n-1, i = 1, \cdots, p$ cannot span the n-dimensional space. That is, a non-zero vector r exists such that

$$\begin{aligned} (r, e_{ki}) &= 0, \\ k &= 0, \cdots, n-1, i = 1, \cdots, p. \end{aligned} \tag{31}$$

Or equivalently,

$$r^TA^kB = r^T(\Lambda_c - B\gamma_c)^kB = 0, \quad k = 0, \cdots, n-1. \tag{32}$$

Evaluating (32) starting with $k = 0$ gives

$$\begin{aligned} &r^TB = 0 \\ &r^T(\Lambda_c - B\gamma_c)B = r^T\Lambda_cB - (r^TB)\gamma_cB = r^T\Lambda_cB = 0 \\ &\quad\vdots \qquad\qquad\qquad\qquad \vdots \\ &r^T(\Lambda_c - B\gamma_c)^{n-1}B = r^T\Lambda_c^{n-1}B = 0. \end{aligned} \tag{33}$$

From Theorem 2 it can be seen that the columns of the matrices Λ_c^kB,

$k = 0, \cdots, n-1$ span the n-dimensional space if and only if B has no zero row. Since by the previous paragraph B has no zero row, this and (33) imply that r is zero. Thus the contradiction is obtained.

The necessity part of ii) is obvious from the discussion at the end of section 4.

The observability part of ii) is proved by starting with Figure 5*b* and S_o. Then S is given by

$$\begin{aligned} \dot{x} &= Ax \\ v &= Cx, \end{aligned} \qquad A = \Lambda_o - \beta_o C, \qquad C = (I + D_a D_b)^{-1}\gamma_0 . \tag{34}$$

The above steps can then be applied to the columns of C with the desired results.

Theorem 4 applied to the determination of S_c and S_o gives iii).

Consider the controllability part of iv). From (28), (29), and (30) it can be seen that if the i-th row of β_c is zero, $\dot{x}_i = \lambda_i x_i$. Thus the uncontrollable coordinate $y_{ic} = x_i$ is unchanged by feedback. Moreover, by Theorem 4 this coordinate must originate in S_b. Similar arguments give the observability part of iv).

The most important result of Theorem 5 is ii). It says that closed-loop controllability and observability can be ascertained from the open-loop systems S_c and S_o. Thus one is not forced to deal with intricate closed-loop equations.

When S_b is static an even simpler situation exists. Then iv) implies that S is controllable and observable if S_a is controllable and observable.

Further information on uncontrollable and unobservable coordinates can be gleaned from Theorems 3, 4, and 5. Let S^u denote the combination of systems S^o, S^c, and S^f, that is, the part of system S which is not controllable and observable. From Theorem 1 it is clear that the coordinates of S_a^u, S_b^u, $\cdots$ are uncontrollable or unobservable in the composite system S. Thus S_a^u, S_b^u, $\cdots$ are part of S^u. To see what happens to the remaining coordinates of S_a, S_b, $\cdots$ it is sufficient to examine by means of Theorems 3, 4 and 5 the interconnection of the controllable and observable system S_a^*, S_b^*, $\cdots$. As an example take the feedback system of Theorem 5: S^u consists of S_a^u and S_b^u plus the coordinates of S_b^* which are uncontrollable in the system S_a^* followed by S_b^* and unobservable in the system S_b^* followed by S_a^*.

5. The transfer-function matrix. The traditional approach to the analysis and synthesis of multivariable systems is based on the transfer-function matrix rather than the differential equations (1). To obtain a transfer-function representation of a system S, it is assumed that the output vector v

is entirely due to input forcing u, i.e., initial conditions are zero. Let Laplace transforms be denoted by upper-case letters.
Then

$$V(s) = H(s)U(s) \tag{35}$$

where s is the Laplace-transform variable and $H(s) = [H_{ij}]$ is the q by p transfer-function matrix. The element $H_{ij}(s)$ is the scalar transfer function which relates the i-th output and the j-th input.

To obtain the transfer-function matrix from the differential-equation representation consider:

THEOREM 6. *Given a system S defined by equations* (1) *and* (4), *the transfer-function matrix is*

$$\begin{aligned} H(s) &= C(Is - A)^{-1}B + D = \gamma(Is - \Lambda)^{-1}\beta + D \\ &= \sum_{i=1}^{n^*} \frac{K_i}{s - \lambda_i^*} + D \end{aligned} \tag{36}$$

where the matrices K_i have rank one.

The first two expressions for H follow directly from the Laplace transform of (1) and (4) with $x(0) = y(0) = 0$. Since $(Is - \Lambda)^{-1}$ is diagonal, the second expression can be written out in terms of the columns of γ, γ_i, and the rows of β, ${\beta_i}^T$:

$$H = \sum_{i=1}^{n} \frac{\gamma_i {\beta_i}^T}{s - \lambda_i} + D. \tag{37}$$

For any i corresponding to an uncontrollable or unobservable coordinate, γ_i or ${\beta_i}^T$ is zero. Thus the sum needs to be taken only over the characteristic roots associated with S^*. $K_i = \gamma_i^* {\beta_i^*}^T$, being a vector outer product, is of rank one.

The important, and not surprising, conclusion of Theorem 6 is that a transfer-function matrix represents the controllable and observable part of S, S^*. It has been noted in Theorems 4 and 5 that controllability and observability of subsystems does not assure the controllability and observability of a composite system. Thus transfer-function matrices may satisfactorily represent all the dynamic modes of the subsystems but fail to represent all those of the composite system. Furthermore, the loss of hidden response modes is not easy to detect because of the complexity of the transfer-function matrices and matrix algebra. Since differential equations offer a safer basis for describing multivariable systems it is valid to ask why transfer-function matrices should be used at all. The answer is that frequency domain design procedures and the smaller size of H (it is $q \times p$ rather than $n \times n$) often make computations more manageable.

If the transfer-function matrix of a physical system is given it is generally impossible to derive the corresponding differential-equation representation. This is because the state variable choice is not unique and all information concerning systems S^c, S^o, and S^f is missing. It is possible, however, to find a set of differential equations (1) or (4) which yield the same $H(s)$ as a prescribed $H(s)$. Procedures for doing this are described below. The main result is stated here as a theorem and gives the required order of the differential equations.

THEOREM 7. *Given a rational transfer-function matrix $H(s)$ whose elements have a finite number of simple poles at $s = \lambda_i$, $i = 1, \cdots, m$ in the finite s-plane. Let the partial fraction expansion of H be*

$$H(s) = \sum_{i=1}^{m} \frac{K_i}{s - \lambda_i} + D, \tag{38}$$

where

$$K_i = \lim_{s \to \lambda_i} (s - \lambda_i) H(s), \tag{39}$$

$$D = \lim_{s \to \infty} H(s). \tag{40}$$

Let the rank of the i-th pole, r_i, be defined as the rank of K_i. Then $H(s)$ can be represented by differential equations (1) or (4) whose order is

$$n = \sum_{i=1}^{m} r_i. \tag{41}$$

The eigenvalues of A and Λ are distinct if and only if all $r_i = 1$. It is impossible to represent $H(s)$ by a differential equation whose order is less than n.

First it will be shown how $H(s)$ can be represented by a set of differential equations.

Since the matrix K_i is of rank r_i there are r_i linearly independent columns in K_i. Let e_{ji}, $j = 1, \cdots, r_i$ be such a set of columns. Then every column of K_i can be expressed as a linear combination of the e_{ji}. A compact notation is

$$K_i = E_i F_i \tag{42}$$

where E_i is a $q \times r_i$ matrix which has columns e_{ji}. To determine F_i premultiply (42) by E^T. Then

$$E_i^T K_i = E_i^T E_i F_i. \tag{43}$$

But the determinant of $E_i^T E_i$ is the Gram determinant [19] of the e_{ji}, and is nonzero because the e_{ji} are linearly independent (this is a good test for

picking a linearly independent set e_{ji}). Thus

$$F_i = (E_i^T E_i)^{-1} E_i^T K_i . \tag{44}$$

Once F_i is known K_i can be expressed as

$$K_i = \sum_{j=1}^{r_i} e_{ji} f_{ij}^T \tag{45}$$

where f_{ij}^T is the j-th row of F_i. Thus

$$H(s) = \sum_{i=1}^{m} \sum_{j=1}^{r_i} \frac{e_{ji} f_{ji}^T}{s - \lambda_i} + D. \tag{46}$$

This formula is similar to (37) except that there are r_i vector outer products for each λ_i. Thus $H(s)$ can be represented by (4) where

$$\Lambda = \begin{bmatrix} \lambda_1 I_1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2 I_2 & & & \vdots \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \lambda_m I_m \end{bmatrix}, \tag{47}$$

$$\beta = \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix}, \qquad \gamma = [E_1 \cdots E_m]$$

and I_i is an identity matrix of order r_i. Thus the root λ_i is of multiplicity r_i and $n = \sum_{i=1}^{m} r_i$.

To show that a realization of lower order is not possible, the Laplace transform of (4) is taken, defining a transfer-function matrix $\bar{H}$. However, to cover all possibilities it is essential that Λ take its most general form, the Jordan normal form. For a characteristic root of multiplicity ℓ, this means that the number of Jordan blocks with this characteristic root is not fixed, only that all the blocks taken together form a matrix of order ℓ. However, $\bar{H}$ shows that all Jordan blocks must be of order one if $\bar{H}$ is to have simple poles (unless some modes are uncontrollable or unobservable, which only increases the order of (4)). Furthermore, the rank of the residue of $\bar{H}$ at $s = \lambda_i$ is no greater than the multiplicity of λ_i. Thus if $\bar{H}$ is to have the form of H in (38) the differential equations (4) must have a minimum order

$$n = \sum_{i=1}^{m} r_i .$$

If the equations (4) have order greater than n, the realization is either uncontrollable, unobservable, or both.

Theorem 7 provides a solution of the synthesis problem, since once the

differential equations (1) or (4) are known, they can be realized as a physical system (example, an electronic differential analyzer). Furthermore, the synthesized system uses a minimum number of dynamic elements.[6] The assumption of simple poles can be relaxed, but at the expense of considerable additional complexity. Kalman [16] gives an alternative procedure for determining n.

Theorem 7 and a simple example illustrate how the order of a system represented by a transfer-function matrix may be underestimated. Let

$$H(s) = \begin{bmatrix} \frac{1}{s+1} & \frac{2}{s+1} \\ \frac{-1}{(s+1)(s+2)} & \frac{1}{s+2} \end{bmatrix} = \frac{\begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}}{s+1} + \frac{\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}}{s+2}. \tag{48}$$

At first glance it might be guessed that the system has order two, but $r_1 + r_2 = 2 + 1 = 3$, so the minimum order is three. One realization of an equivalent third order system is shown in Figure 6. It is possible that the actual order of the system may be greater than three. For example, in Figure 7 the order is five.

Underestimation of system order is the reason why most erroneous stability analyses have gone unnoticed. In a stability analysis the number of characteristic roots considered should at least be equal to the sum of the minimum orders of all the subsystems. This is easily checked by means of Theorem 7—and errors in many references gave been noted.[7]

If a transfer function matrix has any poles of rank greater than one, the assumption of distinct characteristic roots, which was made in all prior developments, is violated. If such transfer functions are encountered, an approximating system may be set up (use approximation to equations (47)) which has poles of rank one. Then all the previous results can be used.

From the above discussion it is clear that each element of $H(s)$ is an integral part of the whole description. Thus it is generally not permissible to partition a transfer-function matrix into several transfer function matrices and treat the resulting matrices as though they describe distinct systems. Yet, this has been done consistently in the representation of plants which have more inputs than outputs [9, 12]. As a consequence erroneous statements have been made concerning the existence of stable feedback systems.[8]

[6] McMillan [18] defines the degree of a square ration al matrix, which is equivalent to n, but the development is more complicated being based on the Smith normal form of a polynomial matrix. He also shows that if the m atrix is an impedance matrix, it may be synthesized by a passive network with n, and no fewer, reactive elements.

[7] See for example [10, 11].

[8] This has been noted by the author in a discussion [12].

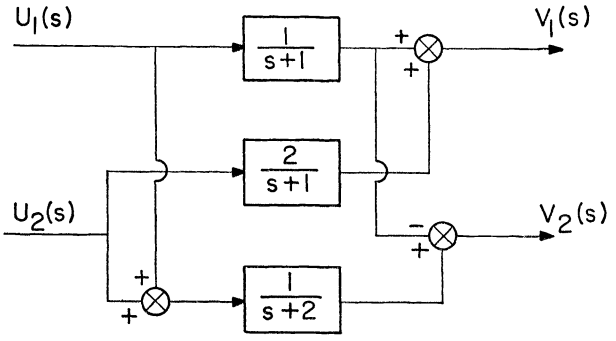

FIG. 6. Third order representation of $H(s)$.

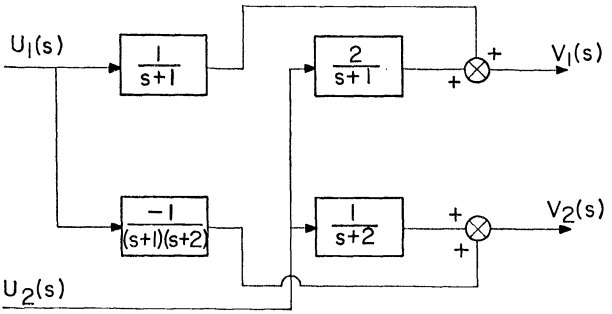

FIG. 7. Fifth order representation of $H(s)$

6. Transfer-function representation of multivariable feedback systems. Once the limitations of transfer-function matrices are recognized, it is possible to apply them successfully to the analysis and synthesis of feedback systems. In what follows it will be assumed that all transfer-function matrices have simple poles of rank one. This will keep the transfer-function representations consistent with the differential-equation representations specified earlier.

Let H_a and H_b be transfer-function matrices representing S_a and S_b in Figure 4. Then the developments,

$$U_a = U - V_b = U - H_bV = U - H_bH_aU_a \tag{49}$$
$$= (I + H_bH_a)^{-1}U,$$

$$V = H_aU_a = H_a(I + H_bH_a)^{-1}U, \tag{50}$$

and

$$V = H_a(U - V_b) = H_aU - H_aH_bV \tag{51}$$

$$= (I + H_aH_b)^{-1}H_aU, \tag{52}$$

give alternative expressions for the transfer-function of the feedback system S,

$$H = H_a(I + H_bH_a)^{-1} = (I + H_aH_b)^{-1}H_a . \tag{53}$$

H represents the controllable and observable part of S, S^*.

The remaining part of S, S^u, was considered at the end of section 4. Systems S_a^u and S_b^u naturally are missing from the representation H because they are not represented in H_a and H_b. The coordinates of S_b^* which are not controllable and/or observable in S correspond to the poles of H_b which do not appear in H_bH_a and/or H_aH_b. In the derivation for H it is easy to see where the poles of H_b are lost: (49) gives those of H_bH_a and (51) gives those of H_aH_b. It is not so easy to see that no additional poles are lost, a difficulty which has to do with complexities in evaluating the inverse of a matrix of rational functions. This is one of the reasons that led to the more careful treatment of section 4.

Suppose that all the subsystems which make up H_a and H_b are controllable and observable. This is a reasonable assumption if transfer-function matrices are to be used. Then from the preceding it is plain that the characteristic roots of the feedback system are given by: 1) the poles of H (these roots correspond to the dynamic modes in S which are controllable and observable), 2) the poles of H_b which do not appear as poles of H_aH_b and/or H_bH_a, 3) the poles of the transfer functions representing the subsystems of S_a and S_b which do not appear respectively in H_a and H_b.

In the course of system synthesis and stability analysis *all* characteristic roots of the feedback system must be considered. Procedures for handling the characteristic roots in category 1) have been developed reasonably well in the literature. Therefore, additional effort here will be directed at 2) and 3). In particular, the problem of pole cancellation in multiplying two transfer-function matrices will be explored. This problem applies directly to 2), and often to 3), since the systems S_a and S_b are usually a cascade connection of subsystems.[9] If S_a or S_b are themselves feedback systems they must first be analyzed as feedback systems before progress can be made on the analysis of the overall feedback system.

7. Pole cancellation. Consider the cascade connection of the controllable and observable systems S_a and S_b (not the S_a and S_b of the previous section, see Figure 3). The transfer-function representation gives

$$H = H_bH_a . \tag{54}$$

If H has fewer poles than the sum of poles in H_a and H_b, pole cancellation

[9] The special case where S_b is static and S_a is the cascade connection of two subsystems has been considered in [9, 12]. The results obtained are not as general as those of the next section.

has occurred and the system S is uncontrollable or unobservable. To go further, a more detailed notation is required.

Let H_a be written as

$$H_a = \frac{\mathcal{H}_a}{h_a}, \tag{55}$$

where h_a is the characteristic polynomial for S_a,

$$h_a = k_a(s - \lambda_{1a}) \cdots (s - \lambda_{n_a a}), \qquad k_a \neq 0. \tag{56}$$

Since it has been assumed that H_a has simple poles of rank one, h_a has no repeated linear factors.[10] The elements of the matrix $\mathcal{H}_a$ are polynomials in s. Such a matrix is said to have a factor, if every element of the matrix has the same factor. Since S_a is controllable and observable $\mathcal{H}_a$ has no factors common with h_a. Similar remarks apply to H_b.

Using the notation

$$\mathcal{H} = \mathcal{H}_b\mathcal{H}_a \tag{57}$$

and

$$h = h_b h_a, \tag{58}$$

system S is controllable and observable if h and $\mathcal{H}$ do not have common factors. Any linear factor of h cancelled in H by a like factor of $\mathcal{H}$ corresponds to an uncontrollable or unobservable mode in S. Unless the elements of $\mathcal{H}_a$ and $\mathcal{H}_b$ are in some way related, the possibility that h and $\mathcal{H}$ will have common factors is remote.

The most common situation which causes $\mathcal{H}_a$ and $\mathcal{H}_b$ to be related is that of compensation where either H_a or H_b is fixed, and the other (the compensator) is chosen to make H equal a desired transfer-function matrix H_d. Clearly, if h_d does not equal $h_b h_a$ the compensated system S will be uncontrollable or unobservable. Thus certain constraints must be imposed on H_d if S is to be controllable and observable. Often it is sufficient to require that only the unstable modes of S_a and S_b be controllable and observable in S. This reduces the number of constraints.

The following treatment of constraints assumes pre-compensation (H_b fixed) and H_b square ($p_b = q_b$). The assumption that S_a is controllable and observable is reasonable because a minimum order realization of H_a must be controllable and observable. Also, it is pointless to consider an H_d which corresponds to an uncontrollable or unobservable system.

Formally, compensation requires

$$H_a = H_b^{-1}H_d. \tag{59}$$

[10] Here the term factor means a non-constant factor.

Therefore $|H_b|$ must not be identically zero. This is assured if B_b and C_b have rank p_b. Expansion of (59) yields

$$H_a = \frac{[\text{adj } H_b]\, H_d}{|H_b|} = \frac{h_b\,[\text{adj } \mathcal{C}_b]\, \mathcal{C}_d}{|\mathcal{C}_b|\, h_d}. \tag{60}$$

Let the greatest common divisor of the numerator and denominator be g. Then

$$\mathcal{C}_a g = h_b\,[\text{adj } \mathcal{C}_b]\, \mathcal{C}_d \tag{61}$$

and

$$h_a g = h_d\,|\,\mathcal{C}_b\,| \tag{62}$$

because $\mathcal{C}_a$ and h_a cannot have a common factor if S_a is to be controllable and observable.

From (57), (58), (61), and (62)

$$\mathcal{C} = h_b \mathcal{C}_b\,[\text{adj } \mathcal{C}_b]\, \mathcal{C}_d g^{-1} = h_b\,|\,\mathcal{C}_b\,|\,\mathcal{C}_d g^{-1} \tag{63}$$

and

$$h = h_b h_d\,|\,\mathcal{C}_b\,|\,g^{-1}. \tag{64}$$

Since $h_b\,|\,\mathcal{C}_b\,|\,g^{-1}$ is the only factor common to both h and $\mathcal{C}$ ($\mathcal{C}_d$ and h_d do not have common factors) its linear factors give all the modes which are uncontrollable and unobservable in S. Suppose all unstable modes of S are to be controllable and observable. Then all linear factors of $h_b\,|\,\mathcal{C}_b\,|$ which go to zero in the right-half s-plane must be included in g, or equivalently as common factors of $h_d\,|\,\mathcal{C}_b\,|$ and $h_b\,[\text{adj } \mathcal{C}_b]\,\mathcal{C}_d$. This happens only if 1) h_d includes the right-half-plane factors of h_b, and 2) $[\text{adj } \mathcal{C}_b]\,\mathcal{C}_d$ includes the right-half-plane factors of $|\,\mathcal{C}_b\,|$.

Very often the constraints cannot be imposed as indicated. Consider the example

$$\mathcal{C}_b = \begin{bmatrix} s & 1 \\ 1 & s \end{bmatrix}, \qquad h_b = (s+1)(s-1). \tag{65}$$

Both poles have rank one. Constraint 1) requires h_d to have the factor $(s-1)$. Suppose 2) is to be satisfied with $\mathcal{C}_d$ diagonal. Then

$$[\text{adj } \mathcal{C}_b]\,\mathcal{C}_d = \begin{bmatrix} s & -1 \\ -1 & s \end{bmatrix}\begin{bmatrix} k_{11d} & 0 \\ 0 & k_{22d} \end{bmatrix} = \begin{bmatrix} sk_{11d} & -k_{22d} \\ -k_{11d} & sk_{22d} \end{bmatrix}. \tag{66}$$

Since $|\,\mathcal{C}_b\,| = (s^2 - 1)$, each element of (66) must include the factor $(s-1)$, which in this case means both k_{11d} and k_{22d} have the factor $(s-1)$. But $(s-1)$ cannot be a common factor of $\mathcal{C}_d$ and h_d. The same problem also occurs if $k_{11d} = k_{22d} = 0$. As will be seen shortly, the difficulty can be

resolved only by letting H_d have a pole of rank two at $s = 1$. This is the same constraint which would result from procedures described in the literature [9, 12]. It has not been noted previously that it may be relaxed if H_d is not diagonal, a fact which is of interest, since present design procedures are based on diagonalization of the open-loop transfer-function matrix.

The above analyses cannot be extended readily when H_a, H_b, or H_d have poles of rank greater than one, because then common factors in the numerator and denominator of H_a, H_b, and H_d do not necessarily imply that the systems are uncontrollable or unobservable. Theorem 7 offers a satisfactory alternative approach. System S is controllable and observable if the order of S as determined from $H = H_d$ is equal to the order of S_a plus the order of S_b. To simplify the application of this statement the following assumptions are made: i) H_a, H_b, and H_d all have simple poles, ii) H_b has poles of rank one, iii) S_a and S_b are controllable and observable, iv) H_d, is diagonal. Then S is controllable and observable if and only if

$$\text{rank}\,[\lim_{s\to\lambda}(s-\lambda)H_a] + \text{rank}\,[\lim_{s\to\lambda}(s-\lambda)H_b] = \text{rank}\,[\lim_{s\to\lambda}(s-\lambda)H_d], \tag{67}$$
$$\lambda = \lambda_{1a}, \cdots, \lambda_{n_a a}, \lambda_{1b}, \cdots, \lambda_{n_b b}.$$

Define

$$G = H_b^{-1} = [g_1 \cdots g_p] \tag{68}$$

and let $\ell_i = 0$ if h_{iid} is analytic at $s = \lambda$, $\ell_i = 1$ if h_{iid} has a simple pole at $s = \lambda$. Using (59) and ii), (67) can be written as

$$\text{rank}\,[\lim_{s\to\lambda}\{(s-\lambda)h_{11d}g_1 \cdots (s-\lambda)h_{ppd}g_p\}]$$
$$= \sum_{i=1}^{p}\ell_i - 1, \qquad \lambda = \lambda_{1b}, \cdots, \lambda_{n_b b} \tag{69}$$
$$= \sum_{i=1}^{p}\ell_i, \qquad \lambda = \lambda_{1a}, \cdots, \lambda_{n_a a}, \quad \lambda \neq \lambda_{1b}, \cdots, \lambda_{n_b b}.$$

Once the g_i are computed, constraints on the h_{iid} such that (69) is satisfied are easily found. For example, with H_b as defined by (65),

$$g_1 = \begin{bmatrix} s \\ -1 \end{bmatrix}, \qquad g_2 = \begin{bmatrix} -1 \\ s \end{bmatrix}. \tag{70}$$

Consider $\lambda = 1$. Clearly, (69) holds only if $\ell_1 = \ell_2 = 1$. Thus both h_{11d} and h_{22d} have simple poles at $s = 1$. The same result is true at $s = -1$. Other values of $s = \lambda$ which must be considered are those where H_a (and also H_d) has poles. Since for $s \neq \pm 1$, g_1 and g_2 are linearly independent, (69) will

be satisfied automatically. If g_1 or g_2 had poles they would have to be included as zeros of h_{11d} or h_{22d}. As before, it is often sufficient to impose the constraints only at λ values which have positive real parts, letting some stable modes be uncontrollable or unobservable. In this example, the only active constraint would then be that h_{11d} and h_{22d} have simple poles at $s = 1$.

Usually, the g_i are analytic at $s = \lambda_{1b}, \cdots, \lambda_{nb}$. In fact with ii), a necessary and sufficient condition for analyticity at $s = \lambda_{ib}$ is that K_{ib} and

$$\sum_{j=1, j\neq i}^{p} \frac{K_{jb}}{\lambda_{ib} - \lambda_{jb}} + D_b$$

have columns which span the p-dimensional coordinate space. Furthermore, if $G(s)$ is analytic at $s = \lambda_{ib}$ then it can be shown that $G(\lambda_{ib})$ is of rank $p - 1$. Thus for $\lambda = \lambda_{ib}$ equation (69) is satisfied if h_{iid}, $i = 1, \cdots, p$ have simple poles at $s = \lambda_{ib}$. Many times, a considerably less severe constraint is sufficient. For example, if $g_1(\lambda_{ib}) = 0$ ($g_2, \cdots, g_p$ are linearly independent) only h_{11d} requires a pole at $s = \lambda_{ib}$. Or suppose $g_3(\lambda_{ib}) = k_1 g_1(\lambda_{ib}) + k_2 g_2(\lambda_{ib})$ where k_1 and k_2 are arbitrary constants; then (69) is true if $\ell_i = 1, i = 1, 2, 3$ and $\ell_i = 0, i = 4, \cdots, p$.

Finally, consider an example where $G(s)$ is not analytic at $s = \lambda_{ib}$.

$$H_b = \begin{bmatrix} \dfrac{2s}{(s-1)(s+1)} & \dfrac{-(s-3)}{2(s+1)} \\ \dfrac{-(s-1)}{(s+1)} & \dfrac{-2(s-1)}{(s+1)} \end{bmatrix} \tag{71}$$

and

$$G = \begin{bmatrix} \dfrac{4(s-1)}{(s+3)} & \dfrac{-(s-3)}{(s+3)} \\ \dfrac{-2(s-1)}{(s+3)} & \dfrac{-4s}{(s+3)(s-1)} \end{bmatrix}. \tag{72}$$

Take $\lambda = 1$. If $\ell_1 = \ell_2 = 1$ (the usual constraint [9, 12]), (69) is not satisfied (S is not controllable and observable even if a multiple pole treatment is considered); but it will be satisfied if $\ell_1 = 1$ and h_{22d} has a zero at $s = 1$. If stable modes are to be controllable and observable $\ell_1 = \ell_2 = 1$ at $\lambda = -1$ and h_{11d} and h_{22d} must have zeros at $s = -3$.

Though the above is a limited treatment, it does allow solution of many compensation problems and indicates the complexity of the situation. With obvious modifications the case of post-compensation (H_α fixed) can be handled.

8. Conclusion. From the foregoing it should be concluded that too great an emphasis on operational methods (the transfer-function matrix) is unwise. Differential equations (1) arise naturally in relating the physical properties of a system to its response characteristics, and any mathematical procedure which neglects information contained in these equations should be viewed skeptically.[11] It is surprising that physical considerations have not raised more doubts about the transfer-function representation earlier. Certainly, the errors of underestimated order would not have occurred if any effort had been made to relate the mathematical representation to the physical world—for example, by means of system simulation.

Finally, it should be noted that the synthesis of a multivariable feedback system is truly a formidable task. Unwieldy calculations, complex compensation constraints, and difficulties in evaluating the effect of disturbance inputs and parameter variations all complicate the search for satisfactory design procedures. The results developed above should at least provide a sound basis for this search.

Acknowledgment. The author wishes to thank Edward O. Gilbert and Bernard S. Morgan for helpful suggestions during the writing of the paper. The work was partially supported by the National Aeronautics and Space Administration under contract number NAS-8-1569.

REFERENCES

[1] A. S. Boksenbom and R. Hood, *General Algebraic Method Applied to Control Analysis of Complex Engine Types*, National Advisory Committee for Aeronautics, Technical Report 980, Washington, D. C., 1949.

[2] M. Golomb and E. Usdin, *A theory of multidimensional servo systems*, J. Franklin Inst., 253 (1952), pp. 29–57.

[3] H. S. Tsien, *Engineering Cybernetics*, McGraw-Hill Book Company, Inc., New York, 1954, Ch. 5, pp. 53–69.

[4] D. J. Povejsil and A. M. Fuchs, *A method for the preliminary synthesis of a complex multiloop control system*, Trans. Amer. Inst. Elec. Engrs., 74, Part II (1955), pp. 129–34.

[5] R. J. Kavanagh, *The application of matrix methods to multivariable control systems*, J. Franklin Inst., 262 (1956), pp. 349–67.

[6] R. J. Kavanagh, *Noninteraction in linear multivariable systems*, Trans.Amer. Inst. Elec. Engrs., 76, Part II (1957), pp. 95–100.

[7] R. J. Kavanagh, *Multivariable control system synthesis*, Trans. Amer. Inst. Elec. Engrs., 77, Part II (1958), pp. 425–429.

[8] H. Freeman, *A synthesis method for multipole control systems*, Trans. Amer. Inst. Elec. Engrs., 76, Part II (1957), pp. 28–31.

[9] H. Freeman, *Stability and physical realizability considerations in the synthesis*

[11] Similar remarks apply to sampled-data systems, except the vector differential equations are replaced by vector difference equations and the Laplace transform is replaced by the z-transform.

of multipole control systems, Trans. Amer. Inst. Elec. Engrs., 77, Part II (1958), pp. 1–5.

[10] I. M. Horowitz, *Synthesis of linear, multivariable feedback control systems*, Trans. I. R. E., Prof. Group on Automatic Control, AC-5, no. 2 (1960), pp. 94–105.

[11] E. V. Bohn, *Stabilization of linear multivariable feedback control systems*, Trans. I. R. E., Prof. Group on Automatic Control, AC-5, no. 4 (1960), pp. 321–327.

[12] K. Chen, R. A. Mathias, and D. M. Sauter, *Design of non-interacting control systems using bode diagrams*, Trans. Amer. Inst. Elec. Engrs., 80, Part II (1961), pp. 336–346.

[13] R. E. Kalman, *On the general theory of control systems*, Proc. First International Congress of Automatic Control, Moscow, USSR, 1960.

[14] R. E. Kalman, Y. C. Ho, and K. S. Narendra, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1, 1961. Vol. 1, No. 1 (1961), John Wiley, New York.

[15] Richard Bellman, *Introduction to Matrix Analysis*, McGraw-Hill Book Company, Inc., New York, 1960.

[16] R. E. Kalman, *Mathematical description of linear dynamical systems*, J. Soc. Indust. Appl. Math. Ser. A: On Control, Vol. 1, No. 2 (1963), pp. 152–192.

[17] Y. C. Ho, *What constitutes a controllable system?*, Trans. I. R. E., Prof. Group on Automatic Control, AC-7, no. 3 (1962), p. 76.

[18] Brockway McMillan, *Introduction to formal realizability theory*, Bell System Tech. J., 31 (1952), pp. 217–279 and pp. 541–600.

[19] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, 1, Inter-Science Publishers, Inc., New York, 1953.

J.S.I.A.M. CONTROL
Ser. A, Vol. 1, No. 2
Printed in U.S.A., 1963

MATHEMATICAL DESCRIPTION OF LINEAR DYNAMICAL SYSTEMS*

R. E. KALMAN†

Abstract. There are two different ways of describing dynamical systems: (i) by means of state variables and (ii) by input/output relations. The first method may be regarded as an axiomatization of Newton's laws of mechanics and is taken to be the basic definition of a system.

It is then shown (in the linear case) that the input/output relations determine only one part of a system, that which is completely observable and completely controllable. Using the theory of controllability and observability, methods are given for calculating irreducible realizations of a given impulse-response matrix. In particular, an explicit procedure is given to determine the minimal number of state variables necessary to realize a given transfer-function matrix. Difficulties arising from the use of reducible realizations are discussed briefly.

1. Introduction and summary. Recent developments in optimal control system theory are based on vector differential equations as models of physical systems. In the older literature on control theory, however, the same systems are modeled by transfer functions (i.e., by the Laplace transforms of the differential equations relating the inputs to the outputs). Two different languages have arisen, both of which purport to talk about the same problem. In the new approach, we talk about state variables, transition equations, etc., and make constant use of abstract linear algebra. In the old approach, the key words are frequency response, pole-zero patterns, etc., and the main mathematical tool is complex function theory.

Is there really a difference between the new and the old? Precisely what are the relations between (linear) vector differential equations and transfer-functions? In the literature, this question is surrounded by confusion [1]. This is bad. Communication between research workers and engineers is impeded. Important results of the "old theory" are not yet fully integrated into the new theory.

In the writer's view—which will be argued at length in this paper—the difficulty is due to insufficient appreciation of the concept of a *dynamical system*. Control theory is supposed to deal with physical systems, and not merely with mathematical objects such as a differential equation or a transfer function. We must therefore pay careful attention to the relationship between physical systems and their representation via differential equations, transfer functions, etc.

* Received by the editors July 7, 1962 and in revised form December 9, 1962.

Presented at the Symposium on Multivariable System Theory, SIAM, November 1, 1962 at Cambridge, Massachusetts.

This research was supported in part under U. S. Air Force Contracts AF 49(638)-382 and AF 33(616)-6952 as well as NASA Contract NASr-103.

† Research Institute for Advanced Studies (RIAS), Baltimore 12, Maryland.

To clear up these issues, we need first of all a precise, abstract definition of a (physical) dynamical system. (See sections 2–3.) The axioms which provide this definition are generalizations of the Newtonian world-view of causality. They have been used for many years in the mathematical literature of dynamical systems. Just as Newtonian mechanics evolved from differential equations, these axioms seek to abstract those properties of differential equations which agree with the "facts" of classical physics. It is hardly surprising that under special assumptions (finite-dimensional state space, continuous time) the axioms turn out to be equivalent to a system of ordinary differential equations. To avoid mathematical difficulties, we shall restrict our attention to linear differential equations.

In section 4 we formulate the central problem of the paper:

Given an (experimentally observed) impulse response matrix, how can we identify the linear dynamical system which generated it?

We propose to call any such system a *realization* of the given impulse response. It is an *irreducible realization* if the dimension of its state space is minimal.

Section 5 is a discussion of the "canonical structure theorem" [2, 14] which describes abstractly the coupling between the external variables (input and output) and the internal variables (state) of any linear dynamical system. As an immediate consequence of this theorem, we find that *a linear dynamical system is an irreducible realization of an impulse-response matrix if and only if the system is completely controllable and completely observable.* This important result provides a link between the present paper and earlier investigations in the theory of controllability and observability [3–5].

Explicit criteria for complete controllability and complete observability are reviewed in a convenient form in section 6.

Section 7 provides a constructive computational technique for determining the canonical structure of a constant linear dynamical system.

In section 8 we present, probably for the first time, a complete and rigorous theory of how to define the state variables of a multi-input/multi-output constant linear dynamical system described by its transfer-function matrix. Since we are interested only in irreducible realizations, there is a certain unique, well-defined number n of state variables which must be used. We give a simple proof of a recent theorem of Gilbert [5] concerning the value of n. We give canonical forms for irreducible realizations in simple cases. We give a constructive procedure (with examples) for finding an irreducible realization in the general case.

Many errors have been committed in the literature of system theory by carelessly regarding transfer functions and systems as equivalent concepts. A list of these has been collected in section 9.

The field of research outlined in this paper is still wide open, except

perhaps in the case of constant linear systems. Very little is known about irreducible realizations of nonconstant linear systems. It is not clear what additional properties—besides complete controllability and complete observability—are required to identify the stability type of a system from its impulse response. Nothing is known about nonlinear problems in this context.

Finally, the writer would like to acknowledge his indebtedness to Professor E. G. Gilbert, University of Michigan, whose work [5] predates this and whose results were instrumental in establishing the canonical structure theorem.

2. Axiomatic definition of a dynamical system. Macroscopic physical phenomena are commonly described in terms of cause-and-effect relationships. This is the "Principle of Causality". The idea involved here is at least as old as Newtonian mechanics. According to the latter, the motion of a system of particles is fully determined for all future time by the present positions and momenta of the particles and by the present and future forces acting on the system. How the particles actually attained their present positions and momenta is immaterial. Future forces can have no effect on what happens at present.

In modern terminology, we say that the numbers which specify the instantaneous position and momentum of each particle represent the *state* of the system. The state is to be regarded always as an abstract quantity. Intuitively speaking, the state is the minimal amount of information about the past history of the system which suffices to predict the effect of the past upon the future. Further, we say that the forces acting on the particles are the *inputs* of the system. Any variable in the system which can be directly observed is an *output*.

The preceding notions can be used to give a precise mathematical definition of a dynamical system [6]. For the present purposes it will be convenient to state this definition in somewhat more general fashion [14].

DEFINITION 1. A dynamical system is a mathematical structure defined by the following axioms:

(D_1) There is given a *state space* Σ and a set of values of *time* Θ at which the behavior of the system is defined; Σ is a topological space and Θ is an ordered topological space which is a subset of the real numbers.

(D_2) There is given a topological space Ω of functions of time defined on Θ, which are the admissible *inputs* to the system.

(D_3) For any initial time t_0 in Θ, any initial state x_0 in Σ, and any input u in Ω defined for $t \geqq t_0$, the future states of the system are determined by the transition function $\varphi: \Omega \times \Theta \times \Theta \times \Sigma \to \Sigma$, which is written as $\varphi_u(t;\, t_0,\, x_0) = x_t$. This function is defined

only for $t \geqq t_0$. Moreover, any $t_0 \leqq t_1 \leqq t_2$ in Θ, any x_0 in Σ, and any fixed u in Ω defined over $[t_0, t_1] \cap \Theta$, the following relations hold:

(D_3-i) $\varphi_u(t_0 ; t_0 , x_0) = x_0 ,$

(D_3-ii) $\varphi_u(t_2 ; t_0 , x_0) = \varphi_u(t_2 ; t_1 , \varphi_u(t_1 ; t_0 , x_0)).$

In addition, the system must be *nonanticipatory*, i.e., if u, $v \in \Omega$ and $u \equiv v$ on $[t_0 , t_1] \cap \Theta$ we have

(D_3-iii) $\varphi_u(t; t_0 , x_0) \equiv \varphi_v(t; t_0 , x_0).$

(D_4) Every *output* of the system is a function $\psi: \Theta \times \Sigma \to$ reals.

(D_5) The functions φ and ψ are continuous, with respect to the topologies defined for Σ, Θ, and Ω and the induced product topologies.

In this paper we will study only a very special subclass of dynamical systems: those which are *real*, *finite-dimensional*, *continuous-time*, and *linear*.

"Real, finite-dimensional" means that $\Sigma = R^n$ = n-dimensional real linear space. "Continuous-time" means that $\Theta = R^1$ = set of real numbers. "Linear" means that φ is linear on $\Omega \times \Sigma$ and ψ is linear on Σ.

By requiring φ and ψ to be sufficiently "smooth" functions, we can deduce from the axioms a set of equations which characterize every real, finite-dimensional, continuous-time, and linear dynamical system. The proof of this fact is outside the scope of the present paper [14]. Here we shall simply assume that every such system is governed by the equations

$$\frac{dx}{dt} = F(t)x + G(t)u(t), \tag{2.1}$$

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

defined on the whole real line $-\infty < t < \infty$, where x, u, and y are n, m, and p-vectors* respectively, and the matrices $F(t)$, $G(t)$, and $H(t)$ are continuous functions of the time t.

We call (2.1–2) the *dynamical equations* of the system.

It is instructive to check whether the axioms are satisfied. (D_1) is obviously true; we have $\Sigma = R^n$, $\Theta = R^1$. The *state* of the system is the vector x. To satisfy (D_2), we must specify the class of all inputs, that is, a subclass of all vector functions $u(t) = (u_1(t), \cdots , u_m(t))$. To define Ω, we shall assume that these functions are piecewise continuous; this is sufficiently

* Vectors will be denoted by small Roman letters, matrices by Roman capitals. The components of a vector x are x_i, components of a matrix A are a_{ij}. On the other hand, $x^1, x^2, \cdots$, are vectors, and F^{AA}, F^{AB} are matrices. A' is the transpose of A.

general for most applications. We have exactly p observations on the system (the components of the vector y) and by (2.2) they are functions of t, x. Hence (D_4) is satisfied. To check (D_3), we recall that the general solution of (2.1) is given by

$$\varphi_u(t; t_0, x_0) \equiv x_t = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)G(\tau)u(\tau)\, d\tau, \tag{2.3}$$

where $\Phi(t, \tau)$ is the transition matrix of the free differential equation defined by $F(t)$ [4, 7]†. Since (2.3) is valid for any $t \geqq t_0$ (in fact, also for $t < t_0$), φ is well defined. Property (D_3-i) is obvious. (D_3-ii) follows from the composition property [4, 7] of the transition matrix:

$$\Phi(t, \sigma) = \Phi(t, \tau)\Phi(\tau, \sigma), \tag{2.4}$$

which holds for every set of real numbers t, τ, σ. Indeed, (2.4) is simply the linear version of (D_3-ii). (D_3-iii) is obvious from formula (2.3). The continuity axiom (D_5) is satisfied by hypothesis.

Evidently φ given by (2.3) is linear on the cartesian product of Σ with the linear space of vector-valued piecewise continuous functions.

We call a linear dynamical system (2.1–2) *constant*, *periodic*, or *analytic* whenever F, G, and H are constant, periodic, or analytic in t.

It is often convenient to have a special name for the couple $(t, x) \mid \in \Theta \times \Sigma$. Giving a fixed value of (t, x) is equivalent to specifying at some time (t) the state (x) of the system. We shall call (t, x) a *phase* and $\Theta \times \Sigma$ the *phase space*. (Recall the popular phrase: "phases" of the Moon.)

To justify our claim—implicit in the above discussion—that equations (2.1–2) are a good model of physical reality, we wish to point out that these equations can be concretely simulated by a simple physical system: a general-purpose analog computer. Indeed, the numbers (or functions) constituting F, G, and H may be regarded as specifying the "wiring diagram" of the analog computer which simulates the system (2.1–2) (see, for instance, [8]).

3. Equivalent dynamical systems. The state vector x must always be regarded as an abstract quantity. By definition, it cannot be directly measured. On the other hand, the inputs and outputs of the system (2.1–2) have concrete physical meaning. Bearing this in mind, equations (2.1–2) admit two interpretations:

(a) They express relations involving the abstract linear transformations $F(t)$, $G(t)$, and $H(t)$.

(*b*) At any fixed time, we take an arbitrary but fixed coördinate system

† I.e., Φ is a solution of $d\Phi/dt = F(t)\Phi$, subject to the initial condition $\Phi(\tau, \tau) = I$ = unit matrix for all τ.

in the (abstract) vector space Σ. Then the symbol $x \equiv (x_1, \cdots, x_n)$ is interpreted as the numerical n-tuple consisting of the coördinates of the abstract state vector which is also denoted by x. F, G, and H are interpreted as the matrix representations of the abstract linear transformations denoted by the same letters under (a).

To describe the behavior of a dynamical system in concrete terms, the second point of view must be used. Then we must also ask ourselves the question: To what extent does the description of a dynamical system depend on the arbitrary choice of the coordinate system in the state space? (No such arbitrariness occurs in the definition of the numerical vectors u, y since the input and output variables u_i and y_j are concrete physical quantities.) This question gives rise to the next definition.

DEFINITION 2. Two linear dynamical systems (2.1–2), with state vectors x, $\bar{x}$, are *algebraically equivalent* whenever their numerical phase vectors are related for all t as

$$(t, \bar{x}) = (t, T(t)x), \tag{3.1}$$

where $T(t)$ is a $n \times n$ matrix, nonsingular for all t and continuously differentiable in t. In other words, there is a 1-1 differentiable correspondence between the phase spaces $\Theta \times \Sigma$ and $\Theta \times \bar{\Sigma}$.

Remark: We could generalize this definition of equivalence to $(\bar{t}, \bar{x}) = (\tau(t), T(t)x)$ where τ is an increasing function of t. But this involves distortion of the time scale which is not permitted in Newtonian physics.

Algebraic equivalence implies the following relations between the defining matrices of the two systems:

$$\begin{aligned} \bar{\Phi}(t, \tau) &= T(t)\Phi(t, \tau)T^{-1}(\tau), \\ \bar{F}(t) &= \dot{T}(t)T^{-1}(t) + T(t)F(t)T^{-1}(t), \\ \bar{G}(t) &= T(t)G(t), \\ \bar{H}(t) &= H(t)T^{-1}(t). \end{aligned} \tag{3.2}$$

In general, algebraic equivalence does not preserve the stability properties of a dynamical system [7, 9, 10]. For this it is necessary and sufficient to have *topological equivalence*: algebraic equivalence plus the condition

$$\| T(t) \| \leqq c_1 \quad \text{and} \quad \| T^{-1}(t) \| \leqq c_2, \tag{3.3}$$

where c_1 and c_2 are fixed constants, and $\| \ \|$ is the euclidean norm*.

A nonconstant system may be algebraically and even topologically equivalent to a constant system. The latter case is called by Markus [11]

* Let Θ, Σ, and $\bar{\Sigma}$ have the usual topologies induced by the euclidean norm. Then the product topologies induced on $\Theta \times \Sigma$ and $\Theta \times \bar{\Sigma}$ are equivalent if and only if (3.3) holds.

"kinematic similarity". Moreover, two constant systems may be algebraically and topologically equivalent without $T(t)$ being a constant. To bypass these complications, we propose

DEFINITION 3. Two constant linear dynamical systems are *strictly equivalent* whenever their numerical phase vectors are related for all t as $(t, \bar{x}) = (t, Tx)$, where T is a nonsingular constant matrix.

Evidently strict equivalence implies topological equivalence.

4. The impulse-response matrix and its realization by a linear dynamical system.

Sections 2–3 were concerned with mathematics, that is, abstract matters. If we now take the point of view of physics, then a dynamical system must be "defined" in terms of quantities which can be directly observed. For linear dynamical systems, this is usually done in the following way.

We consider a system which is at rest at time t_0 ; i.e., one whose input and outputs have been identically zero for all $t \leqq t_0$. We apply at each input in turn a very sharp and narrow pulse. Ideally, we would take $u_i^{(j)}(t) = \delta_{ij}\delta(t - t_0)$, where δ is the Dirac delta function, δ_{ij} is the Kronecker symbol, and $1 \leqq i, j \leqq m$. We then observe the effect of each vector input $u^{(j)}(t)$ on the outputs, which are denoted by $u(t; j)$. The matrix $S(t, t_0) = [s_{ij}(t, t_0)] = [y_i(t; j)]$ so obtained is called the *impulse-response matrix* of the system. Since the system was at rest prior to $t = t_0$, we must define $S(t, t_0) \equiv 0$ for $t < t_0$. We also assume, of course, that S is continuous in t and t_0 for $t > t_0$.

With these conventions, the output of a linear system originally at rest is related to its input by the well-known convolution integral:

$$y(t) = \int_{t_0}^{t} S(t, \tau)u(\tau)\, d\tau. \tag{4.1}$$

In much of the literature of system theory [12] (and also at times in physics) formula (4.1) is the basic definition of a system. The Fourier transform of S is often called "the system function" [13, p. 92].

Unfortunately, this definition does not explain how to treat systems which are not "initially at rest". Hence we may ask, "To what extent, if any, are we justified in equating the physical definition (4.1) of a system with the mathematical one provided by (2.1–2)?"

Suppose that the system in question is actually (2.1–2). Then (2.3) shows that

$$\begin{aligned} S(t, \tau) &= H(t)\Phi(t, \tau)G(\tau), && t \geqq \tau, \\ &= 0, && t < \tau.\dagger \end{aligned} \tag{4.2}$$

† The right-hand side of the first equation (4.2) is defined also for $t < \tau$; then the left-hand side may be regarded as the "backward impulse response", whose physical interpretation is left to the reader.

Thus it is trivial to calculate the impulse-response matrix of a given linear dynamical system. The converse question, however, is non trivial and interesting. *When and how does the impulse-response matrix determine the dynamical equations of the system?*

This problem is commonly called the *identification* of the system from its impulse-response matrix.

Having been given an impulse-response matrix, suppose that we succeed in finding matrices F, G, and H such that (4.2) holds. We have then identified a physical system that may have been the one which actually generated the observed impulse-response matrix. We shall therefore call (2.12) a *realization* of $S(t, \tau)$. This terminology is justified because the axioms given in section 2 are patterned after highly successful models of classical macroscopic physics; in fact, the system defined by (2.1–2) can be concretely realized, actually built, using standard analog-computer techniques in existence today. In short, proceeding from the impulse-response matrix to the dynamical equations we get closer to "physical reality". But we are also left with a problem: Which one of the (possibly very many) realizations of $S(t, \tau)$ is the actual system that we are dealing with?

It is conceivable that certain aspects of a dynamical system cannot ever be identified from knowledge of its impulse response, as our knowledge of the physical world gained from experimental observation must always be regarded as incomplete. Still, it seems sensible to ask how much of the physical world can be determined from a given amount of experimental data.

The first clear problem statement in this complex of ideas and the first results appear to be due to the writer [2, 14].

First of all we note

THEOREM 1. *An impulse-response matrix $S(t, \tau)$ is realizable by a finite-dimensional dynamical system* (2.1–2) *if and only if there exist continuous matrices $P(t)$ and $Q(t)$ such that*

$$S(t, \tau) = P(t)Q(\tau) \quad \textit{for all} \quad t, \tau. \tag{4.3}$$

Proof. Necessity follows by writing the right-hand side of (4.2) as $H(t)\Phi(t, 0)\Phi(0, \tau)G(\tau)$, with the aid of (2.4). Sufficiency is equally obvious. We set $F(t) = 0$, $G(t) = Q(t)$, and $H(t) = P(t)$. Then $\Phi(t, \tau) \equiv I$ and the desired result follows by (4.2).

A realization (2.1–2) of $S(t, \tau)$ is *reducible* if over some interval of time there is a proper (i.e., lower-dimensional) subsystem of (2.1–2) which also realizes $S(t, \tau)$. As will be seen later, a realization of S (particularly the one given in the previous paragraph) is often reducible.

An impulse-response matrix S is *stationary* whenever $S(t, \tau) = S(t + \sigma, \tau + \sigma)$ for all real numbers t, τ, and σ. S is *periodic* whenever

the preceding relation holds for all t, τ, and some σ. An impulse-response matrix is *analytic* whenever S is analytic in t and τ; if (4.3) holds, then P and Q must be analytic in t.

The main result, whose proof will be discussed later, is the following [14]:

THEOREM 2. *Hypothesis: The impulse-response matrix S satisfies* (4.3) *and is either periodic (and continuous) or analytic.*

Conclusions: (i) *There exist irreducible realizations of S, all of which have the same constant dimension n and are algebraically equivalent.* (ii) *If S is periodic [analytic] so are its irreducible realizations.*

Topological equivalence cannot be claimed in general. It may happen that S has one realization which is asymptotically stable and another which is asymptotically unstable [15]. Hence it may be impossible to identify the stability of a dynamical system from its impulse response! This surprising conclusion raises many interesting problems which are as yet unexplored [15]. If S is not periodic or analytic, it may happen that the dimension $n(t)$ of an irreducible realization is constant only over finite time intervals.

In the stationary case, Theorem 2 can be improved [14].

THEOREM 3. *Every stationary impulse-response matrix $S(t, \tau) = W(t - \tau)$ satisfying* (4.3) *has constant irreducible realizations. All such realizations are strictly equivalent.*

In view of this theorem, we may talk indifferently about a stationary impulse-response matrix or the dynamical system which generates it—as has long been the practice in system theory on intuitive grounds. But note that we must require the realization to be irreducible. For nonconstant systems, such a conclusion is at present not justified. The requirement of irreducibility in Theorem 3 is essential; disregarding it can lead—and has led—to serious errors in modeling dynamical systems. (See section 9.)

In many practical cases, it is not the *weighting-function matrix* $W(t - \tau)$ (see Theorem 3) which is given, but its Laplace transform, the *transfer-function matrix* $Z(s) = \mathfrak{L}[W(t)]$. Then condition (4.3) has an interesting equivalent form, which is often used as a "working hypothesis" in engineering texts:

THEOREM 4. *A weighting-function matrix $W(t - \tau)$ satisfies* (4.3) *if and only if its elements are linear combinations of terms of the type $t^i e^{s_j t}$ ($i = 0, 1, \cdots, n - 1, j = 1, \cdots, n$). Hence every element of the transfer-function matrix is a ratio of polynomials in s such that the degree of the denominator polynomial always exceeds the degree of the numerator polynomial.*

This result is provedd in [14]. It implies that the realization of an impulse-response matrix is equivalent to expressing the elements of F, G, and H as functions of the coefficients of the numerator and denominator polynomials of elements of $Z(s)$. (See section 8.)

In the remainder of the paper, we wish to investigate two main problems

arising in the theory sketched above:

(i) Explicit criteria for reducibility.

(ii) Construction of irreducible realizations.

Remark. Elementary expositions of system theory often contain the statement that the operator d/dt ($\equiv s$) is a "system." Is a it system in the same sense as that word is used here? The answer is no. To define such a system rigorously in accordance with the axioms introduced in section 2, one must proceed as follows. The output of the system, which by definition is the derivative of the input, is given by

$$y(t) = \frac{du(t)}{dt} = \psi(t, x(t)), \tag{3.4}$$

so that at any fixed t, $u(t)$ must be a *point* function of $(t, x(t))$. Therefore the state space Σ must include the space Ω of functions on which the operator d/dt is defined. It is simplest to let $\Sigma = \Omega$. Then Σ is usually infinite dimensional because Ω is. Thus we define the state $x \equiv x(t)$ as the function $u(\tau)$, defined for all $\tau \leqq t$. The mapping $\varphi_u(t; t_0, x_{t_0})$ assigns to the function x_0 defined for $\tau \leqq t_0$ the function x_t, which is equal to x_{t_0} on $\tau \leqq t_0$ and equal to u on $t_0 < \tau \leqq t$.

In this paper, the finite dimensionality of Σ is used in an essential way, which rules out consideration of the "system" d/dt in all but trivial cases.

5. Canonical structure of linear dynamical systems. The concept of irreducibility can be understood most readily with the help of the writer's "canonical structure theorem" for linear dynamical systems [2, 14].

Before presenting and illustrating this central result, it is necessary to recall some definitions and facts concerning the *controllability* and *observability* of linear dynamical systems.

DEFINITION 4. A linear dynamical system (2.1-2) is *completely controllable* at time t_0 if it is not algebraically equivalent, for all $t \geqq t_0$, to a system of the type

$$\begin{aligned} &\text{(a)} \quad dx^1/dt = F^{11}(t)x^1 + F^{12}(t)x^2 + G^1(t)u(t) \\ &\text{(b)} \quad dx^2/dt = F^{22}(t)x^2 \\ &\text{(c)} \quad y(t) = H^1(t)x^1(t) + H^2(t)x^2(t). \end{aligned} \tag{5.1}$$

(In (5.1), x^1 and x^2 are vectors of n_1 and $n_2 = n - n_1$ components respectively.)

In other words, it is *not* possible to find a coördinate system in which the state variables x_i are separated into two groups, $x^1 = (x_1, \cdots, x_{n_1})$ and $x^2 = (x_{n_1+1}, \cdots, x_n)$, such that the second group is not affected either by the first group or by the inputs to the system. If one could find such a

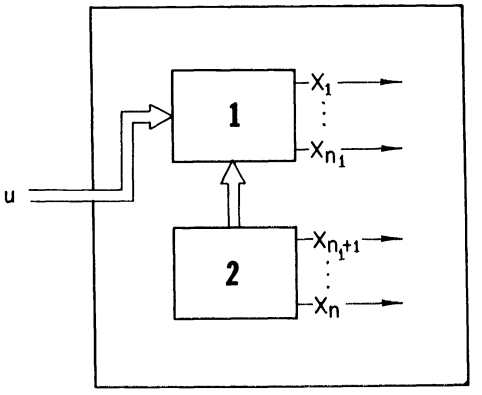

FIGURE 1.

coördinate system, we would have the state of affairs depicted schematically in Fig. 1.

Clearly, controllability is a system property which is completely independent of the way in which the outputs of the system are formed. It is a property of the couple $\{F(t), G(t)\}$.

The "dual" of controllability is observability, which depends only on the outputs but not on the inputs.

DEFINITION 5. A linear dynamical system (2.1–2) is *completely observable* at time t_0 if it is not algebraically equivalent, for all $t \leqq t_0$, to any system of the type

$$\begin{aligned} &(a) \quad dx^1/dt = F^{11}(t)x^1(t) + G^1(t)u(t) \\ (5.2)\quad &(b) \quad dx^2/dt = F^{21}(t)x^1(t) + F^{22}(t)x^2 + G^2(t)u(t) \\ &(c) \quad y(t) = H^1(t)x^1(t). \end{aligned}$$

(Again, x^1 is an n_1-vector and x^2 is an $(n - n_1)$-vector.)

In other words, it is not possible to find a coördinate system in which the state variables x_i are separated into two groups, such that the second group does not affect either the first group or the outputs of the system. If such a coördinate system could be found, we would have the state of affairs depicted in Fig. 2.

The above definitions show that controllability and observability are preserved under algebraic equivalence. These properties are coördinate-free, i.e., independent of the particular choice of basis in the state space.

The equivalence of the present definitions with other more abstract

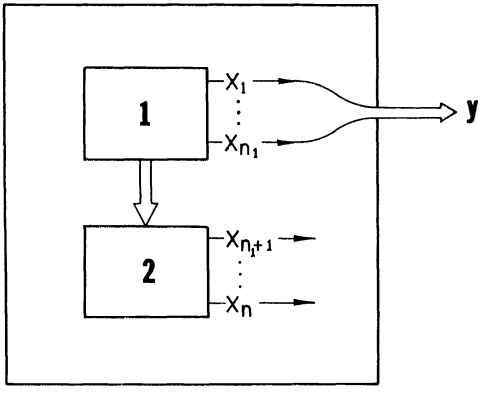

FIGURE 2.

definitions of controllability may be found in [4]. As to observability, we note that the *duality relations*

$$
\begin{aligned}
&\text{(a)} \qquad t - t_0 = t_0 - t', \\
&\text{(b)} \qquad F(t - t_0) \Leftrightarrow F'(t_0 - t'), \\
&\text{(c)} \qquad G(t - t_0) \Leftrightarrow H'(t_0 - t'), \\
&\text{(d)} \qquad H(t - t_0) \Leftrightarrow G'(t_0 - t'),
\end{aligned}
\tag{5.3}
$$

transform the system (5.2) into (5.1). Hence all theorems on controllability can be "dualized" to yield analogous results on observability.

It can be shown that in applying definitions 4–5 to constant systems it is immaterial whether we require algebraic or strict equivalence [14]. Hence—as one would of course expect—for constant systems the notions of complete controllability and complete observability do not depend on the choice of t_0.

EXAMPLE 1. A simple, well-known, and interesting case of a physical system which is neither completely controllable nor completely observable is the so-called constant-resistance network shown in Fig. 3.

Let x_1 be the magnetic flux in the inductor and x_2 the electric charge on the capacitor in Fig. 3, while $u_1(t)$ is a voltage source (zero short-circuit resistance) and $y_1(t)$ is the current into the network. The inductor and capacitor in the network may be time-varying, but we assume—this is the constant-resistance condition—that $L(t)$ and $C(t)$ are related by:

$$L(t)/C(t) = R^2 = 1 \qquad (L(t), C(t) > 0).$$

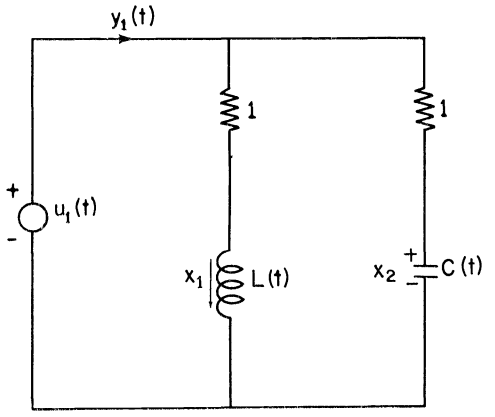

FIGURE 3.

The differential equations of the network are

$$dx_1/dt = -[1/L(t)]x_1 + u_1(t),$$
$$dx_2/dt = -[1/C(t)]x_2 + u_1(t),$$
$$y_1(t) = [1/L(t)]x_1 - [1/C(t)]x_2 + u_1(t).$$

If we let

$$\bar{x}_1 = (x_1 + x_2)/2,$$
$$\bar{x}_2 = (x_1 - x_2)/2,$$

the dynamical equations become

$$d\bar{x}_1/dt = -[1/L(t)]\bar{x}_1 + u_1(t),$$
$$d\bar{x}_2/dt = -[1/L(t)]\bar{x}_2, \tag{5.4}$$
$$y_1(t) = 2[1/L(t)]\bar{x}_2 + u_1(t).^*$$

Here the state variable $\bar{x}_1$ is controllable but not observable, while $\bar{x}_2$ is observable but not controllable.

For obvious reasons, the subsystem (b) of (5.1) may be regarded as (completely) uncontrollable, while subsystem (b) of (5.2) is (completely) unobservable. In view of linearity, it is intuitively clear that it must be possible to arrange the components of the state vector—referred to a

* Note that this equation does not correspond to (2.2) but to $y(t) = H(t)x(t) + J(t)u(t)$. This is a minor point. In fact, Axiom (D_4) may be generalized to: "(D_4): Every output is a function of t, $x(t)$, and $u(t)$." This entails only minor modifications as far as the results and arguments of the present paper are concerned.

suitable (possibly time-varying) coördinate system—into four mutally exclusive parts, as follows:

Part (A): Completely controllable but unobservable.

Part (B): Completely controllable and completely observable.

Part (C): Uncontrollable and unobservable.

Part (D): Uncontrollable but completely observable.

The precise statement of this idea is [2, 14]:

THEOREM 5 (*Canonical Structure Theorem*). *Consider a fixed linear dynamical system* (2.1–2).

(i) *At every fixed instant t of time, there is a coördinate system in the state space relative to which the components of the state vector can be decomposed into four mutually exlusive parts*

$$x = (x^A, x^B, x^C, x^D),$$

which correspond to the scheme outlined above.

(ii) *This decomposition can be achieved in many ways, but the number of state variables* $n_A(t), \cdots, n_D(t)$ *in each part is the same for any such decomposition.*

(iii) *Relative to such a choice of coördinates, the system matrices have the canonical form*

$$F(t) = \begin{bmatrix} F^{AA}(t) & F^{AB}(t) & F^{AC}(t) & F^{AD}(t) \\ 0 & F^{BB}(t) & 0 & F^{BD}(t) \\ 0 & 0 & F^{CC}(t) & F^{CD}(t) \\ 0 & 0 & 0 & F^{DD}(t) \end{bmatrix},$$

$$G(t) = \begin{bmatrix} G^A(t) \\ G^B(t) \\ 0 \\ 0 \end{bmatrix},$$

and

$$H(t) = [0 \quad H^B(t) \quad 0 \quad H^D(t)].$$

In view of this theorem, we shall talk, somewhat loosely, about "Parts $(A), \cdots, (D)$ of the system." Thus the system (5.4) consists of Parts (A) and (D).

The canonical form of F, G, and H can be easily remembered by reference to the causal diagram shown on Fig. 4.

It is intuitively clear (and can be easily proved) that algebraically equivalent systems have the same canonical structure.

Unfortunately, the coördinate system necessary to display the canonical form of F, G, and H will not be continuous in time unless $n_A(t), \cdots, n_D(t)$ are constants. If these dimension numbers vary, we cannot call the various

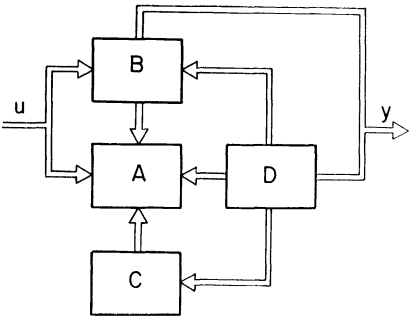

FIGURE 4.

parts of the canonical structure "subsystems." For constant systems this difficulty does not arise. More generally, we have:

THEOREM 6. *For a periodic or analytic linear dynamical system* (2.1–2) *the dimension numbers* $n_A, \cdots, n_D$ *are constants, and the canonical decomposition is continuous with respect to* t.

An illustration of the canonical structure theorem is provided by

EXAMPLE 2. Consider the constant system defined by

$$F = \begin{bmatrix} -3 & -3 & 0 & 1 \\ 26 & 36 & -3 & -25 \\ 30 & 39 & -2 & -27 \\ 30 & 43 & -3 & -32 \end{bmatrix},$$

$$G = \begin{bmatrix} 3 & 3 \\ -2 & -1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$H = [-5 \quad -8 \quad 1 \quad 5].$$

We introduce new coördinates by letting $\bar{x} = Tx$, where

$$T = \begin{bmatrix} 2 & 3 & 0 & -2 \\ 1 & 1 & 0 & -1 \\ -2 & -3 & 0 & 3 \\ -6 & -9 & 1 & 6 \end{bmatrix},$$

and

$$T^{-1} = \begin{bmatrix} 0 & 3 & 1 & 0 \\ 1 & -2 & 0 & 0 \\ 3 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

With respect to these new coördinates the system matrices assume the

canonical form:

$$\bar{F} = TFT^{-1} = \begin{bmatrix} 2 & 4 & 1 & -1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -3 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\bar{G} = TG = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

and

$$\bar{H} = HT^{-1} = [0 \quad 1 \quad 0 \quad 1].$$

On the other hand, if we define the new coördinates by

$$T = \begin{bmatrix} 3 & 4 & 0 & -3 \\ 1 & 1 & 0 & -1 \\ -5 & -7.5 & 0.5 & 6 \\ -6 & 9 & 1 & 6 \end{bmatrix},$$

$$T^{-1} = \begin{bmatrix} 0 & 3 & 1 & -0.5 \\ 1 & -3 & 0 & 0 \\ 3 & -3 & 0 & 1 \\ 1 & -1 & 1 & -0.5 \end{bmatrix},$$

then the system matrices become

$$\bar{F} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\bar{G} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

and

$$\bar{H} = [0 \quad 1 \quad 0 \quad 1].$$

The numerical values of these two canonical forms are different, yet Theorem 5 is verified in both cases. In the second case the connections from Part (D) to Parts (A) and (C) are missing. This is not a contradiction since Theorem 5 does not require that all the indicated casual connections in Fig. 4 be actually present.

The transfer-function matrix of the system is easily found from the canonical representation. The coördinate transformations affect only the

internal (state) variables, but not the external (input and output) variables; consequently the impulse response matrix is invariant under such transformations. We get by inspection:

$$Z(s) = \begin{bmatrix} \dfrac{1}{s+1} & \dfrac{1}{s+1} \end{bmatrix}.$$

It would be rather laborious to determine these transfer functions directly from the signal-flow graph [16] corresponding to F, G, and H.

EXAMPLE 3. A far less trivial illustration of the canonical decomposition theorem is provided by the following dynamical system, which occurs in the solution of a problem in the theory of statistical filtering [17]. Let A be an arbitrary positive function of t and define

$$F = \begin{bmatrix} -t^4/4A & 1 & 0 \\ -t^3/2A & 0 & 1 \\ -t^2/2A & 0 & 0 \end{bmatrix},$$

$$G = \begin{bmatrix} t^4/4A \\ t^3/2A \\ t^2/2A \end{bmatrix},$$

and

$$H = [0 \quad 1 \quad 0].$$

We introduce new state variables

$$\bar{x}(t) = T(t)x(t),$$

where

$$T(t) = \begin{bmatrix} 0 & 0 & 1 \\ 2 & -t & 0 \\ 0 & 1 & -t \end{bmatrix},$$

$$T^{-1}(t) = \begin{bmatrix} t^2/2 & 1/2 & t/2 \\ t & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Then

$$\bar{F}(t) = T(t)F(t)T^{-1}(t) - \dot{T}(t)T^{-1}(t) = \left[\begin{array}{c|c|c} -t^4/4A & -t^3/4A & -t^2/4A \\ \hline 0 & 0 & 1 \\ \hline 0 & 0 & 0 \end{array}\right],$$

$$\bar{G}(t) = T(t)G(t) = \left[\begin{array}{c} t^2/2A \\ \hline 0 \\ \hline 0 \end{array}\right],$$

and

$$H(t) = H(t)T^{-1}(t) = [t \mid 0 \mid 1].$$

Hence the system consists of Parts $(B - D)$, with $n_B = nc = n_D = 1$. It is interesting that the canonical decomposition is of constant dimension, even though the system may be neither periodic nor analytic.

The preceding examples illustrate special cases of a noteworthy general relationship which exists between the canonical structure of a dynamical system and irreducible realizations of an impulse-response matrix. The main facts here are the following:

THEOREM 7. (i) *The impulse-response matrix of a linear dynamical system* (2.1–2) *depends solely on Part* (B) *of the system and is given explicitly by*

$$S(t, \tau) = H^B(t)\Phi^{BB}(t, \tau)G^B(\tau), \tag{5.5}$$

where Φ^{BB} *is the transition matrix corresponding to* F^{BB}.

(ii) *Any two completely controllable and completely observable realizations of* S *are algebraically equivalent.*

(iii) *A realization of* S *is irreducible if and only if at all times it consists of Part* (B) *alone; thus every irreducible realization of* S *is completely controllable and completely observable.*

Proof. The first statement can be read off by inspection from Fig. 4. The second statement is proved in [14]. The necessity of the third statement follows from Theorem 5, while the sufficiency is implied by (ii).

It is clear that Theorem 2 is a consequence of Theorems 5–7.

We can now answer the question posed in section 4 in a definite way:

THEOREM 8 (*Main Result*). *Knowledge of the impulse-response matrix* $S(t, \tau)$ *identifies the completely controllable and completely observable part, and this part alone, of the dynamical system which generated it. This part ("*B*" in Theorem 5) is itself a dynamical system and has the smallest dimension among all realizations of* S. *Moreover, this part is identified by* S *uniquely up to algebraic equivalence.*

Using different words, we may say that an impulse-response matrix is a *faithful representation* of a dynamical system (2.1–2) if and only if the latter is completely controllable and completely observable.

Remark. It is very interesting to compare this result with Theorem 4 of E. F. Moore, in one of the early papers on finite automata [26]:

"*The class of all machines which are indistinguishable from a given strongly connected machine* S *by any single experiment has a unique (up to isomorphism) member with a minimal number of states. This unique machine, called the reduced form of* S, *is strongly connected and has the property that any two of its states are distinguishable.*"

"Indistinguishable machines" in Moore's terminology correspond in ours to alternate realizations of the same input/output relation. "Strongly con-

nected" in his terminology means completely controllable in ours. "Indistinguishable states" in our terminology corresponds to states whose difference, not zero, is an unobservable state in the sense of [3].

Evidently the two theorems are concerned with the same abstract facts, each being stated in a different mathematical framework.

6. Explicit criteria for complete controllability and observability. The canonical structure theorem is so far merely an abstract result, since we have not yet given a constructive procedure for obtaining the coördinate transformation which exhibits the system matrices in canonical form. We shall do this in section 7. The method rests on the possibility of finding explicit criteria for complete controllability and complete observability. The following lemmas, proved in [4], play a central role:

LEMMA 1. $n_A(t_0) + n_B(t_0) = \text{rank } W(t_0, t_1)$ *for* $t_1 > t_0$ *sufficiently large, where*

$$W(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_0, \tau)G(\tau)G'(\tau)\Phi'(t_0, \tau)\, d\tau \tag{6.1}$$

or

$$dW/dt_0 = F(t_0)W + WF'(t_0) - G(t_0)G'(t_0),\ W(t_1) = 0. \tag{6.2}$$

LEMMA 2. $n_c(t_0) + n_D(t_0) = \text{rank } M(t_0, t_{-1})$ *for* $t_{-1} < t_0$ *sufficiently small, where*

$$M(t_0, t_{-1}) = \int_{t_{-1}}^{t_0} \Phi'(\tau, t_0)H'(\tau)H(\tau)\Phi(\tau, t_0)\, d\tau \tag{6.3}$$

or

$$-dM/dt_0 = F'(t_0)M + MF(t_0) - H'(t_0)H(t_0),\ M(t_{-1}) = 0. \tag{6.4}$$

For constant systems, the preceding lemmas can be considerably improved [4]:

LEMMA 3. *For a constant system,*

$$n_A + n_B = \text{rank } [G, FG, \cdots, F^{n-1}G]. \tag{6.5}$$

LEMMA 4. *For a constant system,*

$$nc + n_D = \text{rank } [H', F'H', \cdots, (F')^{n-1}H']. \tag{6.6}$$

EXAMPLE 4. For F and G defined in Example 2, the matrix (6.5) is

$$\begin{bmatrix} 3 & 3 & -3 & -3 & 3 & 3 & -3 & -3 \\ -2 & -1 & 6 & 8 & 2 & 6 & 14 & 22 \\ 0 & 3 & 12 & 18 & 12 & 24 & 36 & 60 \\ 0 & 1 & 4 & 6 & 4 & 8 & 12 & 20 \end{bmatrix}. \tag{6.7}$$

The rank of this matrix is 2, which checks with the fact that $n_A = 1$ and $n_B = 1$ in Example 2.

The determination of the rank of (6.7), while elementary, is laborious. For practical purposes it might be better to compute W; for instance, by solving the differential equation (6.2).

In the constant case, there is another criterion of complete controllability which is particularly useful in theoretical investigations. The most general form of this theorem (which may be found in [14]) is complicated; we state here a simplified version which is adequate for the present purposes:

LEMMA 5. *Hypothesis: The matrix F is similar to a diagonal matrix. In other words, there is a nonsingular coördinate transformation $\bar{x} = Tx$ with the property that in the new coördinate system F has the form*

$$\bar{F} = TFT^{-1} = \begin{bmatrix} \lambda_1 I_{q_1} & & 0 \\ & \ddots & \\ 0 & & \lambda_r I_{q_r} \end{bmatrix},$$

where I_{q_i} is a $q_i \times q_i$ unit matrix,

$$\sum_{i=1}^{r} q_i = n,$$

and the matrix G has the form

$$\bar{G} = TG = \begin{bmatrix} \bar{G}^{(1)} \\ \hline \vdots \\ \hline \bar{G}^{(r)} \end{bmatrix} \begin{matrix} \} \ q_1 \text{ rows} \\ \\ \} \ q_r \text{ rows.} \end{matrix}$$

Conclusion: The system is completely controllable if and only if

$$\text{rank } \bar{G}^{(1)} = q_1, \cdots, \text{rank } \bar{G}^{(r)} = q_r. \tag{6.8}$$

We leave it to the reader to dualize this result to complete observability.

EXAMPLE 5. Consider the special case $q_1 = \cdots = q_r = 1$ of Lemma 5. The eigenvalues of F are then distinct. If condition (6.8) is satisfied, every element of the one-column matrix $\bar{G}$ is nonzero; by a trivial transformation, all of these elements can be made equal to 1, without affecting $\bar{F}$. Thus we can choose a coordinate system in which F, G have the representation:

$$\bar{F} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} (\lambda_i = \lambda_j \Rightarrow i = j), \bar{G} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{6.9}$$

This is the canonical form of Lur'e [18]. It is closely related to the partial-fraction expansion of transfer functions. To illustrate this, consider the 1×1 transfer-function matrix:

$$z_{11}(s) = \frac{s+2}{(s+1)(s+3)(s+4)} = \frac{1/6}{s+1} + \frac{1/2}{s+3} - \frac{2/3}{s+4}.$$

This transfer function is realized by the system:

$$F = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -4 \end{bmatrix}, \tag{6.10}$$

$$G = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{6.11}$$

and

$$H = [\tfrac{1}{6} \quad \tfrac{1}{2} \quad -\tfrac{2}{3}] \tag{6.12}$$

which is in the canonical form of Lur'e.

By Lemma 5, (6.10–11) is completely controllable; by the dual of Lemma 5, (6.10–12) is completely observable.

We can double-check these facts by means of Lemmas 3–4. For (6.9) the matrix (6.5) is

$$\begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} \end{bmatrix}, \tag{6.13}$$

where the λ_j are the diagonal elements (= eigenvalues) of F in (6.9). But the determinant of (6.13) is the well-known Vandermonde determinant. The latter is nonzero if and only if all the λ_i are distinct, which is what we have assumed.

7. Computation of the canonical structure. We show now how to determine explicitly the change of coördinates which reduces F, G, H to the canonical form. We consider only the constant case of (2.1–2). The computations are elementary; it is not necessary to diagonalize the matrix F or even to determine its eigenvalues.

The procedure is as follows:

(a) We compute the controllability matrix $W = W(0, 1)^*$ given by

* It can be shown [4, Theorem 10] that in the constant case one may choose any $t_1 > t_0$ in Lemma 1.

(6.1); for instance, by solving the differential equation (6.2). Then we find a nonsingular matrix T such that

$$T'WT = E = \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix} \tag{7.1}$$

where I_{n_1} is the $n_1 \times n_1$, $0 \leqq n_1 \leqq n$, unit matrix and the 0's are zero matrices of appropriate size. Clearly $n_1 = n_A + n_B$ is the number of controllable state variables.

The matrix T defines the change of coördinates

$$x = T\bar{x}; \tag{7.2}$$

in terms of the new coördinates, the system matrices are

$$\bar{F} = T^{-1}FT, \qquad \bar{G} = T^{-1}G, \qquad \bar{H} = HT, \qquad \bar{W} = E. \tag{7.3}$$

$$\bar{x} = \begin{bmatrix} \bar{x}^1 \\ \bar{x}^2 \end{bmatrix}, F = \begin{bmatrix} \bar{F}^{11} & \bar{F}^{12} \\ 0 & \bar{F}^{22} \end{bmatrix}, \bar{G} = \begin{bmatrix} \bar{G}^1 \\ 0 \end{bmatrix}, \text{ and } H = [\bar{H}^1 \quad \bar{H}^2]. \tag{7.4}$$

This decomposition is trivial (and therefore omitted) if $n_1 = n$, i.e., when the system is completely controllable.

(b) Next we consider the two subsystems defined by

$$\begin{aligned} &\bar{F}^{11}, \bar{G}^1, \quad \text{and} \quad \bar{H}^1; \\ &\bar{F}^{22}, 0, \quad \text{and} \quad \bar{H}^2. \end{aligned} \tag{7.5}$$

We compute the observability matrices $\bar{M}^1 = \bar{M}^1(0, 1)$ and $\bar{M}^2 = \bar{M}^2(0, 1)$ given by (6.3) for both of these subsystems. Then we determine two nonsingular matrices $\bar{U}^1$, $\bar{U}^2$ such that

$$(\bar{U}^1)'\bar{M}^1\bar{U}^1 = \bar{E}^1 = \begin{bmatrix} 0 & 0 \\ 0 & I_{n_B} \end{bmatrix}, \tag{7.6}$$

and

$$(\bar{U}^2)'\bar{M}^2\bar{U}^2 = \bar{E}^2 = \begin{bmatrix} 0 & 0 \\ 0 & I_{n_d} \end{bmatrix}. \tag{7.7}$$

These results define another change of coördinates

$$\bar{x} = \begin{bmatrix} \bar{x}^1 \\ \bar{x}^2 \end{bmatrix} = \bar{U}\tilde{x} = \begin{bmatrix} \bar{U}^1 & 0 \\ 0 & \bar{U}^2 \end{bmatrix} \cdot \begin{bmatrix} \tilde{x}^1 \\ \tilde{x}^2 \end{bmatrix}.$$

One or the other of these transformations is superfluous if $n_B = n_1$ or $n_d = n - n_1$.

After the coördinate changes (7.2) and (7.8), we obtain the following

matrices

$$(7.9)\qquad \tilde{x} = \begin{bmatrix} x^A \\ x^B \\ \hline x^c \\ x^d \end{bmatrix}, \qquad \tilde{F} = \bar{U}^{-1}\bar{F}\bar{U} = \left[\begin{array}{cc|cc} F^{AA} & F^{AB} & F^{Ac} & F^{Ad} \\ 0 & F^{BB} & \tilde{F}^{Bc} & F^{Bd} \\ \hline 0 & 0 & \tilde{F}^{Bc} & F^{cd} \\ 0 & 0 & 0 & F^{dd} \end{array}\right],$$

$$\bar{U}^{-1}G = \tilde{G} = \begin{bmatrix} G^A \\ G^B \\ \hline 0 \\ 0 \end{bmatrix}, \qquad \tilde{H} = \bar{H}\bar{U} = [0 \quad H^B \quad 0 \quad H^d],$$

$$\tilde{M}^1 = \bar{E}^1, \qquad \tilde{M}^2 = \bar{E}^2.$$

Clearly, n_B is the number of state variables which are both controllable and observable. But, in general, $n_d < n_D$ and $n_c > n_C$.

(c) It remains to transform the element $\tilde{F}^{Bc}$ into 0, if this is not already the case. (If $\tilde{F}^{Bc} = 0$, then $n_c = n_C$, $n_d = n_D$ and (7.9) has the desired canonical structure.)

We consider the subsystem

$$(7.10)\qquad \tilde{F}^* = \left[\begin{array}{c|c} F^{BB} & \tilde{F}^{Bc} \\ \hline 0 & \tilde{F}^{cc} \end{array}\right], \quad \tilde{G}^* = \begin{bmatrix} G^B \\ \hline 0 \end{bmatrix}, \quad \text{and} \quad \tilde{H}^* = [H^B \mid 0].$$

The corresponding observability matrix given by (6.3) is

$$\tilde{M}^*(0,1) = \tilde{M}^* = \left[\begin{array}{c|c} I_{n_B} & A \\ \hline A' & Q \end{array}\right], \quad (Q = Q' \text{ nonnegative definite.})$$

(The upper left element of $\tilde{M}^*$ is I_{n_B} in view of (7.9); all we know about the other elements is their symmetry properties.) Letting

$$\tilde{V}^* = \left[\begin{array}{c|c} I_{n_B} & -A \\ \hline 0 & I_{n_c} \end{array}\right],$$

we find that

$$(\tilde{V}^*)'\tilde{M}^*\tilde{V}^* = \tilde{M}^* = \left[\begin{array}{c|c} I_{n_B} & 0 \\ \hline 0 & R \end{array}\right],$$

where $R = Q - A'A$ is a symmetric, nonnegative-definite matrix.

Now let $\tilde{V}^{**}$ be a nonsingular matrix such that

$$(\tilde{V}^{**})'\tilde{M}^{**}\tilde{V}^{**} = \left[\begin{array}{c|cc} I_{n_B} & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & I_{n_e} \end{array}\right],$$

where $n_e = \text{rank } R$. Let $\tilde{V} = \tilde{V}^{*}\tilde{V}^{**}$. Since $\tilde{V}^{*}$ and $\tilde{V}^{**}$ are upper triangular relative to the partitioning in (7.10), so is $\tilde{V}$, which will take $\tilde{F}^{*}$ into the upper triangular form

$$\tilde{V}^{-1}\tilde{F}^{*}\tilde{V} = \left[\begin{array}{c|cc} F^{BB} & F^{BC} & F^{Be} \\ \hline 0 & F^{CC} & F^{Ce} \\ 0 & F^{eC} & F^{ee} \end{array}\right].$$

where $n_C = n_c - n_e$. But these transformations decompose $\tilde{F}^{*}$ into a completely observable and an unobservable part. Hence $F^{BC} = F^{eC} = 0$. Moreover,

$$\tilde{H}^{*}\tilde{V} = [H^{B} \;\vdots\; 0]\tilde{V} = [H^{B} \;\vdots\; 0 \quad H^{e}]$$

THEOREM 9. *The explicit transformation which takes the constant matrices F, G, and H into the canonical form required by Theorem* (5-iii) *is given by $x \to \tilde{V}^{-1}\tilde{U}^{-1}\tilde{T}^{-1}x$. We partition*

$$F^{Ac} = [F^{AC} \quad F^{Ae}],$$

and partition

$$F^{cd} = \begin{bmatrix} F^{Cd} \\ F^{ed} \end{bmatrix}.$$

Then we define $n_D = n_d + n_e$ and find

$$F^{AD} = [F^{Ae} \quad F^{Ad}],$$

$$F^{BD} = [F^{Be} \quad F^{Bd}],$$

$$F^{CD} = [F^{Ce} \quad F^{Cd}],$$

$$F^{DD} = \begin{bmatrix} F^{ee} & F^{ed} \\ 0 & F^{dd} \end{bmatrix},$$

$$H^{D} = [H^{e} \quad H^{d}].$$

8. Construction of irreducible realizations.

Now we give an explicit procedure for the construction of an irreducible realization of a weighting-function matrix $W(t - \tau)$. In view of Theorem 7,

part (iii), we can do this in two stages:

(I) We construct a realization of W, then

(II-A) we prove, using Lemmas 1–5, that the resultant system is completely controllable and completely observable, hence irreducible; or

(II-B) we carry out explicitly the canonical decomposition and remove all parts other than (B).

Instead of the weighting-function matrix W, it is usually more convenient to deal with its Laplace transform Z.

Let us consider the problem with Method A in order of increasing difficulty.

Case 1. $m = p = 1$. This is equivalent to the problem of simulating a single transfer function on an analog computer. There are several well-known solutions. They may be found in textbooks on classical servomechanism theory or analog computation.

Without loss of generality (see Theorem 4) we may consider transfer functions of the form

$$z_{11}(s) = \frac{a_n s^{n-1} + \cdots + a_1}{s^n + b_n s^{n-1} + \cdots + b_1} = \frac{N(s)}{D(s)} \tag{8.1}$$

where the $a_n, \cdots, a_1$; $b_n, \cdots, b_1$ are real numbers. Of course, at least one of the a_i must be different from zero. We assume also that the numerator $N(s)$ and denominator $D(s)$ of $z_{11}(s)$ have no common roots.

There are two basic realizations of (8.1). See Figs. 5–6, where the standard signal-flow-graph notation [16] is used. In either case, one verifies almost by inspection that the transfer functions relating y_1 to u_1 are indeed given by z_{11}.

In Fig. 5, the system matrices are

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -b_1 & -b_2 & -b_3 & \cdots & -b_{n-1} & -b_n \end{bmatrix}, \tag{8.2}$$

$$G = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{8.3}$$

and

$$H = [a_1 \quad a_2 \quad \cdots \quad a_{n-1} \quad a_n]. \tag{8.4}$$

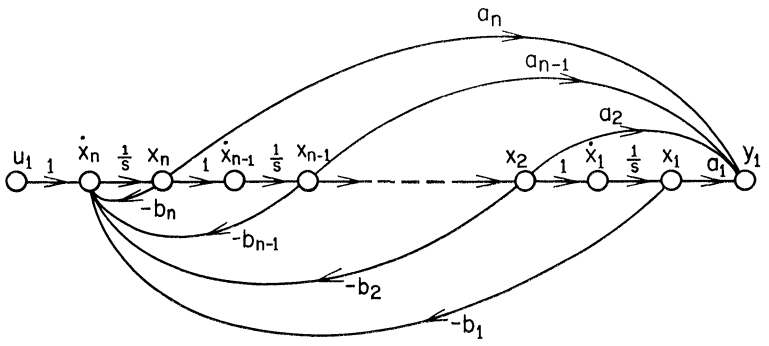

FIGURE 5.

In Fig. 6, the system matrices are

$$F = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -b_1 \\ 1 & 0 & 0 & \cdots & 0 & -b_2 \\ 0 & 1 & 0 & \cdots & 0 & -b_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -b_n \end{bmatrix}, \tag{8.5}$$

$$G = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix}, \tag{8.6}$$

and

$$H = [0 \quad 0 \quad 0 \quad \cdots \quad 1]. \tag{8.7}$$

It is very easy to check by means of (6.5) and (6.6) that the system (8.2, 3) is completely controllable and (8.5, 7) is completely observable.

However, if we attempt to check the controllability of (8.5, 6) by means of (6.5) we get a matrix whose elements are complicated products of the coefficients of $N(s)$ and $D(s)$. To prove that the determinant of this matrix does not vanish, we have only one fact at our disposal: the assumption that $N(s)$ and $D(s)$ have no common roots. Guided by this observation, we find that the following is true:

LEMMA 7. *Suppose F has the form* (8.5) *and G has the form* (8.6). *Then* (i) *we have the relation*

$$K(F, G) = [G \quad FG \quad \cdots \quad F^{n-1}G] = N(F), \tag{8.8}$$

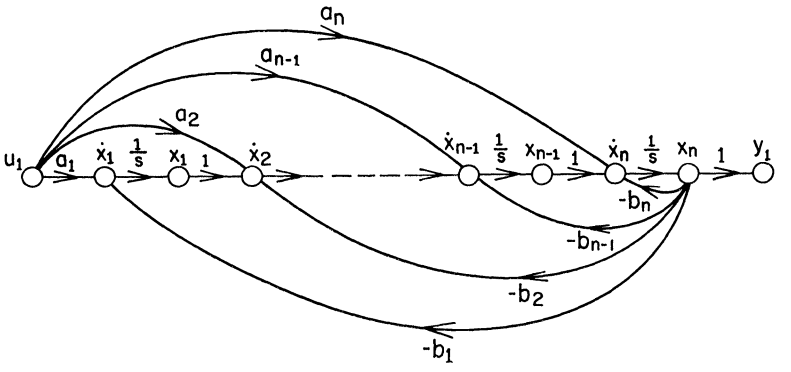

FIGURE 6.

and (ii) the polynomials $N(s)$ and $D(s)$ have no root in common if and only if $\det K(F, G) \neq 0$.

The main fact to be proved is (ii), for then the complete controllability of (8.5, 6) follows by Lemma 3. A straightforward way of establishing (ii) is to transform the standard Euler-Sylvester determinantal criterion [19, p. 84] for the nonexistence of common roots of $N(s)$ and $D(s)$ (the so-called *resolvent* of $N(s)$ and $D(s)$) into the form (8.8). This can be easily done, but the details are not very transparent. Therefore we prefer to give another

Proof. Let e_i, $i = 1, \cdots, n$, be the set of n-vectors in which the j-th component of e_i is δ_{ij}. Since F is given by (8.5), we see that $e_{i+1} = Fe_i$, $1 \leqq n - 1$, and $K(F, e_1) = [e_1, e_2, \cdots, e_n] = I$. Hence $K(F, e_i) = K(F, F^{i-1}e_1) = F^{i-1}K(F, e_1) = F^{i-1}$, when $1 \leqq i \leqq n$. Then (8.8) follows by linearity.

Let $\lambda_i[A]$, $i = 1, \cdots, n$ denote the eigenvalues (not necessarily distinct) of a square matrix A. Then

$$\det K(F, G) = \prod_{i=1}^{n} \lambda_i[N(F)] = \prod_{i=1} N(\lambda_i[F]),$$

where the second equality follows from (8.8) by a well-known identity in matrix theory. Thus $\det K(F, G) = 0$ if and only if $N(\lambda_i[F]) = 0$ for some i; that is, when an eigenvalue of F is a root of $N(\lambda)$. Since the eigenvalues of F are roots of $D(\lambda)$, this proves (ii).*

It is interesting that (8.8) provides a new representation for the resolvent, which is preferable in some respects to the Euler-Sylvester determinant. The latter is a $2n \times 2n$ determinant, whereas $\det K(F, G)$ is $n \times n$.

The complete observability of (8.2, 4) is proved similarly.

The systems given by (8.2–4) and (8.5–7) are duals of one another in

* The present proof of Lemma 6 was suggested by Drs. John C. Stuelpnagel and W. M. Wonham of RIAS.

the sense defined by (5.3). Fig. 6 is a reflection of Fig. 5 about the vertical axis, with all arrows reversed.

A third type of realization in common use is obtained from the partial-fraction expansion of $z_{11}(s)$ (see Example 5). Note, however, that this requires factorization of the denominator of $z_{11}(s)$, whereas the preceding realizations can be written down by inspection, using only the coefficients of $z_{11}(s)$.

These considerations may be summarized as the following result, which is a highly useful fact in control theory:

THEOREM 10. *Consider a linear constant dynamical system with $m = p = 1$, which is completely controllable and completely observable. Then one may always choose a basis in the state space so that F, G, H have the form* (8.2–4) *or* (*with respect to a different basis*) (8.5–7).

Proof. Let (8.1) be the transfer-function matrix of the given dynamical system. By Theorem 8, the given system is an irreducible realization of (8.1). So are the systems specified by (8.2–4) and (8.5–7). By Theorem (7-ii), all three systems are algebraically equivalent and by constancy (Theorem 3) they are even strictly equivalent.

Extensions of this theorem may be found in [14]. For an interesting application to the construction of Lyapunov functions, see [25].

The procedure described here may be generalized to the non-constant case. Assuming the factorization (4.3) of $S(t, \tau)$ is known (with $m = p = 1$), Batkov [20] shows how to determine the coefficients of the differential equation

$$d^n y_1/dt^n + b_n(t)^{n-1}y_1/dt^{n-1} + \cdots + b_1(t)y_1 = a_n(t)d^{n-1}u_1/dt^{n-1} + \cdots + a_1(t)u_1 . \tag{8.9}$$

Laning and Battin [21, p. 191–2] show how one converts (8.9) into a system of first-order differential equations (2.1) with variable coefficients. We shall leave to the reader the proof of the irreducibility of the realization so obtained.

Case 2-a. $m = 1$, $p > 1$. We have a single-input/multi-output system. We can realize $Z(s)$, without factoring the denominators of its transfer functions, by the following generalization of the procedure given by Fig. 5 and (8.2–4).

First, we find the smallest common denominator of the elements of $Z(s)$. (This can be done, of course, without factorization.) $Z(s)$ assumes the form

$$Z(s) = \frac{1}{s^n + b_n s^{n-1} + \cdots + b_1} \begin{bmatrix} a_{1n} s^{n-1} + \cdots + a_{11} \\ \cdots \\ a_{pn} s^{n-1} + \cdots + a_{p1} \end{bmatrix}.$$

Then the following dynamical system provides an irreducible realization

of $Z(s)$: F and G are as in (8.2–3), while H given by (8.4) is generalized to

$$H = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pn} \end{bmatrix}.$$

Complete controllability is trivial; complete observability is established by a straightforward generalization of Lemma 6.

In this case we form p linear functions of the state, rather than merely one as in Fig. 5.

Case 2-b. $m > 1$, $p = 1$. We can realize this multi-input/single-output system analogously to Case 2-a by generalizing the procedure given by Fig. 6 and (8.5–8.7). Let us write the elements of $Z(s)$ in terms of their smallest common denominator:

$$Z(s) = \begin{bmatrix} \dfrac{a_{n1}\, s^{n-1} + \cdots a_{11}}{s^n + b_n\, s^{n-1} + \cdots + b_1} & \cdots & \dfrac{a_{nm}\, s^{n-1} + \cdots + a_{1m}}{s^n + b_n\, s^{n-1} + \cdots + b_1} \end{bmatrix}.$$

Then the desired irreducible realization consists of F and H as defined by (8.5–6), while

$$G = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}.$$

This case is the dual of Case 2-a.

Even in Case 2, it is impractical to give a general formula which expresses the coefficients of F, G, and H in terms of the coefficients of the transfer functions in $Z(s)$ if the denominators are not all the same. When we pass to the general case, determination of F, G, and H often requires extensive numerical computation.

Case 3. m, p arbitrary. Here Method (A) is very complicated if any transfer function in $Z(s)$ has multiple poles [14]. In most practical applications, however, such complications are of no interest. Ruling them out, E. G. Gilbert gave an elegant and relatively simple solution [5].

Let $s_1, \cdots, s_q$ be distinct complex numbers corresponding to the poles of all the elements of $Z(s)$. Assume that all poles are simple. Then

$$R(k) = \lim_{s \to s_k} (s - s_k) Z(s), \quad k = 1, \cdots, q$$

is the k-th residue matrix of $Z(s)$. If $s_\ell = \bar{s}_k$, then $R(s_\ell) = \bar{R}(s_k)$, where the bar denotes the complex conjugate. In terms of the residue matrices, the weighting-function matrix $W(t)$ corresponding to $Z(s)$ has the explicit form

$$W(t) = \mathfrak{L}^{-1}[Z(s)] = \sum_{k=1}^{q} R(k) e^{s_k t}.$$

We have then:

THEOREM 11. (*Gilbert*). *Hypotheses: No element of the transfer-function*

matrix $Z(s)$ has multiple poles. $Z(s)$ has a total of q distinct poles $s_1, \cdots, s_q$, with corresponding residue matrices $R(1), \cdots, R(q)$.

Conclusions: (i) *The dimension of irreducible realizations of $Z(s)$ is*

$$n = \sum_{k=1}^{q} r_k, \text{ where } r_k = \operatorname{rank} R(k). \tag{8.11}$$

(ii) *Write*

$$R(k) = H(k)G(k), \qquad k = 1, \cdots, q, \tag{8.12}$$

where $H(k)$ is a $p \times r_k$ matrix and $G(k)$ is an $r_k \times m$ matrix, both of rank r_k. Then $Z(s)$ has the irreducible realization

$$F = \begin{bmatrix} s_1 I_{r_k} & & 0 \\ & \ddots & \\ 0 & & s_q I_{r_q} \end{bmatrix}, \qquad (I_r = r \times r \text{ unit matrix}), \tag{8.13}$$

$$G = \begin{bmatrix} G(1) \\ \vdots \\ G(q) \end{bmatrix}, \tag{8.14}$$

and

$$H = [H(1) \quad \cdots \quad H(q)]. \tag{8.15}$$

Proof. This is one of the main results in [5]. With the aid of machinery developed here, we can give a shorter (though more abstract) demonstration. The factorization (8.12) is well known in linear algebra. We give in the Appendix various explicit formulae (which are easily machine-computable) for $G(k)$ and $H(k)$. Applying Lemma 5 shows that the dynamical system defined by (8.13–15) is completely controllable and completely observable. Hence it is irreducible, which implies formula (8.11). By elementary changes of variables, (8.13–15) can be transformed into matrices which have only real elements.

A serious disadvantage of Method (A), as expressed by Theorem 11, is that the denominators of the transfer functions in $Z(s)$ must be factored in order to determine the poles. This is not easily done numerically. Moreover, the residue matrices $R(k)$ corresponding to complex poles are complex, which makes the factorization (8.11) more complicated (see Appendix).

Now we turn to Method (B). This method does not require computation of eigenvalues, and it is not bothered by multiple poles. This is a decided advantage in numerical calculations. On the other hand, the method is not convenient for simple illustrative examples. Nor is it possible to display the elements of F, G, and H as simple functions of the coefficients in $z_{ij}(s)$.

An easy way of realizing $Z(s)$ (without guaranteeing irreducibility) is the following. Let α_i be the number of distinct poles (counting each pole with its maximum multiplicity) in the i-th row of $Z(s)$, and let β_i be the number of poles in the i-th column. Then the maximum number n_0 of state variables required to realize $Z(s)$ by repeatedly using the scheme given under Case 2-a or 2-b is

$$n_0 = \min\left\{ \sum_{i=1}^{p} \alpha_i , \quad \sum_{j=1}^{m} \beta_j \right\}.$$

As before, we can determine the α_i and β_j without factoring the transfer functions of $Z(s)$. There is in general no simple way in this method to determine the dimension $n \leqq n_0$ of irreducible realizations without performing the computations outlined in Section 7.

The two methods are best compared via an example. This example must be of fairly high order, since we wish to provide accurate numerical checks.

EXAMPLE 6. Consider the transfer-function matrix

$$Z(s) = \begin{bmatrix} \dfrac{3(s+3)(s+5)}{(s+1)(s+2)(s+4)} & \dfrac{6(s+1)}{(s+2)(s+4)} & \dfrac{2s+7}{(s+3)(s+4)} & \dfrac{2s+5}{(s+2)(s+3)} \\ \dfrac{2}{(s+3)(s+5)} & \dfrac{1}{(s+3)} & \dfrac{2(s-5)}{(s+1)(s+2)(s+3)} & \dfrac{8(s+2)}{(s+1)(s+3)(s+5)} \\ \dfrac{2(s^2+7s+18)}{(s+1)(s+3)(s+5)} & \dfrac{2s}{(s+1)(s+3)} & \dfrac{1}{(s+3)} & \dfrac{2(5s^2+27s+34)}{(s+1)(s+3)(s+5)} \end{bmatrix}.$$

Applying Method (A) first, we find that the residue matrices are:

$$R(1) = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 3 & 1 & 0 & 3 \end{bmatrix}; \qquad r_1 = 3.$$

$$R(2) = \begin{bmatrix} -4.5 & -3 & 0 & 1 \\ 0 & 0 & -6 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}; \qquad r_2 = 2.$$

$$R(3) = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ -3 & -3 & 1 & 1 \end{bmatrix}; \qquad r_3 = 2.$$

$$R(4) = \begin{bmatrix} -0.5 & 9 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}; \qquad r_4 = 1.$$

$$R(5) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -3 \\ 2 & 0 & 0 & 6 \end{bmatrix}; \qquad r_5 = 1.$$

Thus $n = 9$.

Employing the procedure given in the Appendix, we find the following factors for matrices $R(k)$ (the products are accurate up to four places beyond the decimal point):

$$H(1) = \begin{bmatrix} 8.0000 & 0.0000 & 0.0000 \\ 0.0000 & 4.1231 & 0.0000 \\ 3.0000 & 0.7276 & 3.0774 \end{bmatrix},$$

$$G(1) = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.9701 & 0.2425 \\ 0.0000 & 0.3249 & -0.2294 & 0.9175 \end{bmatrix};$$

$$H(2) = \begin{bmatrix} 5.5000 & 0.0000 \\ 0.0000 & 6.0000 \\ 0.0000 & 0.0000 \end{bmatrix},$$

$$G(2) = \begin{bmatrix} -0.8182 & -0.5455 & 0.0000 & 0.1818 \\ 0.0000 & 0.0000 & -1.0000 & 0.0000 \end{bmatrix};$$

$$H(3) = \begin{bmatrix} 1.3416 & 0.4472 \\ 3.1305 & -0.4472 \\ 0.0000 & 4.4721 \end{bmatrix},$$

$$G(3) = \begin{bmatrix} 0.2236 & 0.2236 & 0.6708 & 0.6708 \\ -0.6708 & -0.6708 & 0.2236 & 0.2236 \end{bmatrix};$$

$$H(4) = \begin{bmatrix} 9.0692 \\ 0.0000 \\ 0.0000 \end{bmatrix}, \qquad G(4) = [-0.0551 \quad 0.9924 \quad 0.1103 \quad 0.0000];$$

$$H(5) = \begin{bmatrix} 0.0000 \\ -3.1623 \\ 6.3246 \end{bmatrix}, \qquad G(5) = [0.3162 \quad 0.0000 \quad 0.0000 \quad 0.9487].$$

Using these numerical results, we find that the dynamical equations of the irreducible realization are given by

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 \end{bmatrix},$$

$$G = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.9701 & 0.2425 \\ 0.0000 & 0.3249 & -0.2294 & 0.9175 \\ -0.8182 & -0.5455 & 0.0000 & 0.1818 \\ 0.0000 & 0.0000 & -1.0000 & 0.0000 \\ 0.2236 & 0.2236 & 0.6708 & 0.6708 \\ -0.6708 & 0.6708 & 0.2236 & 0.2236 \\ -0.0551 & 0.9924 & 0.1103 & 0.0000 \\ 0.3162 & 0.0000 & 0.0000 & 0.9487 \end{bmatrix},$$

$$H = \begin{bmatrix} 8.0000 & 0.0000 & 0.0000 & 5.5000 & 0.0000 & 1.3416 & 0.4472 & 9.0692 & 0.0000 \\ 0.0000 & 4.1231 & 0.0000 & 0.0000 & 6.0000 & 3.1305 & -0.4472 & 0.0000 & -3.1623 \\ 3.0000 & 0.7276 & 3.0774 & 0.0000 & 0.0000 & 0.0000 & 4.4721 & 0.0000 & 6.3246 \end{bmatrix}.$$

Now we apply Method (B). First of all we note that $\alpha_1 = \alpha_2 = 4$, $\alpha_3 = 3$, while $\beta_1 = 5$, $\beta_2 = \beta_3 = 4$ (see p. 181). Hence it is best to choose for the preliminary realization three structures of the type discussed under Case 2-b. This will require $n_0 = p(\alpha_1 + \alpha_2 + \alpha_3) = 11$ dimensions.

Next, we find the least common denominator of the rows of $Z(s)$. See Fig. 7.

$$Z(s) = \begin{bmatrix} \frac{3(s^3 + 11s^2 + 39s + 45)}{s^4 + 10s^3 + 35s^2 + 50s + 24} & \frac{6(s^3 + 5s^2 + 7s + 3)}{\cdots} & \frac{2s^3 + 13s^2 + 25s + 14}{\cdots} & \frac{2s^3 + 15s^2 + 33s + 20}{\cdots} \\ \frac{2(s^2 + 3s + 2)}{s^4 + 11s^3 + 41s^2 + 61s + 30} & \frac{s^3 + 8s^2 + 17s + 10}{\cdots} & \frac{2(s^2 + 10s + 25)}{\cdots} & \frac{8(s^2 + 4s + 4)}{\cdots} \\ \frac{2(s^2 + 7s + 18)}{s^3 + 9s^2 + 23s + 15} & \frac{-2(s^2 + 5s)}{\cdots} & \frac{s^2 + 6s + 5}{\cdots} & \frac{2(5s^2 + 27s + 34)}{\cdots} \end{bmatrix}$$

FIGURE 7.

The desired realization of $Z(s)$ can be read off by inspection from Fig. 7, using (8.5) and (8.6):

$$F = \left[\begin{array}{cccc|cccc|cccc} 0 & 0 & 0 & -24 & & & & & & & & \\ 1 & 0 & 0 & -50 & & & & & & & & \\ 0 & 1 & 0 & -35 & & & 0 & & & & 0 & \\ 0 & 0 & 1 & -10 & & & & & & & & \\ \hline & & & & 0 & 0 & 0 & -30 & & & & \\ & & 0 & & 1 & 0 & 0 & -61 & & 0 & & \\ & & & & 0 & 1 & 0 & -41 & & & & \\ & & & & 0 & 0 & 1 & -11 & & & & \\ \hline & & & & & & & & 0 & 0 & -15 & \\ & & 0 & & & & 0 & & 1 & 0 & -23 & \\ & & & & & & & & 0 & 1 & -9 & \end{array}\right],$$

$$G = \left[\begin{array}{rrrr} 135 & 18 & 14 & 20 \\ 117 & 42 & 25 & 33 \\ 33 & 30 & 13 & 15 \\ 3 & 6 & 2 & 2 \\ \hline 4 & 10 & 50 & 32 \\ 6 & 17 & 20 & 32 \\ 2 & 8 & 2 & 8 \\ 0 & 1 & 0 & 0 \\ \hline 36 & 0 & 5 & 68 \\ 14 & -10 & 6 & 54 \\ 2 & -2 & 1 & 10 \end{array}\right],$$

and

$$H = \left[\begin{array}{cccccccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

By virtue of its construction, this system is completely observable but we cannot tell by inspection whether or not it is completely controllable. (From the results obtained above with Method (A), we know that the system is not completely controllable since $11 = n_0 > n = 9$.) Therefore the canonical decomposition may contain Parts (B) and (D).

To see what the dimensions of these parts are, we compute numerically the decomposition of the system into completely controllable and uncontrollable parts according to the method described in Section 8. These calculations involve only the matrices F and G, but the resulting transformations must be applied also to the matrix H.

$$\hat{F} = \left[\begin{array}{rrrrrrrrr|rr} -0.3346 & 0.1182 & 0.0139 & -0.0299 & 0.0097 & -0.0001 & -0.0663 & -0.0113 & 0.0000 & 0.0000 & 0.8334 \\ 0.2455 & -0.2025 & -0.0115 & 0.0268 & -0.0101 & 0.0000 & 0.0734 & 0.0120 & 0.0001 & 0.0000 & 0.7189 \\ -0.8333 & -0.3850 & -0.1023 & 1.0335 & -0.2230 & -0.0005 & -0.9998 & 0.0237 & -0.0120 & 21.5463 & -9.9631 \\ -0.2943 & 0.2032 & 0.0361 & 0.0022 & -0.0610 & -0.0044 & 0.1773 & -0.0569 & 0.0154 & -1.6475 & 0.8726 \\ -0.8896 & 0.8321 & 0.1838 & 0.4999 & -0.4287 & -0.0275 & 1.1199 & 0.0024 & 0.0089 & -1.1882 & 2.5562 \\ 0.2477 & 1.3097 & 2.0439 & -0.3777 & -0.9685 & -0.4965 & -0.4046 & 0.3657 & 0.0199 & -462.2221 & -73.0068 \\ -0.1358 & 0.0429 & 0.0009 & -0.0321 & 0.0315 & 0.0016 & -0.3252 & -0.0152 & -0.0010 & 0.0000 & 0.4698 \\ 1.1634 & -0.0290 & 0.1114 & 1.1645 & -0.4233 & -0.0196 & 0.4057 & -0.0480 & -0.0040 & -25.8717 & -2.1911 \\ -0.2787 & -0.6854 & 2.4604 & 2.4604 & -1.5802 & -0.2801 & 0.8992 & 0.8863 & -0.2649 & -710.3771 & -32.0706 \\ \hline 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.5[illegible]2 & -0.0003 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0401 & -0.2993 \end{array}\right] \times 10$$

FIGURE 8.

$$\hat{G} = \begin{bmatrix} 1.2012 & -0.7886 & -0.3459 & -0.5320 \\ 1.6822 & 0.6039 & 0.3594 & 0.4745 \\ -1.7962 & 0.8640 & 2.5387 & -1.7927 \\ -0.1683 & -0.2420 & 2.2981 & 0.4039 \\ -0.0663 & 0.9596 & 1.4137 & 2.7344 \\ 1.0412 & 1.2734 & 1.5800 & -2.9382 \\ -0.1117 & -0.3946 & -0.0977 & 1.5554 \\ 1.3130 & -2.8589 & 1.9252 & -0.4989 \\ -2.8756 & -1.7637 & 0.9281 & -2.5137 \\ \hline 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix} \times 10$$

$$\hat{H} = \left[\begin{array}{ccccccccc:cc} -0.2928 & 0.4076 & 0.0160 & -0.0373 & 0.0141 & 0.0000 & -0.0981 & -0.0167 & -0.0001 & 0.0000 & 1.0000 \\ -0.0286 & 0.0023 & -0.0116 & -0.0169 & 0.0373 & 0.0022 & -0.0830 & -0.0033 & -0.0007 & 1.0000 & 0.0000 \\ -0.0599 & 0.2173 & 0.0161 & 0.0354 & -0.0603 & -0.0039 & 0.6736 & 0.0178 & 0.0024 & 0.0000 & 0.0000 \end{array}\right]$$

FIGURE 9.

The final results may be seen in Figs. 8–9, which give the matrices $\hat{F}$, $\hat{G}$, and $\hat{H}$. Elements in the lower left-hand corner of $\hat{F}$ should be exactly zero. In fact, they are zero to at least the number of digits indicated in Fig. 8.

To check the accuracy of these two irreducible realizations of the transfer function matrix on p. 181, we have computed the corresponding weighting-function matrices $W^{(1)}(t)$ and $W^{(2)}(t)$. The equality $W^{(1)}(t) = W^{(2)}(t)$ was found to be correct to at least four significant digits.

9. Other applications to system theory. The literature of system theory contains many instances of errors, incomplete or misleading solutions of problems, etc., which can be traced to a lack of understanding of the issues discussed in this paper. This section presents some cases of this known to the writer; other examples may be found in the paper of Gilbert [5].

Analog computers. According to Theorem 8, a linear dynamical system (2.1–2) is a "faithful" realization of an impulse-response matrix if and only if it is irreducible. Suppose the dynamical equations (2.1–2) are programmed on an analog computer. (See [8].) Then it is clear from Theorem 8 that *the computer will simulate the impulse-response matrix correctly if and only if a minimal number of integrators are used.* Otherwise the system programmed on the analog computer will have, besides Part (B), at least one of the Parts (A), (C), or (D). Since the impulse-response matrix determines Part (B), and that alone, the nature of the redundant parts will depend not on the impulse-response matrix but on the particular method used to ob-

tain the dynamical equations. It should be borne in mind that the canonical decomposition is an abstract thing; usually it is not possible to identify the redundant integrators without a change of variables.

The writer is not aware of any book or paper on analog computation where this is explicitly pointed out. But the facts of life seem to be well known (intuitively) to practitioners of the analog art.

That redundancy in the number of integrators used *can* cause positive harm is quite clear from the canonical structure theorem.

EXAMPLE 7. Let the simulated system consist of Parts (A) and (B) and suppose that Part (A) is unstable. Because of noise in the computer, Part (A) will be subject to perturbations; they will be magnified more and more, because of the instability. As long as assumptions of linearity hold exactly, the unstable (A) component of the state vector will not be noticed, but soon the computer will cease to function because its linear range will be exceeded.

Lur'e canonical form. In his book on the Lur'e problem, Letov implies [18; equation (2.4) and (2.23)] that every vector system

$$dx/dt = Fx + g\cdot\sigma \qquad (\sigma = \text{scalar}) \tag{9.1}$$

can be reduced to the canonical form

$$dx_i/dt = \lambda_i x_i + \sigma, \qquad i = 1, \cdots, n \tag{9.2}$$

whenever the eigenvalues λ_i of F are distinct. Since (9.2) is completely controllable, this assertion, if true, would imply that (9.1) is also completely controllable, which is false. In fact, the system defined by

$$F = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \qquad g = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{9.3}$$

is obviously not equivalent to

$$F = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \qquad g = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{9.4}$$

whenever $\lambda \neq \mu$.

In examining the derivation originally given by Lur'e for his canonical form [27; Chapter 1, §2–3], it is clear that the last step before equation (3.5) is valid if and only if $\det [H_k(\lambda_p)] \neq 0$ (in the notation of Lur'e [27].) It is easy to show that this condition is equivalent to complete controllability, whenever the eigenvalues of F are distinct. Unfortunately, the condition $\det [H_k(\lambda_p)] \neq 0$ was not emphasized explicitly by Lur'e [28] in the original publications.

We may thus conclude that *when F has distinct eigenvalues and there is a*

single control variable, the Lur'e-Letov canonical form exists if and only if the pair $\{F, g\}$ is completely controllable.

It is interesting to note that (9.3) can be transformed into (9.4) when $\lambda = \mu$; in other words, when the eigenvalues are not distinct the Lur'e canonical form may exist even if the system is not completely controllable.

Cancellations in the transfer-function. When a mathematical model is derived from physical principles, the equations of the system are in or near the form (2.1–2). Regrettably, it has become widespread practice in system engineering to dispense with differential equations and to replace them by transfer functions $Z(s)$. Later, $Z(s)$ must be converted back into the form (2.1–2) for purposes of analog computation. In the process of algebraic manipulations, some transfer functions may have (exactly or very nearly) common factors in the numerator and denominator, which are then *canceled.* This is an indication that a part of the dynamics of the system is not represented by the transfer function.

Such cancellations are the basic idea of some elementary design methods in control theory. These methods do not bring the system under better control but merely "decouple" some of the undesirable dynamics. But then the closed-loop transfer function is no longer a faithful representation of the (closed-loop) dynamics. Stability difficulties may arise. Similar criticisms may be leveled against the large, but superficial, literature on "noninteracting" control system design.

EXAMPLE 8. Consider the system defined by the matrices

$$F = \begin{bmatrix} 0 & 1 & 0 \\ 5 & 0 & 2 \\ -2 & 0 & -2 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 \\ 0 \\ 0.5 \end{bmatrix}, \qquad H = [-2 \quad 1 \quad 0]. \tag{9.5}$$

The transfer function relating y_1 to u_1 is the sum of two terms:

$$\begin{aligned} \frac{y_1(s)}{u_1(s)} &= -2\frac{x_1(s)}{u_1(s)} + \frac{x_2(s)}{u_1(s)} \\ &= -\frac{2}{s^3 + 2s^2 - 5s - 6} + \frac{s}{s^3 + 2s^2 - 5s - 6} \\ &= \frac{(s-2)}{(s+1)(s-2)(s+3)} = \frac{1}{(s+1)(s+3)}. \end{aligned} \tag{9.6}$$

Thus, by cancellation, the transfer function is reduced from the third to the second order. The system has an unstable "natural mode" (corresponding to $s_3 = 2$) about which the transfer functions gives no information.

Using (6.5) we see that the system (9.5) is completely controllable. By Theorem 5, the system cannot be completely observable: $n_B = 2$ from (9.6) and Case 1, section 8. The canonical structure consists of Parts (A)

and (B). In canonical coördinates the system matrices can be taken as

$$\bar{F} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \qquad \bar{G} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \bar{H} = [0.5 \quad -0.5 \quad 0].$$

We can easily calculate the change of coördinates

$$\bar{x} = Tx$$

by the method of partial fractions discussed in [8]. First we find T^{-1}, then T. The results are

$$T^{-1} = \frac{1}{30}\begin{bmatrix} -5 & 3 & 2 \\ 5 & -9 & 4 \\ 10 & 6 & -1 \end{bmatrix}, \qquad T = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 2 \\ 8 & 4 & 2 \end{bmatrix}.$$

Loss of controllability and observability due to sampling. Consider a single-input/single-output constant linear system. Suppose the output is observed only at the instants $t = kT$ (k = integer, $T > 0$), and that the input is constant over the intervals $kT \leqq t < (k+1)T$. This situation is commonly called "sampling"; it arises when a digital computer is used in control or data processing. T is the *sampling period.* We can regard such a setup as a discrete-time dynamical system. We define here Θ (Axiom (D_1)) as the set of integers and replace (2.1) by a difference equation. All theorems carry over to this situation with small modifications.

The analysis of discrete-time systems by conventional techniques requires the computation of the so-called z-transform of $Z(s)$ [22]. The analysis using z-transforms then proceeds in close analogy with analysis based on Laplace transforms.

A constant linear system which is completely controllable and completely observable will retain these properties even after the introduction of sampling if and only if [4]

$$\text{(9.7)} \qquad \text{Re } s_i = \text{Re } s_j \quad \text{implies} \quad \text{Im } (s_i - s_j) \neq q\pi/T$$

where $i, j = 1, \cdots, n$ and q = positive integer.

If this condition is violated (the sampling process "resonates" with the system dynamics) then cancellations will take place in the z-transform. The z-transform will then no longer afford a faithful representation of the system, so that if (9.7) *is violated, results based on formal manipulation of z-transforms may be invalid.*

This point is not at all clear in the literature. True, Barker [23] has drawn attention to a related phenomenon and called it "hidden oscillation." The textbooks, however, dismiss the problem without providing real insight [22, §5-3; 24, §2.13].

A practical difficulty arises from the fact that near the "resonance" point given by (9.7) it is hard to identify the dynamical equations accurately from the z-transform. Small numerical errors in the computation of the z-transform may have a large effect on the parameters of the dynamical equations.

REFERENCES

[1] R. E. KALMAN, *Discussion of paper by I. Flügge-Lotz*, Proc. 1st International Conference on Automatic Control, Moscow, 1960; Butterworths, London, 1961, Vol. 1, pp. 396–7.

[2] R. E. KALMAN, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci. USA, 48 (1962), pp. 596–600.

[3] R. E. KALMAN, *On the general theory of control systems*, Proc. 1st International Congress on Automatic Control, Moscow, 1960; Butterworths, London, 1961, Vol. 1, pp. 481–492.

[4] R. E. KALMAN, Y. C. HO, AND K. S. NARENDRA, *Controllability of linear dynamical systems*, (to appear in Contributions to Differential Equations, Vol. 1, John Wiley, New York.)

[5] E. G. GILBERT, *Controllability and observability in multivariable control systems*, J. Soc. Indust. Appl. Math. Ser. A: On Control, Vol. 1, No. 2 (1963), pp. 128–151.

[6] V. V. NEMITSKII AND V. V. STEPANOV, *Qualitative Theory Of Differential Equations*, Princeton Univ. Press, Princeton, 1960.

[7] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design via the 'second method' of Lyapunov*, J. Basic Engr. (Trans. A.S.M.E.), 82 D (1960), pp. 371–393.

[8] R. E. KALMAN, *Analysis and design principles of second and higher-order saturating servomechanisms*, Trans. Amer. Inst. Elect. Engrs., 74, II (1955), pp. 294–310.

[9] W. HAHN, *Theorie und Anwendung der direkten Methode von Ljapunov*, Springer, Berlin, 1959.

[10] J. P. LASALLE AND S. LEFSCHETZ, *Stability By Lyapunov's Direct Method*, Academic Press, New York, 1961.

[11] L. MARKUS, *Continuous matrices and the stability of differential systems*, Math. Z., 62 (1955), pp. 310–319.

[12] L. A. ZADEH, *A general theory of linear signal transmission systems*, J. Franklin Inst., 253 (1952), pp. 293–312.

[13] D. MIDDLETON, *An Introduction To Statistical Communication Theory*, McGraw-Hill, New York, 1960.

[14] R. E. KALMAN, *On controllability, observability, and identifiability of linear dynamical systems*, (to appear).

[15] R. E. KALMAN, *On the stability of time-varying linear systems*, Trans. I.R.E. Prof. Gr. Circuit Theory, (CT-9 (1962), pp. 420–422.).

[16] S. J. MASON, *Feedback theory: some properties of signal flow graphs*, Proc. I.R.E., 41 (1953), pp. 1144–56; *Further properties of signal flow graphs*, ibid., 44 (1956), pp. 920–926.

[17] R. E. KALMAN, *New results in filtering and prediction theory*, RIAS Report 61-1, Research Institute for Advanced Studies (RIAS), Baltimore, 1961.

[18] A. M. LETOV, *Stability In Nonlinear Control Systems*, Princeton Univ. Press, Princeton, 1961.

[19] B. L. VAN DER WAERDEN, *Modern Algebra*, Vol. 1, 2nd Ed., Ungar, New York, 1949.
[20] A. M. BATKOV, *On the problem of synthesis of linear dynamic systems with two parameters*, Avtomat. i Telemeh., 19 (1958), pp. 49–54.
[21] J. H. LANING, JR. AND R. H. BATTIN, *Random Processes In Automatic Control*, McGraw-Hill, New York, 1956.
[22] J. R. RAGAZZINI AND G. F. FRANKLIN, *Sampled-Data Control Systems*, McGraw-Hill, New York, 1958.
[23] R. H. BARKER, *The pulse transfer function and its application to sampling servo systems*, Proc. Inst. Elec. Engrs. 99 IV (1952), pp. 302–317.
[24] E. I. JURY, *Sampled-Data Control Systems*, John Wiley, New York, 1957.
[25] R. E. KALMAN, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. USA, 49, (1963), pp. 201–205.
[26] E. F. MOORE, *Gedanken-experiments on sequential machines*, *Automata Studies*, Princeton Univ. Press, Princeton, 1956.
[27] A. I. LUR'E, *Certain Nonlinear Problems in the Theory of Automatic Control.* (in Russian), Gostekhizdat, Moscow, 1951; German translation Akademie-Verlag, Berlin, 1957.
[28] Private communication, Academician A. I. Lur'e.

APPENDIX

Factorization of rectangular matrices. Given an arbitrary, real, $p \times m$ matrix R of rank $q \leqq \min(m, p)$. We wish to find a $p \times q$ matrix H and a $q \times m$ matrix G, both of rank q, such that $R = HG$. The existence of H and G follows almost immediately from the definition of rank. We describe below a constructive procedure for determining H and G numerically from numerical values of R.

Let $p \leqq m$. Form the $p \times p$ matrix $S = RR'$.

As is well known, there exists a nonsingular matrix T such that

$$TRR'T' = TST' = E, \tag{A-1}$$

where precisely q diagonal elements of E are 1, all other elements are 0. T can be calculated by steps similar to the gaussian elimination procedure.

Compute the generalized inverse $R^{\#}$ (in the sense of Penrose [4]) of R. $R^{\#}$ is an $m \times p$ matrix.

Using the properties of $R^{\#}$ ([4]) we obtain

$$R = RR^{\#}R = RR'R^{\#\prime} = SR^{\#\prime} = T^{-1}ET^{-1\prime}R^{\#\prime} = (T^{-1}E)(T^{-1}E)'R^{\#\prime}. \tag{A-2}$$

Now $T^{-1}E$ is a matrix which contains precisely $p - q$ zero columns. Deleting these columns, we obtain a $p \times q$ matrix $(T^{-1}E)^0 = H$. Similarly, deleting $p - q$ zero rows from $(T^{-1}E)'R^{\#\prime} = (R^{\#}T^{-1}E)'$ we obtain a $m \times q$ matrix $G' = (R^{\#}T^{-1}E)^0$. Evidently $R = HG$. Since the ranks of H and G are obviously less than or equal to q, both ranks must be exactly q for otherwise *rank* $R < q$, contrary to hypothesis.

Alternately, let T, U be nonsingular matrices such that

$$TRU = E;$$

then

$$R = (T^{-1}E)^0(EU^{-1})^0 \tag{A-3}$$

is the desired decomposition. However, the computation of (A-3) may require more steps than that of (A-2).

Suppose now that R is complex. Then $S = R\bar{R}' = RR^* = A + iB$ is complex hermitian; it corresponds to the $2n \times 2n$ nonnegative matrix

$$\Sigma = \begin{bmatrix} A & B \\ -B & A \end{bmatrix} \tag{A-4}$$

where $A = A'$ and $B = -B'$. In fact, if $z = x + iy$, the hermitian form z^*RR^*z (which is real-valued) is equal to the quadratic form

$$[x \quad y]\begin{bmatrix} A & B \\ -B & A \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

As is well known, there exists a nonsingular *complex* matrix T such that $TST^* = E$. If $T = U + iV$, it follows further that

$$\begin{bmatrix} U & V \\ -V & U \end{bmatrix}\begin{bmatrix} A & B \\ -B & A \end{bmatrix}\begin{bmatrix} U' & -V' \\ V' & U' \end{bmatrix} = \begin{bmatrix} E & 0 \\ 0 & E \end{bmatrix}.$$

Hence the determination of the complex $n \times n$ matrix has been reduced to the determination of a real $2p \times 2p$ matrix. Similar remarks apply to the calculation of $R^\sharp$. Thus the problem of factoring complex $p \times m$ matrices can be embedded in the problem of factoring real $2p \times 2m$ matrices.

J.S.I.A.M. CONTROL
Ser. A, Vol. 1, No.2
Printed in U.S.A., 1963

OPTIMIZATION AND CONTROL OF NONLINEAR SYSTEMS USING THE SECOND VARIATION*

JOHN V. BREAKWELL†, JASON L. SPEYER‡,
AND ARTHUR E. BRYSON§

Abstract. A feedback control scheme is described that maximizes a terminal quantity while satisfying specified terminal conditions, in the presence of small disturbances. The scheme can also be used in a rapidly converging computation technique to find exact solutions to the nonlinear two-point boundary value problems occurring in the calculus of variations. The scheme is based on a linear perturbation from a nominal optimum path and, as such, involves the second variation of the calculus of variations. A simple analytical example is given for thrust direction control to place a vehicle in orbit. Numerical examples of both the control scheme and the optimization technique are given for a lifting vehicle re-entering the earth's atmosphere at parabolic speed.

1. Introduction. The necessary conditions for the control program of a nonlinear system to maximize a terminal quantity while satisfying specified initial and terminal conditions were set forth in [1]. Numerical computation schemes for determining such extremal control programs were described in [1, 2, and 3]. These control programs are open-loop programs; even slight changes in the initial or final conditions require a new computation of the entire control program. In this paper small changes in the initial and/or final conditions from nominal values are considered which, it is assumed, require only small changes in the control program to preserve optimality. Using the second variation, this results in a linear feedback control scheme or a linear interpolation computation scheme, very similar to the one proposed in [1]. In fact, the second variation provides a quadratic approximation to the terminal quantity being maximized and the differential equations describing perturbations about the nominal optimum path are linear with time-varying coefficients. Such linear systems with quadratic performance indices have been treated extensively in recent years [see, for example, 4 and 5]. An interesting feature of the present scheme is that the weighting matrices in the quadratic performance index are determined as second partial derivatives of the variational Hamiltonian and the terminal quantity being optimized, evaluated on the nominal optimum path (see Appendix A).

The control scheme and the optimization technique can be extended to

* Received by the editors March 22, 1962 and revised February 14, 1963.

† Mechanical and Mathematical Sciences Laboratory, Lockheed Missiles and Space Co., Palo Alto, California.

‡ Space and Information Systems Division, Raytheon Co., Bedford, Mass.

§ Division of Engineering and Applied Physics, Harvard University, Cambridge, Mass.

include inequality constraints on the control and/or state variables based on the work in [6].

Admittedly, many of the operations involved in obtaining the results in section 3 are purely formal; however, with care, it is believed that these steps are justifiable with the assumptions made. Furthermore, these methods have been extensively tested on a number of practical problems.

During the period from submission of the paper to revision for publication, Kelley [7] has presented similar results for the control problem.

2. A neighboring optimum control scheme. The control scheme proposed is identical in form to the one proposed in [5]. The only difference lies in the determination of the matrix of time-varying gains (Λ-matrix) used to multiply the state variable error vector to produce the control variable deviation vector.

Let the differential equations describing the system be

$$\dot{x} = f(x, u, t), \tag{2.1}$$

where x is a column vector of n state variables, f is a column vector of n known functions, u is a column vector of m control variables, t is the independent variable (often time), and $(\dot{}) = d(\ )/dt$.

A set of initial conditions $x(t_0)$ and a set of control variable programs $u(t)$ usually specify a path uniquely. We assume that a set of nominal control variable programs $u^*(t)$ has been determined that maximizes a terminal quantity

$$J = \phi[x(t_f), t_f], \tag{2.2}$$

and yields the specified values of certain other terminal quantities

$$\psi[x(t_f), t_f] \tag{2.3}$$

where ψ is a column vector of q known functions ($q \leqq n - 1$), and t_f is the terminal value of the independent variable. Note that t_f may be specified explicitly or implicitly in (2.3) or it may be the quantity being maximized (or minimized) (2.2).

We consider small deviations from this nominal optimum path which might be caused by disturbances, and/or small changes in the terminal values, $d\psi_f$. Let the deviations of the state variables be

$$\delta x(t) = x(t) - x^*(t), \tag{2.4}$$

where $(\ )^*$ indicates values along the nominal path and unstarred quantities indicate values observed along the actual path (the observed values will often be estimates based on smoothing the observations up to the present time). We shall call δx the "error vector." We wish to determine small deviations from the nominal optimum control program

$$\delta u(t) = u(t) - u^*(t) \tag{2.5}$$

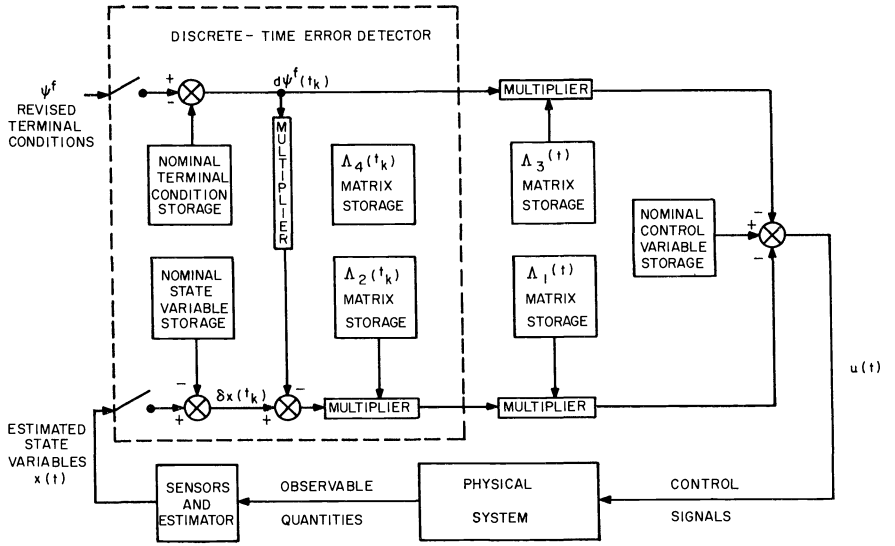

FIG. 1. Block diagram of neighboring optimum terminal control scheme with discrete-time error detection at time t_k, $k = 0, 1, 2, \cdots$. Output of error detector is constant between sampling times. (Note connection missing from Λ_4 to multiplier.)

so that the revised terminal conditions (2.3) are met and ϕ in (2.2) is still maximized when the initial values of the state variables differ from the nominal values of the state variables by $\delta x(t_0)$. It will be assumed in this paper that the nominal optimum control history is continuous and not adjacent to any bound on the controls. It will be shown that a matrix $\Lambda(t_k, t)$ can be pre-calculated so that the control scheme

$$\delta u(t) = -\Lambda(t_k, t)\begin{bmatrix} \delta x(t_k) \\ d\psi_f(t_k) \end{bmatrix} \tag{2.6}$$

yields the desired "neighboring optimum" control program for $t_k \leqq t \leqq t_{k+1}$. Here t_k denotes the k-th sampling time when the error vector is determined. This scheme is shown in a flow chart in Fig. 1 for discrete-time error detection and in Fig. 2 for continuous error detection.

3. Derivation of the feedback gain matrices. The differential equations satisfied by an optimum trajectory are [1]

$$\dot{x} = f(x, u, t), \tag{3.1}$$

$$\dot{\lambda} = -\left(\frac{\partial f}{\partial x}\right)^T \lambda, \tag{3.2}$$

and

$$0 = \left(\frac{\partial f}{\partial u}\right)^T \lambda \tag{3.3}$$

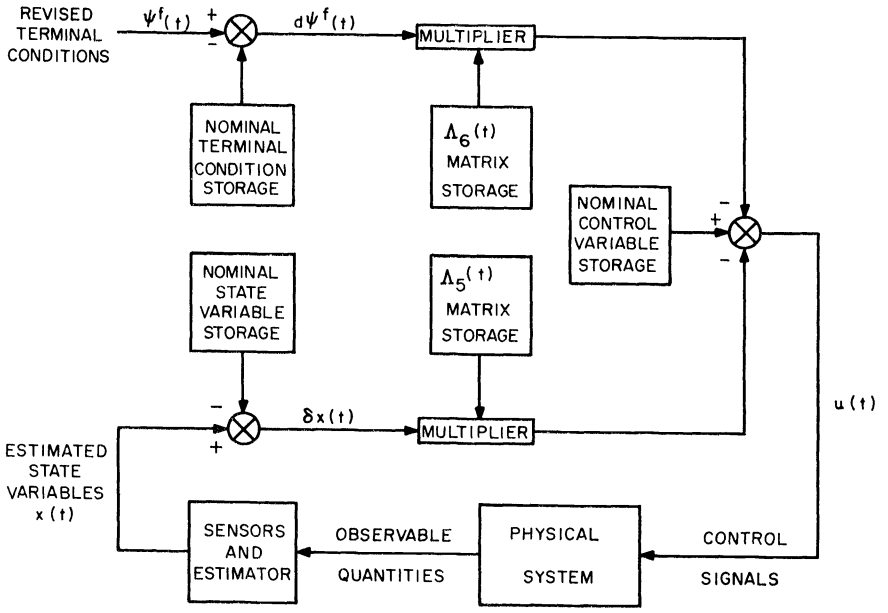

FIG. 2. Block diagram of neighboring optimum terminal control scheme with continuous error detection.

where

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}, \qquad \frac{\partial f}{\partial u} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial u_m} \\ \vdots & & \\ \frac{\partial f_n}{\partial u_1} & \cdots & \frac{\partial f_n}{\partial u_m} \end{bmatrix},$$

and $(\ )^T$ indicates matrix transpose. The initial conditions are:

$$x(t_0) = x^0. \tag{3.4}$$

The terminal conditions for extremalizing $\phi[x(t_f), t_f]$ with the q terminal constraints are:

$$\psi[x(t_f), t_f] = \psi^f, \tag{3.5}$$

$$\lambda(t_f) = \left(\frac{\partial \phi}{\partial x} + \nu^T \frac{\partial \psi}{\partial x}\right)^T_{t=t_f}, \tag{3.6}$$

and

$$\left(\lambda^T f + \frac{\partial \phi}{\partial t} + \nu^T \frac{\partial \psi}{\partial t}\right)_{t=t_f} = 0 \tag{3.7}$$

where

$$\frac{\partial \phi}{\partial x} = \left(\frac{\partial \phi}{\partial x_1}, \cdots, \frac{\partial \phi}{\partial x_n}\right), \qquad \frac{\partial \psi}{\partial x} = \begin{bmatrix} \frac{\partial \psi_1}{\partial x_1} & \cdots & \frac{\partial \psi_1}{\partial x_n} \\ & \cdots & \\ \frac{\partial \psi_q}{\partial x_1} & \cdots & \frac{\partial \psi_q}{\partial x_n} \end{bmatrix},$$

and ν is a column vector of q constants. Equations (3.4)–(3.7) represent $n + q + n + 1$ conditions respectively for the $2n$-th order differential equation system (3.1) and (3.2) and the $q + 1$ unknowns ν and t_f.

Consider a perturbation of this optimal trajectory caused by a perturbation in the initial and/or final conditions (3.4) and (3.5):

$$\frac{d}{dt}(\delta x) - \frac{\partial^2 H}{\partial \lambda \partial x}\delta x = \frac{\partial^2 H}{\partial \lambda \partial u}\delta u, \tag{3.8}$$

$$\frac{d}{dt}(\delta \lambda) + \frac{\partial^2 H}{\partial x^2}\delta x + \frac{\partial^2 H}{\partial x \partial \lambda}\delta \lambda = -\frac{\partial^2 H}{\partial x \partial u}\delta u, \tag{3.9}$$

$$\frac{\partial^2 H}{\partial u \partial x}\delta x + \frac{\partial^2 H}{\partial u \partial \lambda}\delta \lambda = -\frac{\partial^2 H}{\partial u^2}\delta u, \tag{3.10}$$

$$\delta x(t_0) = \delta x^0, \tag{3.11}$$

$$\left(\frac{\partial \psi}{\partial x}\delta x + \frac{D\psi}{Dt}dt_f\right)_{t=t_f} = d\psi^f, \tag{3.12}$$

$$\left\{\delta \lambda - \frac{\partial^2 \Phi}{\partial x^2}\delta x - \left(\frac{\partial \psi}{\partial x}\right)^T d\nu - \left[\frac{\partial H}{\partial x} + \frac{D}{Dt}\left(\frac{\partial \Phi}{\partial x}\right)\right]^T dt_f\right\}_{t=t_f} = 0, \tag{3.13}$$

and

$$\left\{f^T\delta \lambda + \left(\frac{\partial H}{\partial x} + \frac{\partial^2 \Phi}{\partial x \partial t}\right)\delta x + \left(\frac{\partial \psi}{\partial t}\right)^T d\nu + \left[\frac{\partial H}{\partial t} + \frac{D}{Dt}\left(\frac{\partial \Phi}{dt}\right)\right]dt_f\right\}_{t=t_f} = 0 \tag{3.14}$$

where

$$H = \lambda^T f = \text{variational Hamiltonian},$$

$$\Phi = \phi + \nu^T \psi,$$

$$\frac{D(\,)}{Dt} = \frac{\partial(\,)}{\partial t} + \frac{\partial(\,)}{\partial x}f,$$

$$\frac{\partial^2 H}{\partial x \partial u} = \frac{\partial}{\partial u}\left(\frac{\partial H}{\partial x}\right)^T,$$

and

$$\left(\frac{\partial H}{\partial \lambda}\right)^T = f.$$

Now (3.10) can be solved for δu provided $\partial^2 H/\partial u^2$ is nonsingular:

$$\delta u = -\left(\frac{\partial^2 H}{\partial u^2}\right)^{-1}\left[\frac{\partial^2 H}{\partial u \partial x}\,\delta x + \frac{\partial^2 H}{\partial u \partial \lambda}\,\delta\lambda\right]. \tag{3.15}$$

If (3.15) is substituted into (3.8) and (3.9), we have $2n$ coupled first order differential equations in δx and $\delta\lambda$:

$$\frac{d}{dt}\begin{bmatrix}\delta x\\ \delta\lambda\end{bmatrix} = \begin{bmatrix}C_1(t), & C_2(t)\\ C_3(t), & -C_1^T(t)\end{bmatrix}\begin{bmatrix}\delta x\\ \delta\lambda\end{bmatrix}, \tag{3.16}$$

where

$$C_1(t) = \frac{\partial^2 H}{\partial \lambda \partial x} - \frac{\partial^2 H}{\partial \lambda \partial u}\left(\frac{\partial^2 H}{\partial u^2}\right)^{-1}\frac{\partial^2 H}{\partial u \partial x},$$

$$C_2(t) = -\frac{\partial^2 H}{\partial \lambda \partial u}\left(\frac{\partial^2 H}{\partial u^2}\right)^{-1}\frac{\partial^2 H}{\partial u \partial \lambda},$$

and

$$C_3(t) = -\frac{\partial^2 H}{\partial x^2} + \frac{\partial^2 H}{\partial x \partial u}\left(\frac{\partial^2 H}{\partial u^2}\right)^{-1}\frac{\partial^2 H}{\partial u \partial x}.$$

Equations (3.11)–(3.14) represent $n + q + n + 1$ conditions respectively for the $2n$-th order differential equation system (3.16) and the $q + 1$ unknowns $d\nu$ and dt_f. We shall regard (3.11)–(3.16) as *linear* equations, i.e., the coefficients are to be evaluated along the nominal optimum path determined by (3.1)–(3.7). Equations (3.12)–(3.14) are $q + n + 1$ equations in the $2n + q + 1$ unknowns $\delta x(t_f)$, $\delta\lambda(t_f)$, $d\nu$, dt_f. If we assume that the constraints (3.12) are linearly independent and that the transversality condition (3.14) is linearly independent of (3.12) and (3.13), then there are only n independent quantities among the $2n + q + 1$ unknowns. In fact, if the q quantities $d\nu$ and $n - q$ of the quantities δx are specified, the remaining $n + q + 1$ quantities are determined, i.e.

$$[\delta\lambda_1, \cdots, \delta\lambda_n, \delta x_1, \cdots, \delta x_q, dt_f]_{t=t_f} \text{ are linear functions of } d\mu,$$

where

$$d\mu^T = [d\nu_1, \cdots d\nu_q, \delta x_{q+1}, \cdots \delta x_n]_{t=t_f}.$$

We now find $n + q$ solutions to (3.16) with $\delta x(t_f)$ and $\delta\lambda(t_f)$ determined by setting each one of the n components of $d\mu$ and the q components of

$d\psi^f$ equal to unity with the other components zero. Let us call these solutions $X(t, t_f)$ where

$$\begin{bmatrix}\delta x(t)\\ \delta\lambda(t)\end{bmatrix} = \begin{bmatrix}X_{x\mu}(t), & X_{x\psi}(t)\\ X_{\lambda\mu}(t), & X_{\lambda\psi}(t)\end{bmatrix}\begin{bmatrix}d\mu\\ d\psi^f\end{bmatrix}. \tag{3.17}$$

A necessary condition for the existence of neighboring extremal paths is that $X_{x\mu}(t)$ be non-singular over the entire interval $t_0 \leqq t < t_f$ (see [8]). If, at any point $t = t_c$, $X_{x\mu}(t_c)$ is singular, this is called a "conjugate point" in the literature of the calculus of variations [see, for example, [8]). Assuming non-singularity, (3.17) can be inverted at $t = t_0$ to give $d\mu$ in terms of $\delta x(t_0)$ and $d\psi^f$:

$$d\mu = [X_{x\mu}(t_0)]^{-1}[\delta x(t_0) - X_{x\psi}(t_0)\, d\psi_f]. \tag{3.18}$$

Substituting (3.18) into (3.17) we obtain

$$\begin{bmatrix}\delta x(t)\\ \delta\lambda(t)\end{bmatrix} = \begin{bmatrix}X_{x\mu}(t), & X_{x\psi}(t)\\ X_{\lambda\mu}(t), & X_{\lambda\psi}(t)\end{bmatrix}\begin{bmatrix}X_{x\mu}^{-1}(t_0)\delta x_0 - X_{x\mu}^{-1}(t_0)X_{x\psi}(t_0)\, d\psi_f\\ d\psi_f\end{bmatrix}. \tag{3.19}$$

Since the control deviations depend on δx and $\delta\lambda$ from (3.15), it follows from (3.19) that the control deviations are determined by $\delta x(t_0)$ and $d\psi^f$

$$\begin{aligned}\delta u(t) &= -[\Lambda_1(t), \Lambda_3(t)]\begin{bmatrix}\Lambda_2(t_0), & -\Lambda_2(t_0)\Lambda_4(t_0)\\ 0 & I\end{bmatrix}\begin{bmatrix}\delta x(t_0)\\ d\psi_f\end{bmatrix}\\ &= -\Lambda(t, t_0)\begin{bmatrix}\delta x(t_0)\\ d\psi_f\end{bmatrix}\end{aligned} \tag{3.20}$$

where

$$\begin{aligned}\Lambda_1(t) &= \left(\frac{\partial^2 H}{\partial u^2}\right)^{-1}\left[\frac{\partial^2 H}{\partial u \partial x}X_{x\mu} + \frac{\partial^2 H}{\partial u \partial \lambda}X_{\lambda\mu}\right],\\ \Lambda_3(t) &= \left(\frac{\partial^2 H}{\partial u^2}\right)^{-1}\left[\frac{\partial^2 H}{\partial u \partial x}X_{x\psi} + \frac{\partial^2 H}{\partial u \partial \lambda}X_{\lambda\psi}\right],\\ \Lambda_2(t_0) &= [X_{x\mu}(t_0)]^{-1},\\ \Lambda_4(t_0) &= X_{x\psi}(t_0).\end{aligned}$$

4. Sampled data or continuous data. Now, any time $t_k < t_f$, $k = 1, 2, \cdots$, may be considered as an "initial" time as well as t_0. If the error vector δx is estimated at $t = t_k$, the modification of the nominal control program, $\delta u(t)$, is determined for $t_k \leqq t \leqq t_f$. If no further disturbances occur, this modified control program will bring the system to the terminal point with the desired values of $d\psi^f$ and an extremal value of ϕ. Since further disturbances undoubtedly will occur, several sampling times t_0,

t_1, t^2, $\cdots$, may be used (see Fig. 1). Alternatively, δx may be estimated continuously (see Fig. 2), in which case

$$\delta u(t) = -[\Lambda_5(t), \Lambda_6(t)]\begin{bmatrix}\delta x(t)\\ d\psi^f(t)\end{bmatrix}$$

where

$$\Lambda_5 = \Lambda_1(t)\Lambda_2(t),$$

$$\Lambda_6 = \Lambda_3(t) - \Lambda_1(t)\Lambda_2(t)\Lambda_4(t).$$

Revisions in the terminal conditions $d\psi^f(t)$ may be estimated continuously in order to accomplish a desired objective.

5. Comparison with other control schemes. All terminal control schemes based on linearization about an optimal nominal path, which lead to satisfying the terminal constraints, produce the same change in the terminal quantity being maximized *to first order* in δx^0 and $d\psi^{f1}$. The first order terms do *not* depend on the control law and are simply the predictions of first order variational theory.

The present control scheme is the best one *to second order* in δx^0 and $d\psi^f$ (see Appendix A).

6. A neighboring-optimum optimization technique. The numerical solution of the nonlinear two-point boundary value problem (3.1)–(3.7) is often far from trivial. A direct method was presented in [1] and steepest ascent methods were presented in [2, 3]. The control scheme developed in the previous sections is very similar to the direct method of [1]. There, (3.1)–(3.3) were integrated numerically with initial conditions (3.4) and estimated values of $\lambda(t_0)$; the terminal conditions (3.5)–(3.7) were, in general, not satisfied, so the estimated values of $\lambda(t_0)$ were changed by small amounts, one at a time, and (3.1)–(3.3) were integrated n times to determine the effects on the terminal conditions; then a multiple linear interpolation was made to find the correct values of $\lambda(t_0)$ to satisfy the desired terminal conditions. The difficulty with this technique seems to be that the linear interpolation is often inadequate if the trial solution misses the terminal conditions by a substantial amount—in fact, severe instabilities arise quite easily! A simple modification of this procedure seems to cure the instabilities and yields a rapidly convergent computation scheme; if the trial solution misses the terminal conditions by substantial amounts, do *not* try to correct it in one step—instead, correct the solution in *several* small steps, each of which brings the terminal conditions closer to the desired values. Each step determines a neighboring optimal path to the pre-

[1] The authors are indebted to Richard E. Kopp for this observation in November 1961.

vious path, staying sufficiently close to the previous path that the linear interpolation is not stretched beyond the limits of its validity.

Instead of changing the estimated values of the components of $\lambda(t_0)$ by small amounts one at a time and integrating (3.1)–(3.3) n times, it may take less computer time and be more accurate to use the perturbation equations (3.16) and integrate them right along with the trial solution, evaluating the coefficients $C_1(t)$, $C_2(t)$, and $C_3(t)$ on the way. It is necessary to find n solutions to the perturbation equations (3.16) with $\delta x(t_0) = 0$ and each of the n quantities $\delta\lambda(t_0)$ equal to unity with the other components zero. Let us call these solutions $Y(t)$ where

$$\begin{bmatrix} \delta x(t) \\ \delta\lambda(t) \end{bmatrix} = \begin{bmatrix} Y_x(t) \\ Y_\lambda(t) \end{bmatrix} \delta\lambda(t_0). \tag{6.1}$$

If t_f is not given explicitly, some criterion for stopping the integration must be chosen, e.g., when $\psi_1[x(t), t]$ reaches the desired value $\psi_1{}^f$. This determines a nominal value for t_f.

In general $\psi_2, \cdots, \psi_q$ will *not* have the desired values at this nominal value of t_f. From the first q equations of (3.6), nominal values for ν may be found. From the remaining $n - q$ equations of (3.6) and the single equation (3.7), find values for $z_{q+1}, \cdots, z_n$, and $\dot{\Phi}(t_f)$, where

$$\left(\lambda_i - \frac{\partial\phi}{\partial x_i} - \nu^T \frac{\partial\psi}{\partial x_i}\right)_{t=t_f} = z_i, \qquad i = q+1, \cdots, n, \tag{6.2}$$

$$\left(\lambda^T f + \frac{\partial\phi}{\partial t} + \nu^T \frac{\partial\psi}{\partial t}\right)_{t=t_f} = \dot{\Phi}(t_f). \tag{6.3}$$

For a solution to the two-point boundary value problem (3.1)–(3.7), not only must $\psi[x(t_f), (t_f)] = \psi^f$ but also $z = \dot{\Phi}(t_f) = 0$. If these conditions are not met, incorrect values for $\lambda(t_0)$ must have been estimated. To obtain an improved set of values for $\lambda(t_0)$, substitute (6.1) into the perturbation boundary conditions at $t = t_f$, namely (3.12)–(3.14):

$$\begin{bmatrix} \dfrac{\partial\psi}{\partial x} Y_x, & 0, & \dfrac{D\psi}{Dt} \\ Y_\lambda - \dfrac{\partial^2\Phi}{\partial x^2} Y_x, & -\left(\dfrac{\partial\psi}{\partial x}\right)^T, & -\left[\dfrac{\partial H}{\partial x} + \dfrac{D}{Dt}\left(\dfrac{\partial\Phi}{\partial x}\right)\right]^T \\ f^T Y_\lambda + \left(\dfrac{\partial H}{\partial x} + \dfrac{\partial^2\Phi}{\partial x \partial t}\right) Y_x, & \left(\dfrac{\partial\psi}{\partial t}\right)^T, & \dfrac{\partial H}{\partial t} + \dfrac{D}{Dt}\left(\dfrac{\partial\Phi}{\partial t}\right) \end{bmatrix}_{t=t_f} \cdot \begin{bmatrix} \delta\lambda^0 \\ d\nu \\ dt_f \end{bmatrix} = \begin{bmatrix} d\psi^f \\ dz \\ d\dot{\Phi} \end{bmatrix} \tag{6.4}$$

where dz is an n vector whose first q components are zero. The values of $d\psi^f$, dz, and $d\dot{\Phi}$ are chosen as some fraction of the desired change in the terminal conditions, and the matrix (6.4) is inverted to find $\delta\lambda^0$. These are the changes in $\lambda(t_0)$ necessary to change the terminal conditions by the specified amounts. A new forward trajectory is then run with the improved estimates of $\lambda(t_0)$ and the process is repeated until $z = \dot{\Phi}(t_f) = 0$ and $\psi[x(t_f), t_f] = \psi^f$.

Alternatively, the solution may be approached by integrating the equations backwards from $t = t_f$ to $t = t_0$ with the correct terminal conditions ψ^f, and estimated values for the n unknown quantities ν and $x_{q+1}, \cdots, x_n$. Using the perturbation equations (3.16) with the n unit solutions $X_{x\mu}$ and $X_{\lambda\mu}$ of (3.17), improved values for $\mu^T = [\nu_1, \cdots, \nu_q, x_{q+1}, \cdots, x_n]$ can be determined in terms of desired changes in the initial conditions by (3.18) with $d\psi^f = 0$.

The choice of forward or backward integration will depend on the problem; if the terminal conditions are extremely sensitive to variations in the initial conditions and not *vice-versa*, more rapid convergence will be attained by integrating backwards. Such is the case, for instance, in re-entry problems.

It is often very difficult to get the *first* trial solution in new problems where little previous experience is available. In this case steepest ascent methods may have to be used to obtain beginning estimates of the missing boundary conditions.

The present methods have a great advantage over steepest descent methods in that no part of the solution has to be placed in memory storage.

7. An analytical example of the control scheme—thrust direction control for ascent into orbit. Using the approximation of constant gravitational force, we consider the problem of thrust direction control to place a rocket vehicle at a given altitude at a given time with zero vertical velocity and maximum horizontal velocity. The terminal range is not specified. Considering the rocket vehicle as a point mass (see Fig. 3) the equations of motion (neglecting air resistance) are:

$$\dot{v} = a \sin \beta - g, \tag{7.1}$$

$$\dot{u} = a \cos \beta, \tag{7.2}$$

$$\dot{y} = v, \tag{7.3}$$

and

$$\dot{x} = u \tag{7.4}$$

where v = vertical velocity, u = horizontal velocity, y = altitude, x

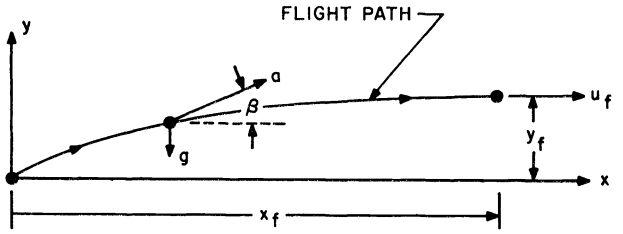

FIG. 3. Geometry and nomenclature of ascent into orbit example problem

= range, a = thrust acceleration, g = gravitational acceleration, and β = thrust direction angle. The initial conditions will be taken as

$$v(0) = u(0) = y(0) = x(0) = 0. \tag{7.5}$$

The first part of the problem is to find $\beta(t)$ to maximize $u_f = u(t_f)$ with

$$\left.\begin{aligned} v(t_f) &= 0 \\ y_f &= y(t_f) \\ t_f &= \text{final time} \end{aligned}\right\} \text{specified.} \tag{7.6}$$

This is done in Appendix B, and the results are:

$$\beta(t) = \tan^{-1}\left\{\tan\beta_0 - (\tan\beta_0 - \tan\beta_f)\frac{t}{t_f}\right\} \tag{7.7}$$

where β_0 and β_f are the initial and final values of β determined by y_f and t_f through the two simultaneous transcendental equations:

$$\sec\beta_0 - \sec\beta_f = \frac{g}{a}(\tan\beta_0 - \tan\beta_f), \tag{7.8}$$

$$\begin{aligned} \frac{y_f}{at_f^2/2} = \frac{1}{(\tan\beta_0 - \tan\beta_f)^2}\Big\{&\tan\beta_f \sec\beta_f - \tan\beta_0 \sec\beta_0 \\ &+ 2\sec\beta_0(\tan\beta_0 - \tan\beta_f) - \log\frac{\tan\beta_0 + \sec\beta_0}{\tan\beta_f + \sec\beta_f}\Big\} \\ &- \frac{g}{a}(\tan\beta_0 - \tan\beta_f)^2. \end{aligned} \tag{7.9}$$

A typical optimal path is shown in Fig. 4, for the case in which $a/g = 3$ and $\dfrac{y_f}{at_f^2/2} = .258$.

The second part of the problem is to determine the feedback gains for neighboring-optimum terminal control. Consider the variation of the

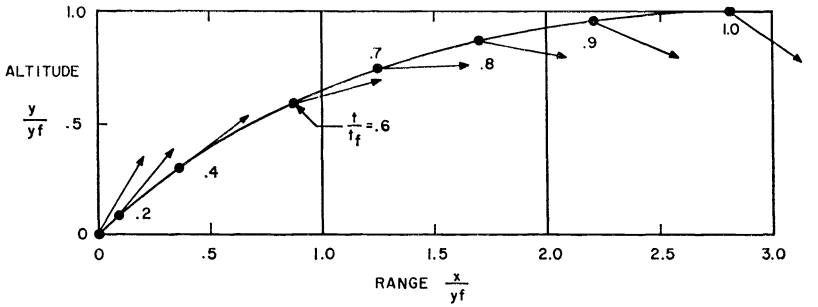

FIG. 4. Optimal ascent trajectory for $a/g = 3$, $y_f/\frac{1}{2}a_f t^2 = .258$

optimum program of (7.7):

$$\sec^2 \beta(\delta\beta) = d\lambda_{v_f} + \left(1 - \frac{t}{t_f}\right) d\bar{\lambda}_y , \tag{7.10}$$

where

$$\lambda_{v_f} = \tan \beta_f , \qquad \bar{\lambda}_y = \tan \beta_0 - \tan \beta_f = \lambda_y t_f .$$

From (7.1) and (7.3)

$$\frac{d}{dt}(\delta v) = a \cos \beta(\delta\beta), \tag{7.11}$$

$$\frac{d}{dt}(\delta y) = \delta v. \tag{7.12}$$

Substituting (7.10) into (7.11) we must solve the two simultaneous differential equations (7.11) and (7.12) with boundary conditions $\delta v = \delta y = 0$ at $t = t_f$. This can be done in closed form, due to the simplicity of the problem, and the result is

$$\begin{bmatrix} \dfrac{\delta v}{a t_f} \\ \dfrac{\delta y}{a t_f^2} \end{bmatrix} = \begin{bmatrix} A_{11}(t), & A_{12}(t) \\ A_{21}(t), & A_{22}(t) \end{bmatrix} \begin{bmatrix} d\lambda_{v_f} \\ d\bar{\lambda}_y \end{bmatrix} \tag{7.13}$$

where

$$A_{11}(t) = \frac{1}{\bar{\lambda}_y}(\sin \beta_f - \sin \beta),$$

$$A_{12}(t) = \frac{-\sec \beta_f}{(\bar{\lambda}_y)^2}[1 - \cos(\beta - \beta_f)],$$

$$A_{21}(t) = \frac{\sec \beta}{(\bar{\lambda}_y)^2}[1 - \cos(\beta - \beta_f)],$$

$$A_{22}(t) = \frac{1}{(\bar{\lambda}_y)^3}\left[\sec \beta \sec \beta_f(\sin \beta - \sin \beta_f) + \log \frac{\tan \beta_f + \sec \beta_f}{\tan \beta + \sec \beta}\right].$$

Next we must invert the expression (7.13):

$$\begin{bmatrix} d\lambda_{v_f} \\ d\lambda_y \end{bmatrix} = \frac{\begin{bmatrix} A_{22}, & -A_{12} \\ -A_{21}, & A_{11} \end{bmatrix}}{D} \begin{bmatrix} \dfrac{\delta v}{at_f} \\ \dfrac{\delta y}{at_f^2} \end{bmatrix}, \tag{7.14}$$

where $D = A_{11}A_{22} - A_{12}A_{21}$ and substitute this relation into the control law (7.10):

$$\delta\beta(t, t_k) = \cos^2\beta^*(t)\left[1, 1 - \frac{t}{t_f}\right] \frac{\begin{bmatrix} A_{22}(t_k), & -A_{12}(t_k) \\ -A_{21}(t_k), & A_{11}(t_k) \end{bmatrix}}{D(t_k)} \begin{bmatrix} \dfrac{\delta v(t_k)}{at_f} \\ \dfrac{\delta y(t_k)}{at_f^2} \end{bmatrix} \tag{7.15}$$

where t_k is the sampling time (i.e., the most recent time when $\delta v(t_k)$ and $\delta y(t_k)$ were measured) and t is the present time. Note $\beta^*(t)$ is the program of β on the nominal path. The data that must be stored to implement the feedback control scheme for discrete-time error detection are:

$$\left.\begin{array}{l} v^*(t_k), \quad y^*(t_k), \\ \dfrac{A_{11}(t_k)}{D(t_k)}, \quad \dfrac{A_{12}(t_k)}{D(t_k)}, \quad \dfrac{A_{21}(t_k)}{D(t_k)}, \quad \dfrac{A_{22}(t_k)}{D(t_k)} \end{array}\right\} \text{at the predetermined sampling times } t_k, \tag{7.16}$$

and β^* at sufficiently frequent intervals to permit interpolation of a continuous function $\beta^*(t)$.

Fig. 5 shows the data necessary to provide neighboring-optimum terminal control for the example of Fig. 4 using discrete-time error detection at five sampling times during the flight (at $t/t_f = 0, .2, .4, .6,$ and $.8$).

Fig. 6 shows part of the data necessary to provide neighboring-optimum terminal control for the example of Fig. 4 using continuous error detection. Also needed would be the nominal time histories $v^*(t)$, $y^*(t)$, and $\beta^*(t)$. All these data and the data of Fig. 6 would have to be stored at sufficiently frequent intervals to permit interpolation of continuous functions.

8. A numerical example of the control scheme—atmospheric re-entry at parabolic speed. Here the control scheme was applied to the problem of guiding a lifting re-entry vehicle to horizontal flight at an altitude of 250,-290 ft. while maximizing the terminal velocity (minimizing the energy loss in the pull-up maneuver). The control problem was assumed to begin when the vehicle had descended to an altitude of 400,000 ft., at which time the velocity was assumed to be close to 36,000 ft. sec.$^{-1}$ and the flight path angle close to $-7.5°$ (nomenclature is shown in Fig. 7).

$$\delta\beta(t,t_k) = \begin{bmatrix} C_1(t), C_2(t) \end{bmatrix} \begin{bmatrix} K_{11}(t_k), K_{12}(t_k) \\ K_{21}(t_k), K_{22}(t_k) \end{bmatrix} \begin{bmatrix} \dfrac{\delta V(t_k)}{a t_f} \\ \dfrac{\delta y(t_k)}{a t_f^2} \end{bmatrix} \quad \begin{cases} \delta\beta = \beta - \beta^* \\ \delta V = V - V^* \\ \delta y = y - y^* \end{cases}$$

t_k/t_f	K_{11}	K_{12}	K_{21}	K_{22}	V^*/at_f	y^*/at_f^2
0	5.56	-11.6	-19.3	-30.9	0	0
.2	5.25	-155	-20.6	-45.3	.097	.0110
.4	5.24	-23.8	-251	-81.6	.173	.0382
.6	6.22	-49.2	-42.8	-230	.199	.0767
.8	6.01	-153	-130	-1420	.142	.1175

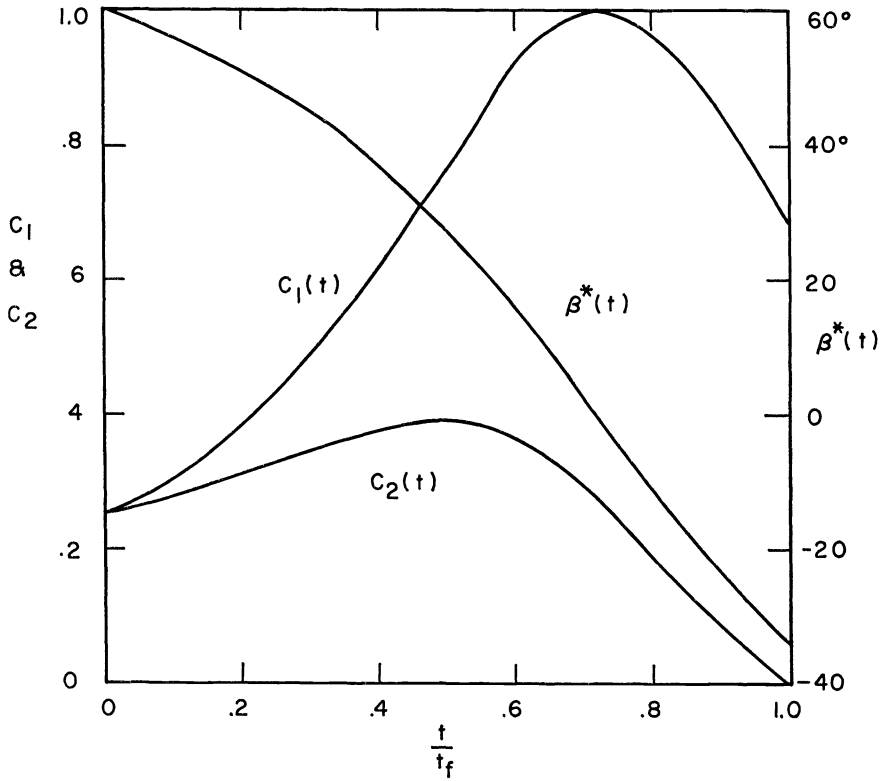

FIG. 5. Feedback gain programs for discrete-time error detection in example problem for $a/g = 3$, $y_f/\frac{1}{2}at_f^2 = .258$.

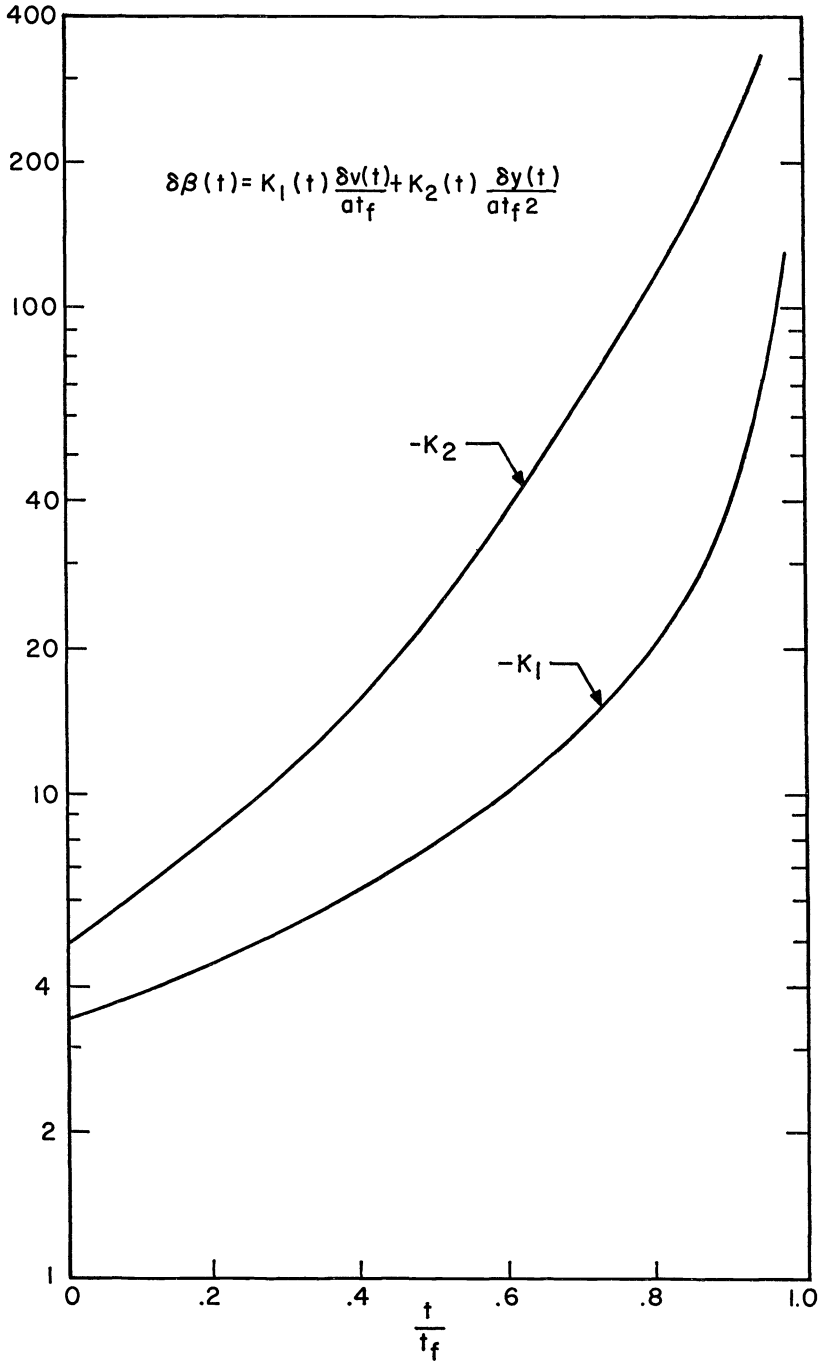

FIG. 6. Feedback gain programs for continuous error detection in example problem for $a/g = 3$, $y_f/\frac{1}{2}at_f^2 = .258$.

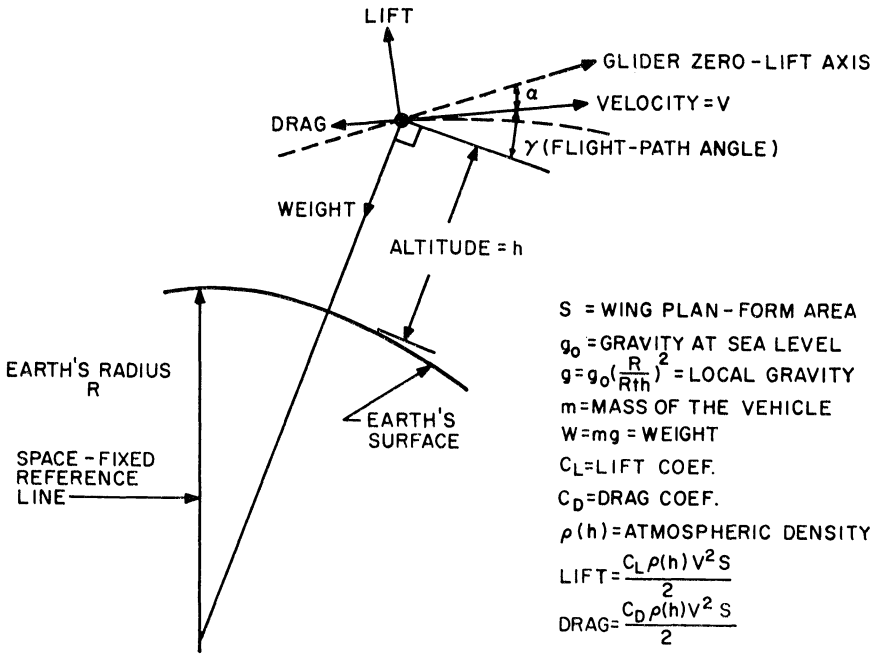

FIG. 7. Geometry and nomenclature of atmospheric re-entry example problem

The wing loading of the vehicle, mg_0/S, was taken as 61.3 lb. ft.$^{-2}$. The 1956 ARDC standard atmosphere model was used. The lift-drag characteristics of the vehicle are shown in Fig. 8. The nonlinear equations of motion are given in Appendix C. The state variable histories for the nominal optimum path (initial velocity exactly 36,000 ft. sec.$^{-1}$, initial flight path angle exactly $-7.5°$) are given in Fig. 9. The maximum terminal velocity for this path is 27,841 ft. sec.$^{-1}$

The control variable in this problem is the angle-of-attack, $\alpha(t)$. Since range (distance along the earth's surface) at the terminal point was not specified and does not appear in the equations of motion, it will not appear in the feedback relation for $\delta\alpha$. The feedback gains for continuous error detection were called K_1, K_2, and K_3 where

$$\delta\alpha = K_1\delta V + K_2\delta\gamma + K_3\delta h. \tag{8.1}$$

These gains were computed and are tabulated in Fig. 9 for every ten seconds on the nominal 238 second flight. Note that the gains in this particular problem are *positive* at the beginning of the flight and then go to negative values. Near the end point the control scheme concentrates on meeting the terminal constraints, i.e., the gains $K2$ and $K3$ tend to large negative values, whereas $K1$ tends to zero.

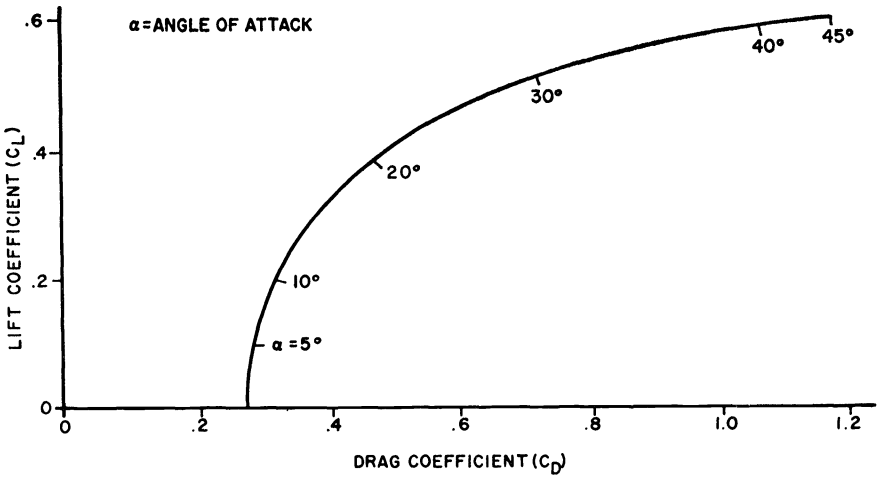

FIG. 8. Lift-drag polar for re-entry vehicle

Time	α	V	γ	h	$K_1 \times 10^3$	K_2	$K_3 \times 10^4$
					$\frac{\text{Deg}}{\text{Ft/Sec}}$	$\frac{\text{Deg}}{\text{Deg}}$	$\frac{\text{Deg}}{\text{Ft}}$
Sec	Deg	Ft/Sec	Deg	Ft			
0	31.69	36,000	-7.50	400,000	.383	1.15	.296
10	31.54	36,039	-7.04	354,424	.342	.859	.252
20	31.33	36,074	-6.57	311,695	.295	.602	.199
30	31.03	36,087	-6.08	271,897	.236	.377	.163
40	30.47	35,939	-5.41	235,623	.165	.098	.248
50	29.06	35,258	-4.13	205,466	.084	-.512	.480
60	25.27	33,807	-1.89	186,828	-.034	-2.01	.459
70	16.73	32,267	.55	183,215	-.293	-5.62	-1.34
80	3.51	31,335	1.81	190,418	-.947	-13.1	-6.72
90	-7.58	30,772	1.92	200,829	-1.96	-22.2	-13.0
100	-14.01	30,331	1.65	210,381	-3.11	-31.1	-17.7
110	-17.78	29,969	1.36	218,268	-4.37	-40.3	-21.7
120	-20.28	29,669	1.11	224,661	-5.74	-50.6	-25.7
130	-22.14	29,418	.92	229,864	-7.20	-62.1	-29.9
140	-23.63	29,202	.76	234,136	-8.72	-75.2	-34.9
150	-24.91	29,013	.64	237,672	-10.3	-90.7	-41.2
160	-26.06	28,843	.53	240,615	-11.9	-109	-49.4
170	-27.10	28,689	.45	243,067	-13.4	-132	-61.1
180	-28.07	28,545	.37	245,101	-14.9	-162	-79.1
190	-28.97	28,411	.30	246,768	-16.3	-204	-109
200	-29.80	28,282	.24	248,101	-17.7	-265	-167
210	-30.57	28,159	.18	249,117	-18.8	-374	-302
220	-31.25	28,038	.11	249,823	-19.9	-602	-746
230	-31.61	27,918	.05	250,213	-20.7	-1490	-4440
238	-31.65	27,841	0	250,290	0	$-\infty$	$-\infty$

$$\delta\alpha = K_1\delta V + K_2\delta\gamma + K_3\delta h$$

FIG. 9. Nominal optimum trajectory and feedback gain programs for continuous error detection in atmospheric re-entry control problem.

To test the control scheme, perturbations in the initial flight path angle, $\delta\gamma(t_0)$, and initial velocity, $\delta V(t_0)$, were introduced. The perturbation in angle-of-attack was obtained from (8.1) and used in the nonlinear differential equations (in place of the "system" in Fig. 1). The integration was stopped when $\gamma = 0$ for the second time. The errors in final altitude are shown in Figs. 10 and 11 plotted against initial flight path angle and initial velocity; within a final altitude error limit of 500 ft., these results indicate an acceptable re-entry corridor of approximately 1.5 degrees in initial flight path angle and 2000 ft. sec.$^{-1}$ in initial velocity, using only *one* nominal optimum trajectory as a reference trajectory. In Fig. 10, note that the control scheme works better for $\delta\gamma(t_0) < 0$ than for $\delta\gamma(t_0) > 0$; this is due, at least in part, to the fact that the final time is greater than the nominal final time when $\delta\gamma(t_0) > 0$, resulting in very large values of the gains, and hence $|\delta\alpha|$, toward the end of the flight. An arbitrary limit of $|\alpha| < 45°$ was introduced to handle this situation; a more satisfactory arrangement might be to "stretch out" the gain histories to fill the predicted time of flight, $t_f - t_0 + dt_f$.

As well as controlling $\gamma(t_f)$ and $h(t_f)$, the control scheme is also supposed to maximize final velocity, $V(t_f)$. How well this aspect was accomplished

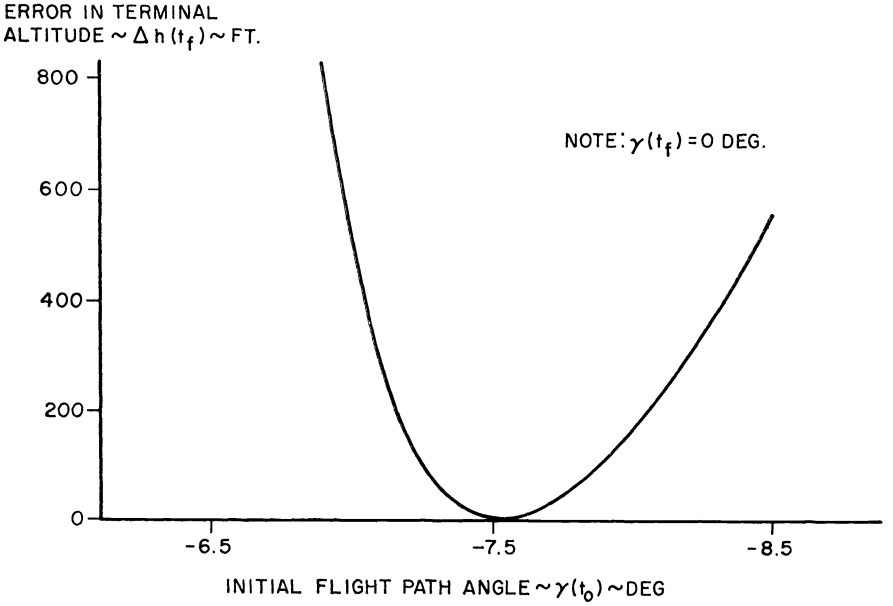

FIG. 10. Error in terminal altitude versus initial flight angle for atmospheric re-entry control problem.

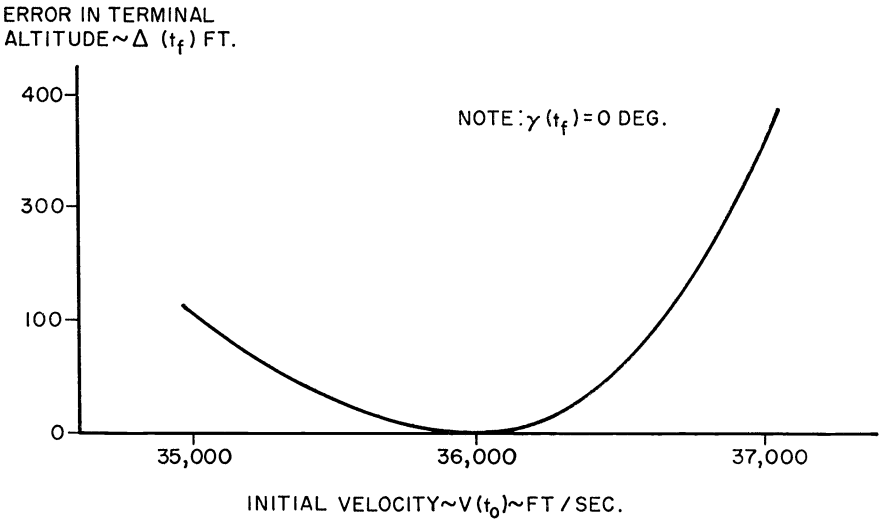

FIG. 11. Error in terminal altitude versus initial velocity for atmospheric re-entry control problem.

is shown in Figs. 12 and 13 where the final velocity is plotted against initial flight path angle, $\gamma(t_0)$, and initial velocity, $V(t_0)$; the exact maximum velocity is also shown in these figures for comparison. The exact maximum final velocity curves were obtained by the optimization technique described in the next section. Again, note that the control scheme works better for $\delta\gamma(t_0) < 0$ than for $\delta\gamma(t_0) > 0$. In fact, for $\delta\gamma(t_0) < 0$ the difference between the "exact" maximum final velocity and the controlled final velocity was so small that the two curves in Fig. 12 are indistinguishable. The same thing is true for $\delta V(t_0) < 0$ in Fig. 13.

9. A numerical example of the optimization technique—atmospheric re-entry at parabolic speeds. Here the neighboring-optimum optimization technique was applied to the problem of determining the nominal optimum trajectory used in the previous section, and also to generate a set of optimum trajectories with slightly different initial conditions. The data and nomenclature are the same as in the previous section; the equations of motion, the Euler-Lagrange equations and the perturbation equations are all presented in Appendix C. The numerical solutions were obtained by the backward-integration technique, discussed in section 6, because we were interested in a parametric study of the initial conditions with final conditions held fixed, namely $\gamma(t_f) = 0$, $h(t_f) = 250{,}290$ ft., $V(t_f)$ maximized.

The power of the optimization scheme lies in its ability to converge

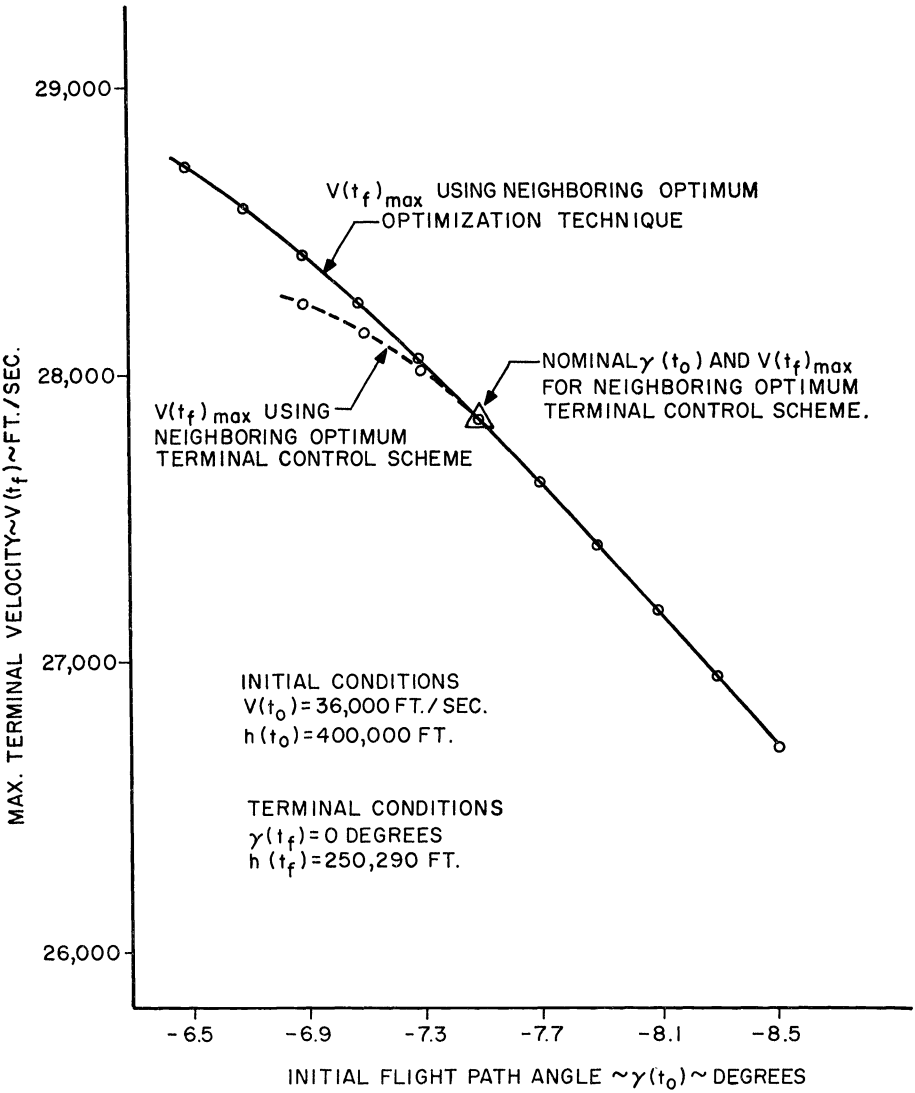

FIG. 12. Terminal velocity versus initial flight path angle for atmospheric re-entry control problem and maximum terminal velocity versus initial flight path angle for atmospheric re-entry optimization problem.

rapidly, to calculate the exact optimum trajectory and to generate a whole family of optimum paths. A family of optimum paths was generated by varying $\gamma(t_0)$ in increments of 0.2 degrees while $V(t_0)$ and $h(t_0)$ were held fixed. The optimization scheme was able to converge to these changes

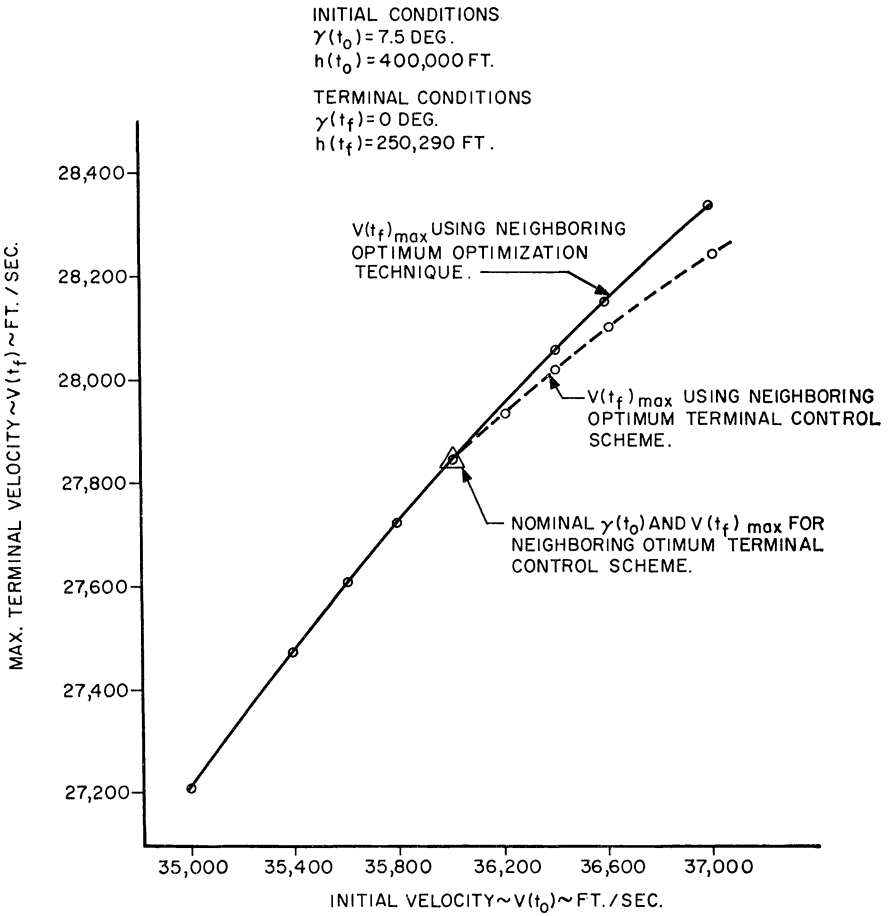

FIG. 13. Terminal velocity versus initial velocity for atmospheric re-entry control problem and maximum terminal velocity versus initial velocity for atmospheric re-entry optimization problem.

in only two iterations. The field of optimum paths is represented in Fig. 12 by $V(t_f)_{\max}$ versus $\gamma(t_0)$. Fig. 13 shows $V(t_f)_{\max}$ versus $V(t_0)$ with $\gamma(t_0)$ and $h(t_0)$ held fixed. Here, the scheme converged to changes in $V(t_0)$ of ± 400 ft./sec. with only two iterations. The control programs generated by the optimization technique for $\gamma(t_0)$ equal to -6.5 deg., -7.5 deg., and -8.5 deg. with $V(t_0) = 36{,}000$ ft./sec. and $h(t_0) = 400{,}000$ ft. are displayed in Fig. 14. These curves are very similar, illustrating the sensitivity of the re-entry trajectory to small changes in α.

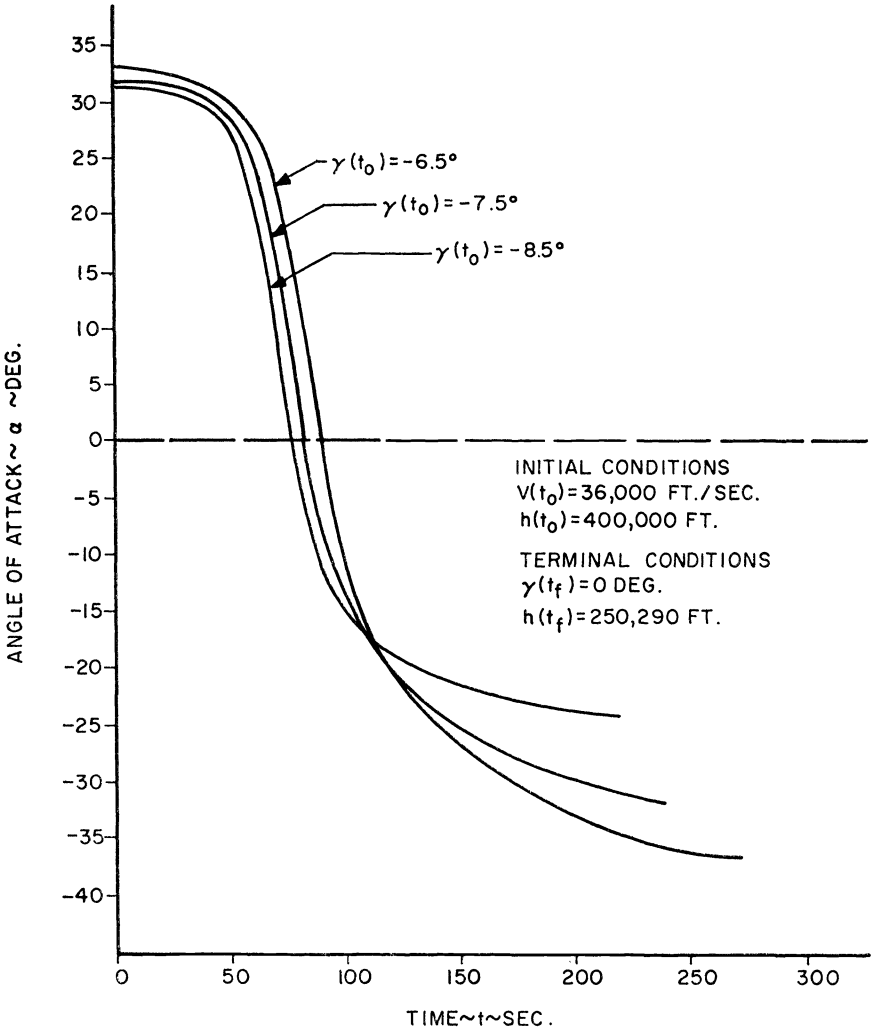

FIG. 14. Angle of attack programs for atmospheric re-entry optimization problems

REFERENCES

[1] J. V. BREAKWELL, *The optimization of trajectories*, J. Soc. Indust. Appl. Math., 7 (1959), pp. 215–247.

[2] H. J. KELLEY, *Gradient theory of optimal flight paths*, J. Amer. Rocket Soc., 30 (1960) pp. 947–953.

[3] A. E. BRYSON AND W. F. DENHAM, *A steepest ascent method for solving optimum programming problems*, J. Appl. Mech., 29, 2 (1962) pp. 247–257.

[4] K. J. ASTRÖM, J. E. BERTRAM, J. J. FLORENTIN, P. D. JOSEPH, AND R. E. KALMAN, *Current status of multivariable linear control system theory*, J. Soc. Indust. Appl. Math. Ser. A: Control. (To appear)

[5] A. E. Bryson and W. F. Denham, *Multivariable terminal control for minimum mean square deviation from a nominal path*, Proc. Inst. Aerospace Sci. Symposium on Vehicle Systems Optimization, Garden City, N. Y., Nov. 1961.

[6] W. F. Denham, and A. E. Bryson, *The solution of optimal programming problems with inequality constraints*, Institute of Aerospace Sciences, Annual Meeting, Jan. 1963. (Also Raytheon Co., Report BR-2121, Nov. 1962.)

[7] H. J. Kelley, *Guidance theory and extremal fields*, Trans. I.R.E., Prof. Group Automatic Control, (1962), pp. 75–82.

[8] G. A. Bliss, *Lectures on the Calculus of Variations*, Univ. of Chicago Press, Chicago, 1946.

APPENDIX

A. Interpretation of the control scheme as a linear system with a quadratic performance index. The derivations of this section are for the most part formal, and are primarily intended to present a heuristic interpretation of the control scheme as a linear system with quadratic performance index. For simplicity we consider only the special case where t_f, as well as t_0, is given. Adjoin the constraints (2.1) and (2.3) to the performance index (2.2) by Lagrange multipliers ν and $\lambda(t)$, as follows:

$$\bar{J} = (\phi + \nu^T \psi)_{t=t_f} - \int_{t_0}^{t_f} \lambda^T [\dot{x} - f(x, u, t)]\, dt. \tag{A.1}$$

Consider the variations of $\bar{J}$ due to a change in control program $\delta u(t)$ and changes in the initial conditions $\delta x(t_0)$ and final conditions $\delta\psi(t_f)$:

$$\begin{aligned}\delta\bar{J} = {} & \left[\frac{\partial\Phi}{\partial x}\,\delta x + \frac{1}{2}\,\delta x^T \frac{\partial^2\Phi}{\partial x^2}\,\delta x + \nu^T\,\delta\psi\right]_{t=t_f} - [\lambda^T\,\delta x]_{t_0}^{t_f} \\ & + \int_{t_0}^{t_f}\left[\left(\dot{\lambda}^T + \lambda^T\frac{\partial f}{\partial x}\right)\delta x + \lambda^T\frac{\partial f}{\partial u}\,\delta u + \frac{1}{2}\,\delta x^T\frac{\partial^2 H}{\partial x^2}\,\delta x\right. \\ & \left. + \frac{1}{2}\,\delta x^T\frac{\partial^2 H}{\partial x\,\partial u}\,\delta u + \frac{1}{2}\,\delta u^T\frac{\partial^2 H}{\partial u\,\partial x}\,\delta u + \frac{1}{2}\,\delta u^T\frac{\partial^2 H}{\partial u^2}\,\delta u\right] dt,\end{aligned} \tag{A.2}$$

where $H = \lambda^T f$ and $\Phi = \phi + \nu^T\psi$. Let us choose $\lambda(t)$ so that

$$\begin{aligned}\dot{\lambda} &= -\left(\frac{\partial f}{\partial x}\right)^T\lambda, \\ \lambda^T(t_f) &= \left(\frac{\partial\Phi}{\partial x}\right)_{t=t_f}.\end{aligned} \tag{A.3}$$

For a first order extremum, we must choose $u(t)$ so that

$$\lambda^T\frac{\partial f}{\partial u} = 0. \tag{A.4}$$

The constants ν must be chosen so that (3.5) is satisfied. We have left then

$$\delta\bar{J} = (\lambda^T \delta x)_{t=t_0} + (\nu^T \delta\psi)_{t=t_f} + \frac{1}{2}\left(\delta x^T \frac{\partial^2 \Phi}{\partial x^2} \delta x\right)_{t=t_f}$$

$$+ \frac{1}{2}\int_{t_0}^{t_f} [\delta x^T, \delta u^T] \begin{bmatrix} \dfrac{\partial^2 H}{\partial x^2}, & \dfrac{\partial^2 H}{\partial x\, \partial u} \\ \dfrac{\partial^2 H}{\partial u\, \partial x}, & \dfrac{\partial^2 H}{\partial u^2} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \end{bmatrix} dt, \tag{A.5}$$

where all the coefficients are evaluated on the nominal optimum path generated by the solution to (3.1)–(3.6). Now, it is apparent from (A.5) that $\lambda^T(t_0)$ and ν^T are the first order influence functions on J for small variations in the initial conditions δx^0 and the final conditions $\delta\psi^f$. We may regard the second order terms on the right hand side of (A.5) as a *quadratic performance index* to be maximized (or minimized) by a choice of $\delta u(t)$ for a neighboring optimum path with slightly different initial and final conditions (δx^0 and $\delta\psi^f$) from the nominal optimum path. The neighboring optimum path must satisfy the perturbation equations

$$\frac{d}{dt}(\delta x) = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial u}\delta u, \tag{A.6}$$

where, again, the coefficients are evaluated on the nominal optimum path; thus equations (A.6) are *linear*. The boundary conditions for this accessory maximum (or minimum) problem are

$$\delta x(t_0) = \delta x^0, \tag{A.7}$$

$$\left(\frac{\partial\psi}{\partial x}\delta x\right)_{t=t_f} = \delta\psi^f. \tag{A.8}$$

Such linear problems with quadratic performance indices have been treated extensively in recent years (see, for example, [4, 5]). Let us adjoin the constraints (A.6) and (A.8) to the performance index (A.5) with Lagrange multipliers $\delta\lambda(t)$ and $d\nu$ as follows:

$$\delta\bar{J} - (\lambda^T \delta x)_{t=t_0} - (\nu^T \delta\psi)_{t=t_f} = \frac{1}{2}\left(\delta x^T \frac{\partial^2\Phi}{\partial x^2}\partial x\right)_{t=t_f} + d\nu^T\left(\frac{\partial\psi}{\partial x}\delta x\right)_{t=t_f}$$

$$+ \int_{t_0}^{t_f} \left\{ \frac{1}{2}[\delta x^T, \delta u^T] \begin{bmatrix} \dfrac{\partial^2 H}{\partial x^2}, & \dfrac{\partial^2 H}{\partial x\, \partial u} \\ \dfrac{\partial^2 H}{\partial u\, \partial x}, & \dfrac{\partial^2 H}{\partial u^2} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta u \end{bmatrix} \right.$$

$$\left. - \delta\lambda^T \left[\frac{d}{dt}(\delta x) - \frac{\partial f}{\partial x}\delta x - \frac{\partial f}{\partial u}\delta u\right] \right\} dt. \tag{A.9}$$

Now consider the variation of $\delta\bar{J}$ (i.e., the second variation, $\delta^2\bar{J}$, due to a change in perturbation control program $\delta^2 u$, for fixed values of δx^0 and $\delta\psi^f$:

$$\begin{aligned}\delta^2\bar{J} = &\left[\left(\delta x^T \frac{\partial^2\Phi}{\partial x^2} + d\nu^T \frac{\partial\psi}{\partial x} - \delta\lambda^T\right)\delta^2 x\right]_{t=t_f} \\ &+ \int_{t_0}^{t_f}\left[\left(\frac{d}{dt}(\delta\lambda^T) + \delta x^T \frac{\partial^2 H}{\partial x^2} + \delta\lambda^T\frac{\partial f}{\partial x}\right.\right. \\ &\left.\left. + \delta u^T \frac{\partial^2 H}{\partial u\,\partial x}\right)\delta^2 x + \left(\delta x^T \frac{\partial^2 H}{\partial x\,\partial u} + \delta\lambda^T \frac{\partial f}{\partial u} + \delta u^T \frac{\partial^2 H}{\partial u^2}\right)\delta^2 u\right] dt.\end{aligned} \tag{A.10}$$

Let us choose $\delta\lambda(t)$ so that

$$\begin{aligned}\frac{d}{dt}(\delta\lambda) &= \frac{\partial^2 H}{\partial x^2}\delta x - \left(\frac{\partial f}{\partial x}\right)^T \delta\lambda - \frac{\partial^2 H}{\partial x\,\partial u}\delta u, \\ \delta\lambda(t_f) &= \left[\frac{\partial^2\Phi}{\partial x^2}\delta x + \left(\frac{\partial\psi}{dx}\right)^T d\nu\right]_{t=t_f}.\end{aligned} \tag{A.11}$$

For an extremum ($\delta^2\bar{J} = 0$ for arbitrary $\delta^2 u$), it is obviously necessary that

$$\frac{\partial^2 H}{\partial u^2}\delta u + \frac{\partial^2 H}{\partial u\,\partial x}\delta x + \left(\frac{\partial f}{\partial u}\right)^T \delta\lambda = 0. \tag{A.12}$$

The constants $d\nu$ must be chosen so that (A.8) is satisfied.

Note that (A.11) and (A.12) are identical to (3.9), (3.10), and (3.13) for $dt_f = 0$. Note also that

$$\left(\frac{\partial f}{\partial u}\right)^T \equiv \frac{\partial^2 H}{\partial u\,\partial\lambda} \quad \text{and} \quad \left(\frac{\partial f}{\partial x}\right)^T \equiv \frac{\partial^2 H}{\partial x\,\partial\lambda}.$$

B. Determination of the nominal optimum path for the ascent into orbit problem. The equations adjoint to equations (7.1)–(7.4) are particularly simple:

$$\dot{\lambda}_v = -\lambda_y, \tag{B.1}$$

$$\dot{\lambda}_u = -\lambda_x, \tag{B.2}$$

$$\dot{\lambda}_y = 0, \tag{B.3}$$

$$\dot{\lambda}_x = 0. \tag{B.4}$$

Since x_f is not specified, $\lambda_x(t_f) = 0$. Since we are maximizing u_f, $\lambda_u(t_f) = 1$.

It follows quite simply that

$$\begin{aligned}\lambda_v &= \lambda_{v_f} + \lambda_y(t_f - t), \\ \lambda_u &= 1, \\ \lambda_y &= \text{constant}, \\ \lambda_x &= 0.\end{aligned} \tag{B.5}$$

The optimality condition is

$$\lambda_v \cos\beta - \lambda_u \sin\beta = 0, \tag{B.6}$$

from which it is seen that the optimum $\beta(t)$ is given by

$$\tan\beta = b - ct, \quad b = \lambda_{v_f} + \lambda_y t_f, \quad c = \lambda_y. \tag{B.7}$$

Substituting this expression for $\beta(t)$ into (7.1)–(7.4) and using the boundary conditions (7.5) we obtain by elementary quadratures

$$v = \frac{a}{c}\left[\sqrt{(1+b^2)} - \sqrt{(1+(b-ct)^2)}\right] - gt, \tag{B.8}$$

$$u = \frac{a}{c}\log\frac{b+\sqrt{(1+b^2)}}{b-ct+\sqrt{(1+(b-ct)^2)}}, \tag{B.9}$$

$$y = \frac{a}{2c^2}\left\{(b-ct)\sqrt{(1+(b-ct)^2)} - b\sqrt{(1+b)^2} - \log\frac{b+\sqrt{(1+b^2)}}{b-ct+\sqrt{(1+(b-ct)^2)}} + 2ct\sqrt{1+b^2}\right\} - \frac{1}{2}gt^2, \tag{B.10}$$

$$x = \frac{a}{c^2}\left[\sqrt{(1+b^2)} - \sqrt{(1+(b-ct)^2)} - (b-ct)\log\frac{b+\sqrt{(1+b^2)}}{b-ct+\sqrt{(1+(b-ct)^2)}}\right]. \tag{B.11}$$

The two constants b and c (and hence, λ_{v_f} and λ_y) are determined by the terminal boundary conditions (7.6):

$$\sqrt{(1+b^2)} - \sqrt{(1+(b-ct_f)^2)} = \frac{g}{a}ct_f, \tag{B.12}$$

$$\frac{y_f}{\frac{1}{2}at_f^2} = \frac{1}{c^2t_f^2}\left\{(b-ct_f)\sqrt{(1+(b-ct_f)^2)} - b\sqrt{(1+b)^2} - \log\frac{b+\sqrt{(1+b^2)}}{b-ct_f+\sqrt{(1+(b-ct_f)^2)}} + 2ct_f\sqrt{(1+b^2)}\right\} - \frac{g}{a}c^2t_f^2. \tag{B.13}$$

Equation (B.12) can be solved for ct_f in terms of b (note t_f is given but c is to be determined):

$$ct_f = \frac{2}{1-g^2/a^2}\left(b - \frac{g}{a}\sqrt{(1+b^2)}\right). \tag{B.14}$$

Substituting (B.14) into (B.13) yields a transcendental equation for b in terms of y_f. Note that

$$\begin{aligned} \tan\beta_0 &= b = \lambda_{v_f} + \lambda_y t_f, \qquad \beta_0 = \beta(0), \\ \tan\beta_f &= b - ct_f + \lambda_{v_f}, \qquad \beta_f = \beta(t_f), \end{aligned} \tag{B.15}$$

$$\tan\beta(t) = \tan\beta_0 - (\tan\beta_0 - \tan\beta_f)\frac{t}{t_f}. \tag{B.16}$$

The maximum value of $u_f = u(t_f)$ is given by

$$\frac{u_f}{at_f} = \frac{1}{ct_f}\log\frac{\tan\beta_0 + \sec\beta_0}{\tan\beta_f + \sec\beta_f}. \tag{B.17}$$

The corresponding value of $x_f = x(t_f)$ is given by

$$\frac{x_f}{\frac{1}{2}at_f^2} = \frac{2}{(ct_f)^2}\left(\sec\beta_0 - \sec\beta_f - \tan\beta_f \log\left(\frac{\tan\beta_0 + \sec\beta_0}{\tan\beta_f + \sec\beta_f}\right)\right). \tag{B.18}$$

Note that the maximum value of y_f in time t_f with $v_f = 0$ is obtained by ascending vertically using

$$\beta(t) = \begin{cases} \frac{\pi}{2}; & 0 < t < \frac{1}{2}\left(1 + \frac{g}{a}\right)t_f \\ -\frac{\pi}{2}; & \frac{1}{2}\left(1 + \frac{g}{a}\right)t_f < t < t_f, \end{cases} \tag{B.19}$$

and this program gives

$$(y_f)_{\max} = \frac{1}{4}\left(1 - \frac{g^2}{a^2}\right)at_f^2. \tag{B.20}$$

C. Equations used in the atmospheric re-entry problems. The equations of motion of a point mass about a spherical non-rotating earth (time being the independent variable) are

$$\dot{V} = -\frac{C_D\,\rho V^2 S}{2m} - g\sin\gamma, \tag{C.1}$$

$$\dot{\gamma} = \frac{C_L\,\rho V S}{2m} + \left(\frac{V}{R+h} - \frac{g}{V}\right)\cos\gamma, \tag{C.2}$$

and

$$\dot{h} = V\sin\gamma. \tag{C.3}$$

The differential equations for the associated Lagrange multipliers are

$$\dot{\lambda}_V = \frac{C_D\,\rho V S}{m}\lambda_V - \left[\frac{C_L\,\rho S}{2m} + \left(\frac{1}{R+h} + \frac{g}{V^2}\right)\cos\gamma\right]\lambda_\gamma - \sin\gamma\,\lambda_h, \tag{C.4}$$

$$\dot{\lambda}_\gamma = g\cos\gamma\,\lambda_V + \left(\frac{V}{R+h} - \frac{g}{V}\right)\sin\gamma\,\lambda_\gamma - V\cos\gamma\,\lambda_h, \tag{C.5}$$

and

$$\dot{\lambda}_h = \left[-\frac{2g}{R+h}\sin\gamma + \frac{C_D V^2 S}{2m}\frac{\partial\rho}{\partial h}\right]\lambda_V + \left[\left(\frac{V}{(R+h)^2} - \frac{2g}{V(R+h)}\right)\cos\gamma - \frac{C_L VS}{2m}\frac{\partial\rho}{\partial h}\right]\lambda_\gamma . \tag{C.6}$$

The optimality condition is

$$\frac{\partial C_L}{\partial\alpha}\lambda_\gamma - V\frac{\partial C_D}{\partial\alpha}\lambda_V = 0. \tag{C.7}$$

The initial conditions ($t = t_0$) are

$$V(t_0) = V^0, \tag{C.8}$$

$$\gamma(t_0) = \gamma^0, \tag{C.9}$$

$$h(t_0) = h^0. \tag{C.10}$$

Terminal conditions ($t = t_f$) are

$$\gamma(t_f) = \gamma_f , \tag{C.11}$$

$$h(t_f) = h_f , \tag{C.12}$$

$$\lambda_V(t_f) = \lambda_{V_f} = 1, \tag{C.13}$$

$$\lambda_\gamma(t_f) = \lambda_{\gamma f} = -\nu_\gamma , \tag{C.14}$$

$$\lambda_h(t_f) = \lambda_{hf} = -\nu_h , \tag{C.15}$$

and

$$\left\{-\lambda_V\left[\frac{C_D\rho V^2 S}{2m} + g\sin\gamma\right] + \lambda_\gamma\left[\frac{C_L\rho VS}{2m} + \left(\frac{V}{R+h} - \frac{g}{V}\right)\cos\gamma\right] + \lambda_h(V\sin\gamma)\right\}_{t=t_f} = 0. \tag{C.16}$$

The unknowns $\lambda_\gamma(t_f)$, $\lambda_h(t_f)$, and $V(t_f)$ were estimated from an approximate optimum trajectory using the *method of steepest-ascent*

Consider perturbations of the optimum trajectory caused by perturbations in the initial and/or terminal conditions as found in equations (3.8) through (3.15). Perturbations on the equations of motion are

$$\delta\dot{V} = -\left[\frac{C_D\,\rho VS}{m}\right]\delta V - g\cos\gamma\,\delta\gamma + \left[\frac{2g\sin\gamma}{R+h} - \frac{C_D\,V^2S}{2m}\frac{\partial\rho}{\partial h}\right]\delta h - \frac{\rho V^2S}{2m}\frac{\partial C_D}{\partial\alpha}\delta\alpha, \tag{C.17}$$

$$\delta\dot{\gamma} = \left[\frac{C_L\,\rho S}{2m} + \left(\frac{1}{R+h} + \frac{g}{V^2}\right)\cos\gamma\right]\delta V - \left(\frac{V}{R+h} - \frac{g}{V}\right)\sin\gamma\,\delta\gamma + \left[\frac{C_L\,VS}{2m}\frac{\partial\rho}{\partial h} - \left(\frac{V}{(R+h)^2} - \frac{2g}{V(R+h)}\right)\cos\gamma\right]\delta h + \frac{\rho VS}{2m}\frac{\partial C_L}{\partial\alpha}\delta\alpha, \tag{C.18}$$

and

$$\delta\dot{h} = \sin\gamma\,\delta V + \cos\gamma\,\delta\gamma. \tag{C.19}$$

Perturbations on the Lagrange multiplier equations are

$$\begin{aligned}\delta\dot{\lambda}_V = {} & \left[\frac{C_D\,\rho S}{m}\lambda_V + \frac{2g}{V^3}\cos\gamma\,\lambda_\gamma\right]\delta V \\ & + \left[\left(\frac{1}{R+h} + \frac{g}{V^2}\right)\sin\gamma\,\lambda_\gamma - \cos\gamma\,\lambda_h\right]\delta\gamma \\ & + \left[\frac{C_D\,VS}{m}\frac{\partial\rho}{\partial h}\lambda_V - \left(\frac{C_L(\partial\rho/\partial h)S}{2m} - \left[\frac{1}{(R+h)^2} + \frac{2g}{V^2(R+h)}\right]\cos\gamma\right)\lambda_\gamma\right]\delta h \\ & + \left[\frac{\rho VS}{m}\frac{\partial C_D}{\partial\alpha}\lambda_V - \frac{\rho S}{2m}\frac{\partial C_L}{\partial\alpha}\lambda_\gamma\right]\delta\alpha + \frac{C_D\,\rho VS}{m}\delta\lambda_V \\ & - \left[\frac{C_L\,\rho S}{2m} + \left(\frac{1}{R+h} + \frac{g}{V^2}\right)\cos\gamma\right]\delta\lambda_\gamma - \sin\gamma\,\delta\lambda_h,\end{aligned} \tag{C.20}$$

$$\begin{aligned}\delta\dot{\lambda}_\gamma = {} & \left[\left(\frac{1}{R+h} + \frac{g}{V^2}\right)\sin\gamma\,\lambda_\gamma - \cos\gamma\,\lambda_h\right]\delta V \\ & + \left[-g\sin\gamma\,\lambda_V + \left(\frac{V}{R+h} - \frac{g}{V}\right)\cos\gamma\,\lambda_\gamma + V\sin\gamma\,\lambda_h\right]\delta\gamma \\ & - \left[\frac{2g}{R+h}\cos\gamma\,\lambda_V + \left(\frac{V}{(R+h)^2} - \frac{2g}{V(R+h)}\right)\sin\gamma\,\lambda_\gamma\right]\delta h \\ & + g\cos\gamma\,\delta\lambda_V + \left(\frac{V}{R+h} - \frac{g}{V}\right)\sin\gamma\,\delta\lambda_\gamma - V\cos\gamma\,\delta\lambda_h,\end{aligned} \tag{C.21}$$

$$
\begin{aligned}
\delta\dot{\lambda}_h = &\left\{\frac{C_D\, VS}{m}\frac{\partial\rho}{\partial h}\lambda_V + \left[\left(\frac{1}{(R+h)^2} + \frac{2g}{V^2(R+h)}\right)\cos\gamma\right.\right. \\
&\left.\left. - \frac{C_L\, S}{2m}\frac{\partial\rho}{\partial h}\right]\lambda_\gamma\right\}\delta V \\
&+ \left[-\frac{2g}{(R+h)}\cos\gamma\,\lambda_V + \left(\frac{2g}{V(R+h)} - \frac{V}{(R+h)^2}\right)\sin\gamma\,\lambda_\gamma\right]\delta\gamma \\
&+ \left\{\left[\frac{6g\sin\gamma}{(R+h)^2} + \frac{C_D\, V^2 S}{2m}\frac{\partial^2\rho}{\partial h^2}\right]\lambda_V\right. \\
&\left. + \left[\left(\frac{6g}{V(R+h)^2} - \frac{2V}{(R+h)^3}\right)\cos\gamma - \frac{C_L\, VS}{2m}\frac{\partial^2\rho}{\partial h^2}\right]\lambda_\gamma\right\}\delta h \\
&+ \left[\frac{V^2 S}{2m}\frac{\partial\rho}{\partial h}\frac{\partial C_D}{\partial\alpha}\lambda_V - \frac{VS}{2m}\frac{\partial\rho}{\partial h}\frac{\partial C_L}{\partial\alpha}\lambda_\gamma\right]\delta\alpha \\
&+ \left[\frac{C_D\, V^2 S}{2m}\frac{\partial\rho}{\partial h} - \frac{2g\sin\gamma}{R+h}\right]\delta\lambda_V \\
&+ \left[\left(\frac{V}{(R+h)^2} - \frac{2g}{V(R+h)}\right)\cos\gamma - \frac{C_L\, VS}{2m}\frac{\partial\rho}{\partial h}\right]\delta\lambda_\gamma\,.
\end{aligned}
\tag{C.22}
$$

The perturbations on the initial conditions are

$$\delta V(t_0) = \delta V^0, \tag{C.23}$$

$$\delta\gamma(t_0) = \delta\gamma^0, \tag{C.24}$$

$$\delta h(t_0) = \delta h^0. \tag{C.25}$$

The perturbations in the final conditions are

$$\delta\gamma(t_f) + \left[C_L\frac{\rho VS}{2m} + \left(\frac{V}{R+h} - \frac{g}{V}\right)\cos\gamma\right]_{t=t_f} dt_f = d\gamma^f, \tag{C.26}$$

$$\delta h(t_f) + [V\sin\gamma]_{t=t_f}\, dt_f = dh^f, \tag{C.27}$$

$$
\begin{aligned}
\delta\lambda_V(t_f) + \left\{\frac{C_D\rho VS}{m}\lambda_V + \left[C_L\frac{\rho S}{2m} + \left(\frac{1}{R+h} + \frac{g}{V^2}\right)\cos\gamma\right]\lambda_\gamma\right. \\
\left. - \sin\gamma\,\lambda_h\right\}_{t=t_f} dt_f = 0,
\end{aligned}
\tag{C.28}
$$

$$
\begin{aligned}
\delta\lambda_\gamma(t_f) + \left[g\cos\gamma\,\lambda_V + \left(\frac{V}{R+h} - \frac{g}{V}\right)\sin\gamma\,\lambda_\gamma\right. \\
\left. - V\cos\gamma\,\lambda_h\right]_{t=t_f} dt_f + d\nu_\gamma = 0,
\end{aligned}
\tag{C.29}
$$

$$
\begin{aligned}
&\delta\lambda_h(t_f) + \left\{\left[-\frac{2g\sin\gamma}{(R+h)} + \frac{C_D V^2 S}{2m}\frac{\partial\rho}{\partial h}\right]\lambda_V + \left[\left(\frac{V}{(R+h)^2}\right.\right.\right.\\
&\qquad \left.\left.\left. - \frac{2g}{V(R+h)}\right)\cos\gamma - \frac{C_L VS}{2m}\frac{\partial\rho}{\partial h}\right]\lambda_\gamma\right\}_{t=t_f} dt_f + d\nu_h = 0, \qquad \text{(C.30)}\\
&\left(\left\{-\frac{C_D \rho VS}{m}\lambda_V + \left[\frac{C_L \rho S}{2m} + \left(\frac{1}{R+h} + \frac{g}{V^2}\right)\cos\gamma\right]\lambda_\gamma\right.\right.\\
&\qquad \left.\left. + \sin\gamma\,\lambda_h\right\}\delta V\right.\\
&\qquad - \left[g\cos\gamma\,\lambda_V + \left(\frac{V}{R+h} - \frac{g}{V}\right)\sin\gamma\,\lambda_\gamma - V\cos\gamma\,\lambda_h\right]\delta\gamma\\
&\qquad + \left\{\left[\frac{C_L VS}{2m}\frac{\partial\rho}{\partial h} - \left(\frac{V}{(R+h)^2} - \frac{2g}{V(R+h)}\right)\cos\gamma\right]\lambda_\gamma\right.\\
&\qquad \left. + \left[\frac{2g\sin\gamma}{R+h} - \frac{C_D V^2 S}{2m}\frac{\partial\rho}{\partial h}\right]\lambda_V\right\}\delta h\\
&\qquad + \left[\frac{C_L \rho VS}{2m} + \left(\frac{V}{R+h} - \frac{g}{V}\right)\cos\gamma\right]\delta\lambda_\gamma\\
&\qquad \left. - \left[\frac{C_D \rho V^2 S}{2m} + g\sin\gamma\right]\delta\lambda_V + V\sin\gamma\,\delta\lambda_h\right)_{t=t_f} = 0. \qquad \text{(C.31)}
\end{aligned}
$$

The variation in the angle of attack is found by perturbing equation (C.7):

$$\delta\alpha = -\left(\frac{1}{\partial^2 H/\partial\alpha^2}\right)(D_1\,\delta V + D_2\,\delta h + D_3\,\delta\lambda_V + D_4\,\delta\lambda_\gamma), \qquad \text{(C.32)}$$

where

$$
\begin{aligned}
D_1 &= \lambda_\gamma \frac{\rho S}{2m}\frac{\partial C_L}{\partial\alpha} - \lambda_V \frac{\rho VS}{m}\frac{\partial C_D}{\partial\alpha},\\
D_2 &= \left[\lambda_\gamma \frac{VS}{2m}\frac{\partial C_L}{\partial\alpha} - \frac{V^2 S}{2m}\lambda_V \frac{\partial C_D}{\partial\alpha}\right]\frac{\partial\rho}{\partial h},\\
D_3 &= -\frac{\rho V^2 S}{2m}\frac{\partial C_D}{\partial\alpha},\\
D_4 &= \frac{\rho VS}{2m}\frac{\partial C_L}{\partial\alpha},
\end{aligned}
$$

and

$$\frac{\partial^2 H}{\partial\alpha^2} = \lambda_\gamma \frac{\rho VS}{2m}\frac{\partial^2 C_L}{\partial\alpha^2} - \lambda_V \frac{\rho V^2 S}{2m}\frac{\partial^2 C_D}{\partial\alpha^2} < 0 \qquad \text{(C.33)}$$

for maximum velocity.

J.S.I.A.M. CONTROL
Ser. A, Vol. 1, No. 2
Printed in U.S.A., 1963

ON THE CORRECTNESS OF THE FORMULATION OF AN OPTIMAL CONTROL PROBLEM*

F. M. KIRILLOVA†

In this note we shall prove that the solution of an optimal control problem [1–4] depends continuously on the initial data and on the parameters of the system (under certain assumptions).

1. Let the control system be described by the differential equations:

$$\frac{dx_i}{dt} = \sum_{k=1}^{n} a_{ik}(t, c_1, \cdots, c_e)x_k(t) + \sum_{k=1}^{r} b_{ik}(t, c_1, \cdots, c_e)u_k(t), \quad i = 1, \cdots, n, \tag{1.1}$$

where $x_1, \cdots, x_n$ are the coordinates of the representative point in phase space, $c_1, \cdots, c_e$ are parameters, and $u_1(t), \cdots, u_r(t)$ are control functions. The functions $a_{ik}(t, c_1, \cdots, c_e)$, and $b_{ik}(t, c_1, \cdots, c_e)$ are continuous in the time t and in the system parameters $c_1, \cdots, c_e$.

We shall assume that the $u_k(t)$ are piecewise continuous, and that the constraint $\max |u_k(t)| \leqq N$, $k = 1, \cdots, r$, has been imposed on these functions, where N is a certain constant. In evaluating $\max |u_k(t)|$, we shall not take into account values $|u_k(t)|$ at isolated points t, if these points constitute a set of measure zero.[1]

For given fixed parameters $c_1, \cdots, c_e$ in (1.1), the problem consists in the following: Given a time t_0 and a point $x_i(t_0) = x_{i0}$, choose control functions $u_k(t)$ such that the point $x_i(x_{10}, \cdots, x_{n0}, t_0, u_1, \cdots, u_r, t)$, moving along a trajectory of the system, attains the origin from the initial position $x_{10}, \cdots, x_{n0}$ in the shortest time T.

The functions $u_k(t)$ which satisfy the requirements of this problem will be called optimal control functions, and the time T will be called the optimal time for the response of the process.

It was shown in [1–4] that a solution of the given problem exists (under certain restrictions). It is of interest to investigate how the optimal control

* Originally published in Izvestia VUZ, Matematika, No. 4 (5), 1958, pp. 113–126. Submitted Jan. 24, 1958. Translated by L. W. Neustadt, Aerospace Corporation, El Segundo, California. Some minor changes—for the sake of clarity and completness—have been made by the translator, particularly in Lemmas 3.1 and 4.1.

† The S. M. Kirov Ural Polytechnic Institute

[1] In such circumstances it is customary to say: $|u_k(t)| \leqq N$ almost everywhere in $[t_0, \tau]$. In practice, one may consider that $|u_k(t)| > N$ only at isolated points, the number of which, on each interval $[t_0, \tau]$, is finite. These points should simply be neglected in the given problem.

functions and the optimal control time depend on the initial data $x_{10}, \cdots, x_{n0}$ and on the system parameters $c_1, \cdots, c_e$.

In this note we shall prove that the solution $u_k(x_{10}, \cdots, x_{n0}, t_0, c_1, \cdots, c_e, t)$, $T(x_{10}, \cdots, x_{n0}, t_0, c_1, \cdots, c_e)$ depends continuously on the initial data $x_{10}, \cdots, x_{n0}$ and on the parameters $c_1, \cdots, c_e$ (under certain assumptions). In other words, we shall settle below the question of the correctness of the formulation of the optimal control problem.

We introduce some notations. If z is a certain function $z(x_{10}, \cdots, x_{n0}, t_0, c_1, \cdots, c_e)$ of the initial data and of the system parameters, we shall in the sequel write $z = z(x_0, t_0, c)$. In particular, in place of $u_k(x_{10}, \cdots, x_{n0}, t_0, c_1, \cdots, c_e, t)$ and $T(x_{10}, \cdots, x_{n0}, t_0, c_1, \cdots, c_e)$, we shall use the symbols $u_k(x_0, t_0, c, t)$ and $T(x_0, t_0, c)$.

The optimal control problem may be considered to be correctly formulated if the optimal control functions $u_k(x_0, t_0, c, t)$ and the time $T(x_0, t_0, c)$ (for fixed values of $c_1, \cdots, c_e$) are continuous in the initial data $x_{10}, \cdots, x_{n0}$; that is, for every $\sigma > 0$ and $\epsilon > 0$ there exists a $\delta > 0$ such that[2]

$$\text{meas}\,(\,|\,u_k(x_0, t_0, c, t) - u_k(x_0', t_0, c, t)\,| \geqq \sigma) < \epsilon,$$
$$|\,T(x_0, t_0, c) - T(x_0', t_0, c)\,| < \epsilon$$

whenever $|\,x_{i0} - x_{i0}'\,| < \delta$.

Also, for given initial conditions $x_{10}, \cdots, x_{n0}$, we shall say that the solution $u_k(x_0, t_0, c, t)$, $T(x_0, t_0, c)$ is continuous in the parameters $c_1, \cdots, c_e$, if, for every $\sigma > 0$ and $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\text{meas}\,(\,|\,u_k(x_0, t_0, c, t) - u_k(x_0, t_0, c', t)\,| \geqq \sigma) < \epsilon,$$
$$|\,T(x_0, t_0, c) - T(x_0, t_0, c')\,| < \epsilon$$

whenever $|\,c_i - c_i'\,| < \delta$.

The proof that the solution of the optimal problem has these properties is given in §§3, 4. In §2 some auxiliary material is presented.

2. Let $u_k(t) = \eta_k(t)$, $k = 1, \cdots, r$, in equations (1.1), where the $\eta_k(t)$ are piecewise continuous, essentially bounded functions.

The solution of (1.1) has the form (see [6])

$$(2.1)\quad x_i(t) = \sum_{k=1}^{n} \varphi_{ik}(c, t)x_{k0} + \int_{t_0}^{t} \sum_{s,l=1}^{n} \sum_{k=1}^{r} \varphi_{is}(c, t)\psi_{sl}(c, \tau)b_{lk}(c, \tau)\eta_k(\tau)\, d\tau,$$

[2] The symbol meas$(|\,u_k(x_0, t_0, c, t) - u_k(x_0', t_0, c, t)\,| \geqq \sigma)$ denotes the measure of the set of t for which $|\,u_k(x_0, t_0, c, t) - u_k(x_0', t_0, c, t)\,| \geqq \sigma$. In other words, we are concerned with convergence in measure of the control functions. This arises from the specific character of the problem, whose solution $u_k(x_0, t_0, c, t)$ is made up of discontinuous functions.

where the $\varphi_{ik}(c, t)$ are the functions in the fundamental matrix of solutions of system (1.1) with $b_{ik} = 0$, and the $\psi_{sl}(c, \tau)$ are the functions in the matrix which is the inverse of the fundamental matrix.

Multiplying relation (2.1) by $\psi_{mi}(c, t)$, and summing over i, we obtain

$$\sum_{i=1}^{n} \psi_{mi}(c, t)x_i(t) = x_{m0} + \int_{t_0}^{t} \sum_{k=1}^{r} \sum_{l=1}^{n} \psi_{ml}(c, \tau)b_{lk}(c, \tau)\eta_k(\tau)\, d\tau, \; m = 1, \cdots, n.$$

Suppose that the point $x_i(x_0, t_0, c, \eta_1, \cdots, \eta_r, t)$ of the trajectory of (1.1) attains the origin for some $t > t_0$. Then the functions $\eta_k(\tau)$ must be a solution of the system

$$-x_{m0} = \int_{t_0}^{t} \sum_{k=1}^{r} \gamma_{mk}(c, \tau)\eta_k(\tau)\, d\tau, \qquad m = 1, \cdots, n, \tag{2.2}$$

where

$$\gamma_{mk}(c, \tau) = \sum_{l=1}^{n} \psi_{ml}(c, \tau)b_{lk}(c, \tau).$$

Let us consider the space $L(t_0, t)$ of systems of functions $\gamma(\tau) = \{\gamma_1(\tau), \cdots, \gamma_r(\tau)\}$, i.e. the space of Lebesgue integrable functions

$$\gamma_k(\tau)\left(\int_{t_0}^{t} |\, \gamma_k(\tau)\,|\, d\tau < +\infty\right).$$

Let $\|\, \gamma(\tau)\,\|$ denote the norm of the element $\gamma(\tau)$. Namely,

$$\|\, \gamma(t)\| = \int_{t_0}^{t} \sum_{k=1}^{r} |\, \gamma_k(\tau)|\, d\tau.$$

Then, the general form of a linear functional which is defined on the given normed linear space is

$$f(\gamma) = \int_{t_0}^{t} \sum_{k=1}^{r} \gamma_k(\tau)\eta_k(\tau)\, d\tau,$$

where the $\eta_k(t)$ are piecewise continuous, essentially bounded functions. Moreover, the norm of the functional f is given by the formula

$$\|f\| = \max |\, \eta_k(\tau)\,|, \quad t_0 \leqq \tau \leqq t, k = 1, \cdots, r.$$

Returning to relation (2.2), the optimal control problem can be formulated as follows: Among the linear functionals which are defined on the space $L(t_0, t)$ of systems of functions

$$\gamma(c, \tau) = \{\gamma_1(c, \tau), \cdots, \gamma_r(c, \tau)\}$$

with norm

$$\|\, \gamma(c, t)\| = \int_{t_0}^{t} \sum_{k=1}^{r} |\, \gamma_k(c, \tau)|\, d\tau,$$

where $c_1, \cdots, c_e$ are given fixed parameters, find the functional f of least norm which is a solution of (2.2). If there exist values of t for which $\min \| f \| \leqq N$, then, for the given initial conditions $x_{10}, \cdots, x_{n0}$, and the constraints $| u_k(\tau) | \leqq N$, the origin is attainable[3] in the time $\theta = t - t_0$. If there is a smallest θ for which $\min \| f \| \leqq N$, then an optimal control does exist.

We shall consider the given problem for the case where the systems of functions $\{\gamma_{i1}(c, \tau), \cdots, \gamma_{ir}(c, \tau)\}$ are completely independent [7]; i.e., we shall suppose that, for each $k = 1, \cdots, r$, the relations

$$\sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) = 0, \qquad \sum_{i=1}^{n} \lambda_i^2 \neq 0$$

hold only on a set of measure zero.

Following [7], we shall say that the element

$$\gamma(c, \tau) = \left(\sum_{i=1}^{n} \lambda_i \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \lambda_i \gamma_{ir}(c, \tau)\right)$$

is minimizing on the interval $[t_0, t]$ if

$$\min \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) \right| d\tau = \| \gamma(c, t) \|,$$

the minimum being taken over all $\lambda_1, \cdots, \lambda_n$ satisfying the relation

$$\sum_{i=1}^{n} \lambda_i x_{i0} = -1, \qquad \sum_{i=1}^{n} x_{i0}^2 \neq 0.$$

As is shown in [7], a solution of (2.2) exists (provided the systems $\{\gamma_{i1}(c, \tau), \cdots, \gamma_{ir}(c, \tau)\}$ are completely independent) if and only if we allow $\| f \| \geqq \lambda(x_0, t_0, c, t)$, where the function $\lambda(x_0, t_0, c, t)$ is defined by the condition

$$\frac{1}{\lambda(x_0, t_0, c, t)} = \min \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) \right| d\tau,$$

$$\sum_{i=1}^{n} \lambda_i x_{i0} = -1, \qquad \sum_{i=1}^{n} x_{i0}^2 \neq 0.$$

Since

$$\min \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) \right| d\tau$$

is always attained [7] with values $\lambda_1, \cdots, \lambda_n$ satisfying the conditions

$$\sum_{i=1}^{n} \lambda_i x_{i0} = -1, \qquad \sum_{i=1}^{n} x_{i0}^2 \neq 0,$$

[3] This will, for example, always be true if the a_{ik} are constant, and if the eigenvalues of the matrix (a_{ik}) have negative real parts.

we may consider a functional f which is a solution of (2.2) and has norm $\lambda(x_0, t_0, c, t)$. Let $\|f\| = \lambda(x_0, t_0, c, t)$, and let f be a solution of (2.2). If the element

$$\gamma(c, \tau) = \left(\sum_{i=1}^{n} \lambda_i \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \lambda_i \gamma_{ir}(c, \tau)\right)$$

is minimizing on $[t_0, t]$, then it is obvious that

$$\int_{t_0}^{t} \sum_{k=1}^{r} \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau)\eta_k(\tau)\, d\tau = 1$$

and

$$\int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) \right| d\tau = \frac{1}{\lambda(x_0, t_0, c, t)}.$$

Hence, $f(\gamma) = \|\gamma\| \, \|f\|$. Consequently, if f is a solution of (2.2), and if $\|f\| = \lambda(x_0, t_0, c, t)$, then, for every element

$$\gamma(c, \tau) = \left(\sum_{i=1}^{n} \lambda_i \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \lambda_i \gamma_{ir}(c, \tau)\right),$$

which is minimizing on $[t_0, t]$, we have

$$\int_{t_0}^{t} \sum_{k=1}^{r} \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau)\eta_k(\tau)\, d\tau = \lambda(x_0, t_0, c, t) \cdot \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) \right| d\tau.$$

Hence, if we neglect a set of measure zero, it follows that the functions $\eta_k(\tau)$ can be found from the formulas [3]:

$$\begin{aligned} &\eta_k(x_0, t_0, c, \tau) \\ &= \lambda(x_0, t_0, c, t) \operatorname{sign} \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau), \quad \sum_{i=1}^{n} \lambda_i x_{i0} = -1, \, k = 1, \cdots, r. \end{aligned} \tag{2.3}$$

Thus, if the origin can be attained from $(x_{10}, \cdots, x_{n0})$, the control functions are determined uniquely by relations (2.3).

It can be shown that the function $\lambda(x_0, t_0, c, t)$ is continuous and strictly monotonic in t (for fixed $x_{10}, \cdots, x_{n0}, t_0, c_1, \cdots, c_e$). In fact, if the element

$$\gamma_1(c, \tau) = \left(\sum_{i=1}^{n} \lambda_i^1 \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \lambda_i^1 \gamma_{ir}(c, \tau)\right)$$

is minimizing on $[t_0, t_1]$, and if the element

$$\gamma_2(c, \tau) = \left(\sum_{i=1}^{n} \lambda_i^2 \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \lambda_i^2 \gamma_{ir}(c, \tau)\right)$$

is minimizing on $[t_0, t_2]$, then it is obvious that

$$\int_{t_0}^{t_1} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{\,2} \gamma_{ik}(c, \tau) \right| d\tau \geqq \int_{t_0}^{t_1} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{\,1} \gamma_{ik}(c, \tau) \right| d\tau.$$

Let $t_1 < t_2$. Then,

$$\int_{t_0}^{t_2} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{\,2} \gamma_{ik}(c, \tau) \right| d\tau > \int_{t_0}^{t_1} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{\,2} \gamma_{ik}(c, \tau) \right| d\tau.$$

Hence,

$$\int_{t_0}^{t_2} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{\,2} \gamma_{ik}(c, \tau) \right| d\tau > \int_{t_0}^{t_1} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{\,1} \gamma_{ik}(c, \tau) \right| d\tau,$$

i.e., $F(x_0, t_0, c, t_2) > F(x_0, t_0, c, t_1)$, where $F(x_0, t_0, c, t) = [\lambda(x_0, t_0, c, t)]^{-1}$, which also proves the monotonicity of $F(x_0, t_0, c, t)$ (and, consequently, also of $\lambda(x_0, t_0, c, t)$).

Let us now prove that $F(x_0, t_0, c, t)$ is continuous. Since

$$\int_{t_0}^{t_2} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{2} \gamma_{ik}(c, \tau) \right| d\tau = \min \int_{t_0}^{t_2} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) \right| d\tau,$$

$$\sum_{i=1}^{n} \lambda_i x_{i0} = -1,$$

it follows that

$$\int_{t_0}^{t_2} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{1} \gamma_{ik}(c, \tau) \right| d\tau \geqq \int_{t_0}^{t_2} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{2} \gamma_{ik}(c, \tau) \right| d\tau,$$

from which we see that

$$F(x_0, t_0, c, t_2) - F(x_0, t_0, c, t_1) \leqq \int_{t_1}^{t_2} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{1} \gamma_{ik}(c, \tau) \right| d\tau.$$

If $t_2 \to t_1$, $F(x_0, t_0, c, t_2) \to F(x_0, t_0, c, t_1)$. This proves that $F(x_0, t_0, c, t)$ is continuous from the right. To show that $F(x_0, t_0, c, t)$ is continuous from the left, we use the inequality

$$\int_{t_0}^{t_1} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{2} \gamma_{ik}(c, \tau) \right| d\tau \geqq \int_{t_0}^{t_1} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{1} \gamma_{ik}(c, \tau) \right| d\tau.$$

If $t_2 < t_1$,

$$F(x_0, t_0, c, t_1) - F(x_0, t_0, c, t_2) \leqq \int_{t_2}^{t_1} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{2} \gamma_{ik}(c, \tau) \right| d\tau.$$

We shall show below in Lemma 3.1 that the numbers $\lambda_i^{\,2}$ (for the values $t_2 \to t_1$) are uniformly bounded. Therefore, the last inequality implies that $F(x_0, t_0, c, t_2) \to F(x_0, t_0, c, t_1)$ as $t_2 \to t_1$.

From the fact that $F(x_0, t_0, c, t)$ is continuous in t it follows that if the origin can be attained[4] for at least one value of t—with the initial data $x_{10}, \cdots, x_{n0}$ and bounds $|u_k(\tau)| \leqq N$—then there exists a unique optimal control defined by the formulas obtained from (2.3):

$$u_k(x_0, t_0, c, \tau) = N \operatorname{sign} \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau), \qquad \sum_{i=1}^{n} \lambda_i x_{i0} = -1, \qquad k = 1, \cdots, r. \tag{2.4}$$

3. The proof of the basic assertions which were stated in §1 is based on an auxiliary lemma. By the symbol $\gamma_{im}(\tau)_t$, $m = 1, \cdots, r$, we shall denote continuous functions of the argument τ which are defined on the interval $t_0 \leqq \tau \leqq t \leqq \theta$.

LEMMA 3.1. *Let there be given a sequence*

$$\gamma_k(\tau)_{t_k} = \left(\sum_{i=1}^{n} \lambda_i^k \gamma_{i1}^k(\tau)_{t_k}, \cdots, \sum_{i=1}^{n} \lambda_i^k \gamma_{ir}^k(\tau)_{t_k} \right), \qquad k = 1, 2, \cdots,$$

such that $\| \gamma_k(\tau)_{t_k} \| \leqq K$ *for every* k, *where* K *is a constant. If the* $\gamma_{im}^k(\tau)_{t_k}$, $m = 1, \cdots, r$, $k = 1, 2, \cdots$, *are continuous functions such that* $\gamma_{im}^k(\tau) \to \gamma_{im}^0(\tau)$ *uniformly as* $k \to \infty$ *for every* i *and* m, *and if the systems* $\{\gamma_{i1}^0(\tau), \cdots, \gamma_{ir}^0(\tau)\}$ *are completely independent, then the numbers* λ_i^k, $k = 1, 2, \cdots$, $i = 1, \cdots, n$, *are uniformly bounded.*

Proof. Assume the contrary. Without loss of generality, we shall assume that

$$c_k = |\lambda_1^k| + \cdots + |\lambda_n^k| \to \infty \quad \text{as} \quad k \to \infty.$$

Since $t_0 \leqq t_k \leqq \theta$, there is no loss of generality in assuming that t_k tends monotonically to some limit t as $k \to \infty$.

By hypothesis, $\| \gamma_k(\tau)_{t_k} \| \leqq K$. Consequently, the sequence of the norms of the elements

$$\bar{\gamma}_k(\tau)_{t_k} = \left(\sum_{i=1}^{n} \frac{\lambda_i^k}{c_k} \gamma_{i1}^k(\tau)_{t_k}, \cdots, \sum_{i=1}^{n} \frac{\lambda_i^k}{c_k} \gamma_{ir}^k(\tau)_{t_k} \right)$$

tends to zero. Since

$$\sum_{i=1}^{n} \left| \frac{\lambda_i^k}{c_k} \right| = 1,$$

we shall assume, again without loss of generality, that

$$\frac{\lambda_i^k}{c_k} \to \alpha_i$$

for each i, as $k \to \infty$.

[4] See footnote 3.

Consider the element

$$\gamma_0(\tau)_t = \left(\sum_{i=1}^{n} \alpha_i \gamma_{i1}^0(\tau)_t, \cdots, \sum_{i=1}^{n} \alpha_i \gamma_{ir}^0(\tau)_t\right).$$

Suppose that the sequence t_k is increasing. Then,

$$\left| \int_{t_0}^{t_k} \sum_{m=1}^{r} \left| \sum_{i=1}^{n} \frac{\lambda_i^k}{c_k} \gamma_{im}^k(\tau)_{t_k} \right| d\tau - \int_{t_0}^{t} \sum_{m=1}^{r} \left| \sum_{i=1}^{n} \alpha_i \gamma_{im}^0(\tau)_t \right| d\tau \right|$$

$$\leqq \int_{t_0}^{t_k} \sum_{m=1}^{r} \left| \sum_{i=1}^{n} \left\{ \left| \frac{\lambda_i^k}{c_k} \right| | \gamma_{im}^k(\tau)_{t_k} - \gamma_{im}^0(\tau)_{t_k} | + | \gamma_{im}^0(\tau)_{t_k} | \left| \frac{\lambda_i^k}{c_k} - \alpha_i \right| \right\} d\tau$$

$$+ \int_{t_k}^{t} \sum_{m=1}^{r} \left| \sum_{i=1}^{n} \alpha_i \gamma_{im}^0(\tau)_t \right| d\tau$$

(an analogous estimate for the difference of the norms of the elements $\gamma_0(\tau)_t$ and $\bar{\gamma}_k(\tau)_{t_k}$ can be obtained if the sequence t_k is decreasing).

Since the functions γ_{im}^k are continuous in τ and

$$\lim \gamma_{im}^k(\tau) = \gamma_{im}^0(\tau) \quad \text{as} \quad k \to \infty, \quad \text{uniformly in } \tau,$$

it follows that

$$\| \bar{\gamma}_k(\tau)_{t_k} \| \to \| \gamma_0(\tau)_t \| \quad \text{as} \quad k \to \infty.$$

But $\| \bar{\gamma}_k(\tau)_{t_k} \| \to 0$, so that $\| \gamma_0(\tau)_t \| = 0$. However, $\sum_{i=1}^{n} | \alpha_i | = 1$, and the systems of functions $\{\gamma_{i1}^0(\tau)_t, \cdots, \gamma_{ir}^0(\tau)_t\}$ are completely independent, so that, for each k,

$$\sum_{i=1}^{n} \alpha_i \gamma_{ik}^0(\tau) = 0$$

only on a set of measure zero. The contradiction proves the lemma.

We shall now show that, for fixed parameters $c_1, \cdots, c_e$, the solution $u_k(x_0, t_0, c, \tau)$, $T(x_0, t_0, c)$ of the optimal problem is continuous in the initial data $x_{10}, \cdots, x_{n0}$.

THEOREM 1. *If the systems of functions* $\{\gamma_{i1}(c, \tau), \cdots, \gamma_{ir}(c, \tau)\}$ *of* (2.2) *are completely independent, and if there exists, for certain initial conditions* $x_{10}, \cdots, x_{n0}$, *an optimal control* $u_k(x_0, t_0, c, \tau)$ *with optimal time* $T(x_0, t_0, c)$, *then, for every* $\sigma > 0$ *and* $\epsilon > 0$, *there exists a* $\delta > 0$ *such that*

$$\text{meas } (| u_k(x_0, t_0, c, \tau) - u_k(x_0', t_0, c, \tau) | \geqq \sigma) < \epsilon,$$

$$| T(x_0, t_0, c) - T(x_0', t_0, c) | < \epsilon$$

whenever $| x_{i0} - x_{i0}' | < \delta$.

Proof. Let T be the optimal control time for the initial conditions $x_{10}, \cdots, x_{n0}$. By the definition of $F(x_0, t_0, c, t)$, $F(x_0, t_0, c, T) = 1/N$. As was shown in §2, $F(x_0, t_0, c, t)$ is strictly monotonic and continuous in

t. Consequently, for an arbitrary $\epsilon > 0$, there exists a $\beta > 0$ such that

$$F(x_0, t_0, c, T - \epsilon) < \frac{1}{N} - \beta, \qquad F(x_0, t_0, c, T + \epsilon) > \frac{1}{N} + \beta.$$

In [7] it is shown that, for fixed $c_1, \cdots, c_e$, and t, the function $F(x_0, t_0, c, t)$ is continuous in $x_{10}, \cdots, x_{n0}$. Therefore, for a sequence of initial data $x_{10}^m, \cdots, x_{n0}^m$ converging to $x_{10}, \cdots, x_{n0}$, we have

$$| F(x_0, t_0, c, T - \epsilon) - F(x_0^m, t_0, c, T - \epsilon) | < \beta,$$

$$| F(x_0, t_0, c, T + \epsilon) - F(x_0^m, t_0, c, T + \epsilon) | < \beta,$$

when $m > M$. Taking the preceding inequalities into account, we conclude that

$$F(x_0^m, t_0, c, T - \epsilon) < \frac{1}{N}, \qquad F(x_0^m, t_0, c, T + \epsilon) > \frac{1}{N}$$

when $m > M$. But, for fixed m, the function $F(x_0^m, t_0, c, t)$ is strictly monotonic and continuous in t. Consequently, for every $m > M$, there is a unique $\theta_m = T(x_0^m, t_0, c)$ such that

$$F(x_0^m, t_0, c, \theta_m) = \frac{1}{N},$$

and, hence, $| T(x_0, t_0, c) - T(x_0^m, t_0, c) | < \epsilon$ when $m > M$. Because of the arbitrariness in the choice of $\epsilon > 0$, this also proves that the optimal time $T(x_0, t_0, c)$ depends continuously on $x_{10}, \cdots, x_{n0}$.

We note that because of the continuity of the optimal control time in the initial data, it follows that the set of points in phase space from which the origin is attainable (with given t_0) is open. For systems of equations with constant coefficients, this fact was noted in [2].

Let us now prove that the optimal control functions $u_k(x_0^m, t_0, c, \tau)$ converge in measure to the functions $u_k(x_0, t_0, c, \tau)$ as $x_{i0}^m \to x_{i0}$. As was shown above,

$$u_k(x_0, t_0, c, \tau) = N \operatorname{sign} \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau), \quad k = 1, \cdots, r,$$

$$\sum_{i=1}^{n} \lambda_i^0 x_{i0} = -1, \qquad t_0 \leqq \tau \leqq t_0 + T,$$

$$u_k(x_0^m, t_0, c, \tau) = N \operatorname{sign} \sum_{i=1}^{n} \lambda_i^m \gamma_{ik}(c, \tau), \quad k = 1, \cdots, r,$$

$$\sum_{i=1}^{n} \lambda_i^m x_{i0}^m = -1, \qquad t_0 \leqq \tau \leqq t_0 + T_m$$

(the symbols T and T_m denote $T(x_0, t_0, c)$ and $T(x_0^m, t_0, c)$, respectively).

The element

$$\gamma_0(c, \tau) = \left(\sum_{i=1}^{n} \lambda_i^0 \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \lambda_i^0 \gamma_{ir}(c, \tau)\right)$$

is minimizing for the condition

$$\sum_{i=1}^{n} \lambda_i^0 x_{i0} = -1$$

on the interval $[t_0, t_0 + T]$. The elements

$$\gamma_m(c, \tau) = \left(\sum_{i=1}^{n} \lambda_i^m \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \lambda_i^m \gamma_{ir}(c, \tau)\right)$$

are respectively minimizing for the conditions

$$\sum_{i=1}^{n} \lambda_i^m x_{i0}^m = -1$$

on the intervals $[t_0, t_0 + T_m]$.

Since

$$\| \gamma_0(c, \tau) \| = \frac{1}{N}, \qquad \| \gamma_m(c, \tau) \| = \frac{1}{N} \quad \text{when} \quad m > M,$$

we conclude on the basis of Lemma 3.1 that the numbers λ_i^m are uniformly bounded. Let $\lambda_i^{m_s}$ be convergent subsequences of the sequences λ_i^m:

$$\lambda_i^{m_s} \to \bar{\lambda}_i \quad \text{as} \quad s \to \infty, \qquad i = 1, \cdots, n.$$

Then, the element

$$\bar{\gamma}(c, \tau) = \left(\sum_{i=1}^{n} \bar{\lambda}_i \gamma_{i1}(c, \tau), \cdots, \sum_{i=1}^{n} \bar{\lambda}_i \gamma_{ir}(c, \tau)\right)$$

is obviously minimizing for the condition

$$\sum_{i=1}^{n} \lambda_i x_{i0} = -1$$

on the interval $[t_0, t_0 + T]$. Consequently,

$$\operatorname{sign} \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau) = \operatorname{sign} \sum_{i=1}^{n} \bar{\lambda}_i \gamma_{ik}(c, \tau), \quad k = 1, \cdots, r,$$

almost everywhere on $[t_0, t_0 + T]$ (see [7]).

But this means that, beginning with some number $m > M$, all the zeros of the functions

$$\sum_{i=1}^{n} \lambda_i^m \gamma_{ik}(c, \tau), \qquad k = 1, \cdots, r,$$

are found in an arbitrarily small ϵ-neighborhood of the zeros of the functions

$$\sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau),$$

respectively (when speaking of zeros here, we refer only to those zeros where the functions actually change sign).

Indeed, if, for some $\epsilon > 0$, there were no M satisfying the preceding condition, it would follow from the uniform boundedness of the numbers λ_i^m that there exist convergent subsequences $\lambda_i^{m_s}$; and if

$$\lambda_i^{m_s} \to \lambda_i^1 \quad \text{as} \quad s \to \infty, \qquad i = 1, \cdots, n,$$

then

$$\operatorname{sign} \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau) = \operatorname{sign} \sum_{i=1}^{n} \lambda_i^1 \gamma_{ik}(c, \tau), \quad k = 1, \cdots, r,$$

almost everywhere on $[t_0, t_0 + T]$. Thus, our assumption leads to a contradiction. Consequently, for every $\epsilon > 0$, there exists a number M beginning with which all of the zeros of the functions

$$\sum_{i=1}^{n} \lambda_i^m \gamma_{ik}(c, \tau)$$

fall within an ϵ-neighborhood of the zeros of the functions

$$\sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau),$$

and outside of this ϵ-neighborhood the signs of the functions

$$\sum_{i=1}^{n} \lambda_i^m \gamma_{ik}(c, \tau) \quad \text{and} \quad \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau) \quad \text{for} \quad m > M$$

agree. But this means that, for any $\sigma > 0$,

$$\operatorname{meas} (| u_k(x_0, t_0, c, \tau) - u_k(x_0^m, t_0, c, \tau) | \geqq \sigma) < \epsilon, m > M.$$

Thus, the optimal control time $T(x_0, t_0, c)$ and the control functions $u_k(x_0, t_0, c, \tau)$, for fixed parameters $c_1, \cdots, c_e$, are continuous in the initial data $x_{10}, \cdots, x_{n0}$.

Note. Inasmuch as an actual automatic control system can never be identified exactly, it follows, in particular, that there will always be an error in the determination of the initial data $x_{10}, \cdots, x_{n0}$. Therefore, the following problem is of interest: Let the optimal control $u_k(x_0, t_0, \tau)$ achieve the origin in the time $T(x_0, t_0) = T$. Will the point on a trajectory of (1.1) be near the point $x_i = 0$ at the time $t_0 + T$ (t_0 being the initial time) if the initial data $x_{10}, \cdots, x_{n0}, t_0$ are the very same, but the control

functions correspond to initial conditions $x_{10}', \cdots, x_{n0}'$ near $x_{10}, \cdots, x_{n0}$?

The theorem proved above permits us to answer this question in the affirmative (under our assumptions).

Indeed, if $u_k(x_0, t_0, \tau)$ are the optimal control functions for the initial conditions $x_{10}, \cdots, x_{n0}$, and if T is the control time, then, according to formulas in [6], we have

$$0 = \sum_{i=1}^{n} \varphi_{ik}(t_0 + T)x_{k0} + \int_{t_0}^{t_0+T} \sum_{k=1}^{r} \sum_{s,l=1}^{n} \varphi_{is}(t_0 + T)\psi_{sl}(\tau)b_{lk}(\tau)u_k(x_0, t_0, \tau)\, d\tau. \tag{3.2}$$

The position of the point $x_i(x_0, t_0, u_1, \cdots, u_r, t)$—moving along a trajectory of (1.1) with the control $u_k(x_0', t_0, \tau)$—at the time $t = t_0 + T$ is determined by the relations (according to formulas in [6])

$$x_i(t_0 + T) = \sum_{i=1}^{n} \varphi_{ik}(t_0 + T)x_{k0} + \int_{t_0}^{t_0+T} \sum_{k=1}^{r} \sum_{s,l=1}^{n} \varphi_{is}(t_0 + T)\psi_{sl}(\tau)b_{lk}(\tau)u_k(x_0', t_0, \tau)\, d\tau. \tag{3.3}$$

From (3.2) and (3.3) we obtain

$$x_i(t_0 + T) = \int_{t_0}^{t_0+T} \sum_{k=1}^{r} \sum_{s,l=1}^{n} \varphi_{is}(t_0 + T)\psi_{sl}(\tau)b_{lk}(\tau)(u_k(x_0', t_0, \tau) - u_k(x_0, t_0, \tau))\, d\tau. \tag{3.4}$$

Let us estimate $x_i(t_0 + T)$. Since the functions $\varphi_{is}(\tau)$, $\psi_{sl}(\tau)$, and $b_{lk}(\tau)$ are continuous in τ, and, consequently, bounded on $[t_0, t_0 + T]$, it follows from (3.4) that

$$| x_i(t_0 + T) | \leqq A \sum_{k=1}^{r} \int_{t_0}^{t_0+T} | u_k(x_0, t_0, \tau) - u_k(x_0', t_0, \tau) |\, d\tau,$$

where A is some constant.

Let us partition the interval $[t_0, t_0 + T]$ into subsets:

$$E_k^1 = \{\tau : | u_k(x_0, t_0, \tau) - u_k(x_0', t_0, \tau) | \geqq \sigma\},$$
$$E_k^2 = \{\tau : | u_k(x_0, t_0, \tau) - u_k(x_0', t_0, \tau) | < \sigma\}$$

where σ is a positive constant defined below. Then,

$$| x_i(t_0 + T) | \leqq A \sum_{k=1}^{r} \left[\int_{E_k^1} | u_k(x_0, t_0, \tau) - u_k(x_0', t_0, \tau) |\, d\tau + \int_{E_k^2} | u_k(x_0, t_0, \tau) - u_k(x_0', t_0, \tau) |\, d\tau \right].$$

Now let β be an arbitrary, as small as desired, positive number. By Theorem 1, if $\epsilon = \beta/(4ArN)$, there is a $\delta > 0$ such that

$$\text{meas } (| u_k(x_0, t_0, \tau) - u_k(x_0', t_0, \tau) | \geqq \sigma) < \epsilon, \quad k = 1, \cdots r, | x_{i0} - x_{i0}' | < \delta.$$

If $\sigma = \beta/(2ATr)$, then

$$| x_i(t_0 + T) | \leqq A[2N\epsilon r + \sigma \cdot Tr],$$

or, finally,

$$| x_i(t_0 + T) | < \beta, \quad \text{if} \quad | x_{i0} - x_{i0}' | < \delta.$$

Thus, if the control functions of the control system are introduced with a certain error dependent on the initial data ($| x_{i0} - x_{i0}' | < \delta$), a process which is close to the optimal process will be obtained.

4. Let the parameters $c_1, \cdots, c_e$ of the system vary in the presence of the initial data $x_{10}, \cdots, x_{n0}$. We shall prove that the optimal control functions $u_k(x_0, t_0, c, \tau)$ and the optimal control time $T(x_0, t_0, c)$ are continuous in $c_1, \cdots, c_e$.

THEOREM 2. *If the systems of functions*

$$\{\gamma_{i1}(c, \tau), \cdots, \gamma_{ir}(c, \tau)\}$$

are completely independent, and if there exists, for some values of the parameters $c_1, \cdots, c_e$, an optimal control $u_k(x_0, t_0, c, \tau)$ with optimal time $T(x_0, t_0, c)$, then, for every $\sigma > 0$ and $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\text{meas } (| u_k(x_0, t_0, c, \tau) - u_k(x_0, t_0, c', \tau) | \geqq \sigma) < \epsilon,$$
$$| T(x_0, t_0, c) - T(x_0, t_0, c') | < \epsilon$$

whenever $| c_i - c_i' | < \delta$.

Let us first derive a property of the function $F(x_0, t_0, c, t)$.

LEMMA 4.1 *For fixed $x_{10}, \cdots, x_{n0}, t_0$, and t, the function $F(x_0, t_0, c, t)$ is continuous in the system parameters $c_1, \cdots, c_e$.*

Proof. Let c_i^m be a sequence of parameters which converge to c_i as $m \to \infty$. Since

$$F(x_0, t_0, c, t) = \min \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c, \tau) \right| d\tau$$
$$= \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau) \right| d\tau$$

under the condition $\sum_{i=1}^{n} \lambda_i x_{i0} = -1$, and

$$F(x_0, t_0, c^m, t) = \min \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i \gamma_{ik}(c^m, \tau) \right| d\tau$$
$$= \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^m \gamma_{ik}(c^m, \tau) \right| d\tau$$

under the conditions

$$\sum_{i=1}^{n} \lambda_i^m x_{i0} = -1, \qquad \sum_{i=1}^{n} x_{i0}^2 \neq 0,$$

it follows that

$$0 < F(x_0, t_0, c^m, t) \leqq \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c^m, \tau) \right| d\tau.$$

Since the functions $\gamma_{ik}(c, \tau)$ are continuous in t as well as the parameters $c_1, \cdots, c_e$ (because of the theorems on the continuity of solutions of differential equations with respect to parameters [6]), the functions $\gamma_{ik}(c^m, \tau)$, for $0 \leqq \tau \leqq t$ and $m = 1, 2, \cdots$, are uniformly bounded. Hence, the numbers $F(x_0, t_0, c^m, t)$, for $m = 1, 2, \cdots$, are uniformly bounded. It follows from Lemma 3.1 that the numbers λ_i^m are also uniformly bounded (for $m = 1, 2, \cdots,$), so that we can find a subsequence m_s such that $\lambda_i^{m_s} \to \bar{\lambda}_i$ as $s \to \infty$ (for $i = 1, \cdots, n$). Consequently, since the functions $\gamma_{ik}(c^{m_s}, \tau)$ converge uniformly to $\gamma_{ik}(c, \tau)$,

$$\lim \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{m_s} \gamma_{ik}(c^{m_s}, \tau) \right| d\tau = \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \bar{\lambda}_i \gamma_{ik}(c, \tau) \right| d\tau \text{ as } s \to \infty.$$

By the definition of $F(x_0, t_0, c, t)$, we have (since $\sum_{i=1}^{n} \bar{\lambda}_i x_{i0} = -1$)

$$\int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \bar{\lambda}_i \gamma_{ik}(c, \tau) \right| d\tau \geqq \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau) \right| d\tau,$$

and, similarly,

$$\int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c^{m_s}, \tau) \right| d\tau \geqq \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^{m_s} \gamma_{ik}(c^{m_s}, \tau) \right| d\tau.$$

Passing to the limit under the integral signs, we obtain

$$\int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau) \right| d\tau \geqq \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \bar{\lambda}_i \gamma_{ik}(c, \tau) \right| d\tau,$$

so that these two integrals are equal.

Consequently, as $s \to \infty$,

$$F(x_0, t_0, c^{m_s}, t) \to \int_{t_0}^{t} \sum_{k=1}^{r} \left| \sum_{i=1}^{n} \bar{\lambda}_i \gamma_{ik}(c, \tau) \right| d\tau = F(x_0, t_0, c, t).$$

Obviously, for every convergent subsequence of the values $F(x_0, t_0, c^m, t)$, we obtain $F(x_0, t_0, c, t)$ in the limit. Since the values $F(x_0, t_0, c^m, t)$ are uniformly bounded in m, every subsequence of the values $F(x_0, t_0, c^m, t)$ converges to $F(x_0, t_0, c, t)$.

This means that $F(x_0, t_0, c, t)$ is continuous in the parameters $c_1, \cdots, c_e$ (for fixed $x_{10}, \cdots, x_{n0}, t_0, t$).

The continuity of the optimal time $T(x_0, t_0, c)$ and of the control functions $u_k(x_0, t_0, c, \tau)$ in the parameters $c_1, \cdots, c_e$ can now be derived from Lemmas 3.1 and 4.1 just as was done in the proof of Theorem 1. Therefore, we shall present only a brief proof of Theorem 2.

Let T be the optimal control time. Then, from the monotonicity and continuity in t of $F(x_0, t_0, c, t)$, it follows that

$$F(x_0, t_0, c, T - \epsilon) < \frac{1}{N} - \beta, \qquad F(x_0, t_0, c, T + \epsilon) > \frac{1}{N} + \beta,$$

where ϵ is an arbitrary, as small as desired, positive number.

If $c_i^m \to c_i$ as $m \to \infty$, then

$$| F(x_0, t_0, c, T \pm \epsilon) - F(x_0, t_0, c^m, T \pm \epsilon) | < \beta$$

when $m > M$ (Lemma 4.1).

Hence,

$$| T(x_0, t_0, c) - T(x_0, t_0, c^m) | < \epsilon \quad \text{for} \quad m > M.$$

The control functions (according to (2.4)) can be computed from the formulas

$$u_k(x_0, t_0, c, \tau) = N \operatorname{sign} \sum_{i=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau), \qquad \sum_{i=1}^{n} \lambda_i^0 x_{i0} = -1, \; t_0 \leqq \tau \leqq t_0 + T,$$

$$u_k(x_0, t_0, c^m, \tau) = N \operatorname{sign} \sum_{i=1}^{n} \lambda_i^m \gamma_{ik}(c^m, \tau), \qquad \sum_{i=1}^{n} \lambda_i^m x_{i0} = -1, \; t_0 \leqq \tau \leqq t_0 + T_m.$$

Since the $\gamma_{ik}(c, \tau)$ are continuous in $c_1, \cdots, c_e$, we may conclude on the basis of Lemma 3.1, that the λ_i^m are uniformly bounded. Therefore, repeating the arguments of Theorem 1 with insignificant modifications, we may convince ourselves that all the zeros of the functions

$$\sum_{i=1}^{n} \lambda_i^m \gamma_{ik}(c^m, \tau),$$

beginning with some number $m > M$, are found in an ϵ-neighborhood of the zeros of the functions

$$\sum_{k=1}^{n} \lambda_i^0 \gamma_{ik}(c, \tau),$$

and that, outside of this neighborhood, the signs of the functions

$$\sum_{i=1}^{n} \lambda_i{}^m \gamma_{ik}(c^m, \tau) \quad \text{and} \quad \sum_{i=1}^{n} \lambda_i{}^0 \gamma_{ik}(c, \tau) \quad \text{for} \quad m > M$$

agree, i.e., for any $\sigma > 0$,

$$\text{meas } (| u_k(x_0, t_0, c, \tau) - u_k(x_0, t_0, c^m, \tau) | \geqq \sigma) \to 0$$

as $m \to \infty$.

With this, the proof of our assertion is complete. The author considers it her duty to note that she became acquainted with optimal control problems through N. N. Krasovskii.

REFERENCES

[1] V. G. Boltyanskii, R. V. Gamkrelidze and L. S. Pontryagin, *On the theory of optimal processes*, Dokl. Akad. Nauk SSSR, 110 (1956), pp. 7–10. (English translation in Report STL-T-Ru-22-60-5111-102, Space Technology Laboratories, Redondo Beach, California.)

[2] R. V. Gamkrelidze, *On the theory of optimal processes in linear systems*, Dokl. Akad. Nauk SSSR, 116 (1957), pp. 9–11.

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

[4] N. N. Krasovskii, *On an optimal control problem*, Prikl. Mat. Meh., 21 (1956), pp. 670–677. (English translation in Report AR61-0007, Space Technology Laboratories, Redondo Beach, California.)

[5] L. A. Liusternik and V. G. Sobolev, *Elements of Functional Analysis*, F. Ungar, New York, N. Y., 1961. (Translation from the Russian.)

[6] V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton U. Press, Princeton, N. J., 1960. (Translation from the Russian.)

[7] N. Ahiezer and M. G. Krein, *Some Questions in the Theory of Moments*, Vol. II of translations of mathematical monographs, American Math. Society, Providence, R. I., 1962. (Translation from the Russian.)

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

A SUFFICIENT CONDITION IN THE THEORY OF OPTIMAL CONTROL*

E. B. LEE†

This note sets forth results which cover most of the known cases where L. S. Pontriagin's Maximum Principle [1] is a sufficient, as well as a necessary, condition for optimal control.

Consider the system

$$\dot{x}^0 = f^0(x, t) + h^0(u, t) \tag{a}$$

$$\dot{x} = A(t)x + h(u, t) \tag{b}$$

with $x(t_0) = x_0$ and $x^0(t_0) = 0$. Here f^0, h^0, A, and h are continuous in all arguments. x is the system state, an n-vector, and u is the control, an m-vector. x^0 is a scalar variable which measures the quality of control.

If $u(s)$ is any control function on the interval $[t_0, t]$ we will write the corresponding response of equations(a) and (b) as $\hat{x}_u(t) = (x_u^0(t), x_u(t))$. The control u is restricted to a set $\Omega \subset R^m$. It is assumed that either Ω is compact or that h, and h^0 are such that

$$\max_{u \in \Omega} \{ \lambda \cdot h(u, t) + \lambda^0 h^0(u, t) \}$$

exists for each $t \in [t_0, T]$ and $\hat{\lambda} = (\lambda^0, \lambda) \in R^{n+1}$ with $\lambda^0 < 0$.

It is further assumed that $f^0(x, t)$ is a single valued, convex function of x for each $t \in [t_0, T]$, that is,

$$\frac{\partial f^0}{\partial x}(x, t) \cdot (\omega - x) + f^0(x, t) \leqq f^0(\omega, t)$$

for all x, $\omega \in R^n$ and $t \in [t_0, T]$.

Definitions,

(a). $u^*(s)$ on $[t_0, T]$ is an extremal control if for some

$$\lambda(t_0) = (\lambda_0^1, \lambda_0^2 \cdots \lambda_0^n) \quad \text{and} \quad \lambda^0 = \text{constant} < 0,$$

$$\lambda^0 h^0(u^*(s), s) + \lambda(s) \cdot h(u^*(s), s) = \operatorname*{Max}_{u \in \Omega} \{\lambda^0 h^0(u, s) + \lambda(s) \cdot h(u, s)\}$$

where

* Received by the editors May 1, 1963 and in revised form June 21, 1963. The work was done at the Minneapolis-Honeywell Regulator Company with support from NASA contract NASw-563.

† Institute of Technology, University of Minnesota, Minneapolis, Minnesota.

$$\dot{\lambda} = -A'(t)\lambda - \lambda^0 \frac{\partial f^0}{\partial x}(x_{u^*}(t), t), \qquad (' \text{ denotes transpose}).$$

Here $x_{u^*}(t)$ is the response corresponding to $u^*(s)$, $t_0 \leqq s \leqq t \leqq T$.

(b). The control $u(s)$ is allowable if it is a measurable real valued vector function with range in Ω on $[t_0, T]$.

(c). The set of attainability $K(T, x_0)$ is the collection of end points of the responses $\hat{x}_u(t) = (x_u{}^0(t), x_u(t))$ which initiate at $(0, x_0)$ for all allowable controls $u(s)$ on $[t_0, T]$.

The problem of optimal control studied here is to select allowable controls $u(s)$ which "steer" the response $x_u(t)$ from the initial point x_0 at time t_0 to a prescribed target set G at time $T < \infty$ and minimize the cost functional of control $C(u) = g(x(T)) + x^0(T)$. Here g is a continuously differentiable function of x. An allowable control which provides an absolute minimum for $C(u)$ amongst the set of all allowable controls which steer $x_u(t)$ from x_0 to G is called an optimal control. Note, the free end point problem results when $G = R^n$.

Let us establish a basic inequality for this problem.

LEMMA. *Let $u^*(s)$, $t_0 \leqq s \leqq T$, be an allowable extremal control with corresponding response $\hat{x}_{u^*}(t)$ which initiates at $\hat{x}(t_0) = (0, x(t_0))$, then $\hat{\lambda}(T) \cdot \hat{x}_{u^*}(T) \geqq \hat{\lambda}(T) \cdot \hat{\omega}$ for $\lambda^0 < 0$ and all $\hat{\omega} \in K(T, x_0)$.*

Proof.

Consider

$$\frac{d(\hat{\lambda}\cdot\hat{x}_{u^*})}{dt} = \lambda^0 \dot{x}^0_{u^*} + \dot{\lambda}\cdot x_{u^*} + \lambda \cdot \dot{x}_{u^*}$$

$$= \lambda^0(f^0(x_{u^*}, t) + h^0(u^*, t)) + \left(-A'(t)\lambda - \lambda^0 \frac{\partial f^0}{dx}(x_{u^*}, t)\right)\cdot x_{u^*} + \lambda\cdot(A(t)\, x_{u^*} + h(u^*, t)).$$

Upon integrating both sides between t_0 and T we obtain

$$\lambda^0 x^0_{u^*}(T) + \lambda(T)\cdot x_{u^*}(T) - \lambda(t_0)\cdot x(t_0) = \int_{t_0}^{T} \left\{\lambda^0\left(f^0(x_{u^*}(t), t) - \frac{\partial f^0}{\partial x}(x_{u^*}(t), t)\cdot x_{u^*}(t)\right) + \lambda^0 h^0(u^*(t), t) + \lambda(t)\cdot h(u^*(t), t)\right\} dt.$$

Let $x_u(t)$ be any other response with initial value $x_0 = x(t_0)$ for which we calculate

$$\lambda^0 x_u{}^0(T) + \lambda(T)\cdot x_u(T) - \lambda(t_0)\cdot x(t_0) = \int_{t_0}^{T} \left\{\lambda^0\left(f^0(x_u(t), t) - \frac{\partial f^0}{\partial x}(x_{u^*}(t), t)\cdot x_u(t)\right) + \lambda^0 h^0(u(t), t) + \lambda(t)\cdot h(u(t), t)\right\} dt.$$

But

$$\lambda^0 h^0(u^*(t), t) + \lambda(t) \cdot h(u^*(t), t) \geqq \lambda^0 h^0(u(t), t) + \lambda(t) \cdot h(u(t), t).$$

Thus if

$$\lambda^0\left(f^0(x_{u^*}(t), t) - \frac{\partial f^0}{\partial x}(x_{u^*}(t), t) \cdot x_{u^*}(t)\right) \geqq \lambda^0\left(f^0(x_u(t), t) - \frac{\partial f^0}{\partial x}(x_{u^*}(t), t) \cdot x_u(t)\right),$$

we obtain the desired inequality. This is certainly true if $\lambda^0 < 0$ and

$$\frac{\partial f^0}{\partial x}(x_{u^*}(t), t) \cdot x_{u^*}(t) - f^0(x_{u^*}(t), t) \geqq \frac{\partial f^0}{\partial x}(x_{u^*}(t), t) \cdot x_u(t) - f^0(x_u(t), t),$$

which is the convexity condition on f^0.

Thus we have

$$\lambda^0 x_{u^*}{}^0(T) + \lambda(T) \cdot x_{u^*}(T) \geqq \lambda^0 x_u{}^0(T) + \lambda(T) \cdot x_u(T)$$

or

$$\hat{\lambda}(T) \cdot \hat{x}_{u^*}(T) \geqq \hat{\lambda}(T) \cdot \hat{\omega}$$

all $\hat{\omega} \in K(T, x_0)$ and the lemma is established.

The basic inequality of the lemma enables us to establish the sufficiency of the maximum principle in a number of cases. These results are summarized as

THEOREM.

(A). *Consider the cost functional of control* $C(u) = x^0(T)$ *and as target set* G *a point* x_1. *Let* $u^*(s)$, $t_0 \leqq s \leqq T$, *be an allowable extremal control which steers the corresponding response* $x_{u^*}(t)$ *from* x_0 *at* t_0 *to* x_1 *at* T, *then* $u^*(s)$ *is an optimal control.*

(B). *Consider the cost functional* $C(u) = g(x(T)) + x^0(T)$ *with* $g(x)$ *a convex function of* x *and consider the target set* $G = R^n$, (*this is the free end point problem*). *Let* $x(t_0) = x_0$. *Then* $u^*(s)$, $t_0 \leqq s \leqq T$, *is an optimal control if it is an allowable extremal control with* $\hat{\lambda}(T) = (-1, -\frac{\partial g}{\partial x}(x_{u^*}(T)))$, (*The condition on* $\hat{\lambda}(T)$ *is a so called transversality condition*).

(C). *Consider the cost functional* $C(u) = x^0(T)$ *and the convex, closed, target set* $G = \{x | \gamma(x) \leqq c\} \subset R^n$, *where* γ *is differentiable and* c *a constant. Let* $u^*(s)$, $t_0 \leqq s \leqq T$, *be an allowable extremal control which steers* $x_{u^*}(t)$ *from* x_0 *at* t_0 *to* $x_1 \in G$ *at* T *with* $\lambda(T)$ *an interior normal*[1] *to* G *at* $x_{u^*}(T)$ *on*

[1] λ is an interior normal to G at x on ∂G if the vector λ is orthogonal to a support plane of G at x and is directed into the halfspace containing G. Thus G need not have an interior to have interior normals. Note if G does not have an interior we can still approximate it by a $\gamma(x)$.

∂G, then $u^*(s)$ is optimal if such a control exists. (If there is no such $u^*(s)$ then the minimum may occur interior to G in which case (B) applies with $\hat{\lambda}(T) = (-1, 0, 0 \cdots 0)$ and if G is just one point part (A) is obtained).

Proof.

(A). From the lemma

$$\hat{\lambda}(T) \cdot \hat{x}_{u^*}(T) \geqq \hat{\lambda}(T) \cdot \hat{\omega} \quad \text{for} \quad \lambda^0 < 0 \quad \text{all} \quad \hat{\omega} \in K(T, x_0).$$

Thus

$$\lambda(T) \cdot x_{u^*}(T) + \lambda^0 x_{u^*}{}^0(T) \geqq \lambda(T) \cdot x_u(T) + \lambda^0 x_u{}^0(T).$$

But, comparing only those responses that end at x_1, that is, those for which $x_{u^*}(T) = x_u(T) = x_1$, the basic inequality becomes

$$\lambda^0 x_{u^*}{}^0(T) \geqq \lambda^0 x_u(T).$$

Since $\lambda^0 < 0$ we have $C(u^*) = x_{u^*}{}^0(T) \leqq x_u{}^0(T) = C(u)$ and therefore $u^*(s)$ is optimal.

(B) With $\hat{\lambda}(T) = \left(-1, \dfrac{-\partial g}{\partial x}(x_{u^*}(T))\right)$ the inequality of the lemma is

$$\frac{-\partial g}{\partial x}(x_{u^*}(T)) \cdot x_{u^*}(T) - x^0_{u^*}(T) \geqq \frac{-\partial g}{\partial x}(x_{u^*}(T)) \cdot x_u(T) - x_u{}^0(T)$$

Adding and subtracting $g(x_{u^*}(T))$ on the left side and $g(x_u(T))$ on the right side the last inequality becomes

$$-x^0_{u^*}(T) - g(x_{u^*}(T)) + g(x_{u^*}(T)) - \frac{\partial g}{\partial x}(x_{u^*}(T)) \cdot x_{u^*}(T) \geqq$$

$$-x_u{}^0(T) - g(x_u(T)) + g(x_u(T)) - \frac{\partial g}{\partial x}(x_{u^*}(T)) \cdot x_u(T).$$

But,

$$\frac{\partial g}{\partial x}(x_{u^*}(T)) \cdot (x_{u^*}(T) - x_u(T)) + g(x_u(T)) \geq g(x_{u^*}(T))$$

if g is a connex function of x. Therefore

$$-C(u^*) = -x_{u^*}{}^0(T) - g(x_{u^*}(T)) \geqq -x_u{}^0(T) - g(x_u(T)) = -C(u),$$

or $C(u^*) \leqq C(u)$. Hence part (B) is established.

(C). Assume for simplicity that γ was picked to be a convex function on G with $\partial G = \{x \mid \gamma(x) = c\}$. Consider only boundary points $x_{u^*}(T)$ where it is required that $\lambda(T) = k\left\{-\dfrac{\partial \gamma}{\partial x}(x_{u^*}(T))\right\}$ in order for $\lambda(T)$ to be an interior normal to G at $x_{u^*}(T)$ on ∂G, (let $k = 1$).

The inequality of the lemma can then be written

$$\hat{\lambda}(T)\cdot\hat{x}_{u^*}(T) = \lambda^0 x^0_{u^*}(T) + \left\{-\frac{\partial\gamma}{\partial x}(x_{u^*}(T))\right\}\cdot x_{u^*}(T) \geqq \hat{\lambda}(T)\cdot\hat{x}_u(T)$$

$$= \lambda^0 x_u{}^0(T) + \left\{-\frac{\partial\gamma}{\partial x}(x_{u^*}(T))\right\}\cdot x_u(T)$$

If $x_{u^*}(T)$ is on the boundary of G it is also true that $\gamma(x_{u^*}(T)) = c \geqq \gamma(x_u(T))$ for all allowable responses $x_u(T)$. Adding the last two inequalities we obtain

$$\lambda^0 x^0_{u^*}(T) - \frac{\partial\gamma}{\partial x}(x_{u^*}(T))\cdot x_{u^*}(T) + \gamma(x_{u^*}(T)) \geqq \lambda^0 x_u{}^0(T)$$

$$- \frac{\partial\gamma}{\partial x}(x_{u^*}(T))\cdot x_u(T) + \gamma(x_u(T)).$$

But again if $x_{u^*}(T)$ in on ∂G and $x_u(t)$ is in the connex set G we have

$$\frac{\partial\gamma}{\partial x}(x_{u^*}(T)) \cdot [x_{u^*}(T) - x_u(T)] + \gamma(x_u(T)) \geq \gamma(x_{u^*}(T))$$

and therefore

$$C(u^*) \leqq C(u).$$

Remarks. If the set of attainability is closed there will exist optimum control provided there is at least one control that steers the response to the desired end point $x_1 \in G$, assuming G is also closed. The property of closure is discussed in [2] in which a bibliography and discussion of cases are presented. The set of attainability is also known to be closed if $h(u, t) = B(t)u$, $f^0(x, t) = x \cdot W(t)x$ and $h^0(u, t) = u \cdot U(t)u$, for $W(t)$, $U(t)$ positive definite on $[t_0, T]$.

When the set of attainability is closed, in the above case, the inequality of the lemma establishes that its lower (exterior normal with $\lambda^0 < 0$) surface is convex. For if it was otherwise we could be led to a contradiction of the maximum principle. Note that the transversality condition [1] follows from the established inequality of the lemma since $\hat{\lambda}(T)$ must be an exterior normal of $\bar{K}(T, x_0)$ at the corresponding response end point, $\hat{x}_{u^*}(T)$.

REFERENCES

[1] L. S. Pontriagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Control Processes*, Interscience, New York, 1962.

[2] L. W. Neustadt, *The existence of optimal controls in the absence of convexity condition*, J. Math. Anal. Appl. (to appear) 1963.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

GENERALIZED CURVES AND THE EXISTENCE OF OPTIMAL CONTROLS*

R. A. GAMBILL†

Introduction. Consider the following problem in optimal control: Determine if there exists a vector function $u(t)$ belonging to a prescribed class of functions $\mathfrak{A}$, which minimizes the functional

$$I(u) = \int_{t_0}^{t_1} f(t, x, u)\, dt + f^0(t_0, x(t_0), t_1, x(t_1)) \tag{1}$$

subject to the conditions;

(a) differential equations

$$\dot{x} = g(t, x, u), \qquad \cdot = d/dt \tag{2}$$

(b) inequalities

$$h(t, x, u) \geqq 0 \tag{3}$$

(c) end conditions

$$(t_0, x(t_0)) \in E_0, \qquad (t_1, x(t_1)) \in E_1, \tag{4}$$

E_0 and E_1 being prescribed closed sets in $n + 1$ dimensional number space.

x is an n-vector (the state vector) $x = (x^1, \cdots, x^n)$, u an m-vector (the control vector) $u = (u^1, \cdots, u^m)$, g an n-vector, h a q-vector ($q \geqq 0$), t will be called time. We restrict the class of functions $\mathfrak{A}$ to be the class of bounded measurable functions, that is, each component of the vector $u(t)$ is a bounded measurable function defined on the interval $[t_0, t_1]$. Also, as soon as a particular control function $u(t) \in \mathfrak{A}$ is chosen, we consider only the resulting absolutely continuous solutions of (2), satisfying (3) and (4) (if such solutions exist).

This problem has received attention in several recent papers, we mention here only a few, [2–4, 10, Chap. III, 11, 13]. Very briefly, the nature of these theorems is as follows: If the functions f, f^0, g, h satisfy specified smoothness properties and are of a certain form and if there is one function

* Received by the editors May 18, 1963 and in revised form August 5, 1963. This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research, Office of Aerospace Research, under Contract No. AF 49(638)-382, in part by the National Aeronautics and Space Administration under Contract No. NASr-103, and in part by the National Science Foundation under contract No. NSF-6-9666. Reproduction in whole or in part is permitted for any purpose of the United States Government.

† Department of Mathematics, Purdue University, West Lafayette, Indiana

$u \in \mathfrak{A}$ for which the conditions (2), (3), (4) are satisfied, then there is a function $u^* \in \mathfrak{A}$ and satisfying (2), (3), (4) for which the functional $I(u)$ attains its minimum with respect to all such functions u. It seems that the most difficult step in such theorems is to show that the set of all solutions of (2) obtained by letting u vary in $\mathfrak{A}$, and satisfying (3), (4), form a closed set.

It is the purpose of the present note to examine these theorems from the point of view of the direct methods of the calculus of variations, in particular, employing the concept of "generalized curve" invented by L. C. Young [12], and developed further by E. J. McShane to study problems of Bolza in parametric form [5–7, 9]. It is a rather striking property of the space of generalized curves that a topology may be defined on the space which leaves every integral of the calculus of variations continuous while at the same time preserving compactness of certain sets of generalized curves (§1). Thus, since differential equations (2) and inequalities (3) can be expressed as isoperimetric conditions (§2), the problem mentioned above concerning closure of the solution space of (2) is immediately solved in the space of generalized curves.

A simple observation is needed to transform the above problem directly into the type of problem which we shall analyze. Put

$$u^j(t) = \dot{x}^{j+n}(t), x^{j+n}(t_0) = 0, \qquad j = 1, 2, \cdots, m.$$

We now identify $x(t)$ as an $n + m$ vector

$$\begin{aligned} x(t) &= (x^1(t), \cdots, x^n(t), x^{n+1}(t), \cdots, x^{n+m}(t)), \\ \dot{x}(t) &= (\dot{x}^1(t), \cdots, \dot{x}^n(t), \dot{x}^{n+1}(t), \cdots, \dot{x}^{n+m}(t)) \\ &= (\dot{x}^1(t), \cdots, \dot{x}^n(t), u^1(t), \cdots, u^m(t)). \end{aligned}$$

The sets E_0, E_1 of (4) are replaced by closed sets $\bar{E}_0 = E_0 \times 0$, $\bar{E}_1 = E_1 \times V$, where 0 is the m-dimensional 0-vector and V is a closed set in m-dimensional number space containing all the points $(x^{n+1}(t_1), \cdots, x^{n+m}(t_1))$, for $u \in \mathfrak{A}$. We may now pose the above problem in the following way: Let K_0 represent the class of absolutely continuous vector functions $x(t)$ satisfying

(a) differential equations

$$(2') \qquad \dot{x}^j = g^j(t, x^1, \cdots, x^n, \dot{x}^{n+1}, \cdots, \dot{x}^{n+m}) \; j = 1, 2, \cdots, n.$$

(b) differential inequalities

$$(3') \qquad \begin{aligned} &h^j(t, x^1, \cdots, x^n, \dot{x}^{n+1}, \cdots, \dot{x}^{n+m}) \geqq 0, \\ &j \in I, \qquad I \text{ a countable index set.} \end{aligned}$$

(c) end conditions

$$(4') \qquad (t_0, x(t_0)) \in \bar{E}_0, \qquad (t_1, x(t_1)) \in \bar{E}_1.$$

Determine if there exists a function $x^*(t)$ in K_0 which gives the functional

$$(1') \qquad I = \int_{t_0}^{t_1} f(t, x^1, \cdots, x^n, \dot{x}^{n+1}, \cdots, \dot{x}^{n+m})\, dt + f^0(t_0, x^1(t_0), \cdots, x^n(t_0), t_1, x^1(t_1), \cdots, x^n(t_1))$$

its least value with respect to all functions $x(t)$ in K_0.

Generally, the results of this note are simple translations of results already obtained by McShane [5] for generalized curves in parametric form. Thus, we will use the notation of that paper and of the paper of Botts [1] in our development. The proofs of many of the lemmas and theorems which we state are already contained in the above mentioned papers, and we shall often appeal to a particular paper and page number for a proof. It is not the purpose of this note to establish new theorems, but only perhaps to draw attention to this method in the calculus of variations of analyzing problems in optimal control.

1. Properties of generalized curves. Let x, r be ν-vectors and t a real number. Let $|r|$ stand for the Euclidean length of the vector r. Let Q be the set of all continuous real valued functions $\varphi(r)$ defined for all r. We consider a class of functionals $\mathfrak{M}$ defined on Q with the following properties:

(5) $\mathfrak{M}$ is *linear*; that is, for each pair of real numbers α_1, α_2 and each pair of functions φ_1, $\varphi_2 \in Q$, $\mathfrak{M}(\alpha_1\varphi_1 + \alpha_2\varphi_2) = \alpha_1 \mathfrak{M}(\varphi_1) + \alpha_2 \mathfrak{M}(\varphi_2)$.

(6) $\mathfrak{M}(1) = 1$

(7) $\mathfrak{M}$ is *non-negative*; that is, if $\varphi \geqq 0$ then $\mathfrak{M}(\varphi) \geqq 0$

(8) $\mathfrak{M}$ is *determined* by the values of φ on a bounded set $B(\mathfrak{M})$ of values r; that is, if φ_1, $\varphi_2 \in Q$ and $\varphi_1(B) = \varphi_2(B)$ then $\mathfrak{M}(\varphi_1) = \mathfrak{M}(\varphi_2)$.

The following two lemmas are immediate

LEMMA 1. *$\mathfrak{M}$ is determined by the values of φ on B if and only if $\mathfrak{M}(\varphi) = 0$ for every function $\varphi \in Q$ which vanishes on B.*

LEMMA 2.

(a) $\mathfrak{M}(\varphi_1) \leqq \mathfrak{M}(\varphi_2)$ *if* $\varphi_1(r) \leqq \varphi_2(r)$ *for all* $r \in B$.

(b) $|\mathfrak{M}(\varphi)| \leqq \mathfrak{M}(|\varphi|)$

(c) $|\mathfrak{M}(\varphi)| \leqq lub_{r \in B} |\varphi|$

(d) *if* $\varphi_n \to \varphi$ *uniformly on* B, *then* $\lim_{n\to\infty} \mathfrak{M}(\varphi_n) = \mathfrak{M}(\varphi)$.

We are now ready to define a generalized curve. For background and a heuristic discussion of these objects, see [5; pp. 513–515].

DEFINITION 1. A generalized curve (hereafter referred to as GC) is a system consisting of a vector function $x(t) = (x^1(t), \cdots, x^\nu(t))$, $t_0 \leqq t \leqq t_1$, a subset M of the interval $[t_0, t_1]$ of measure $t_1 - t_0$, and a one

parameter family of functionals $\mathfrak{M}(t; \varphi)$ defined for all t in M and all φ in Q with the properties;

(9) The functions $x^i(t)$ are absolutely continuous on $[t_0, t_1]$ with finite derivatives on M.

(10) For each t in M, $\mathfrak{M}(t; \varphi)$ satisfies (5) – (8).

(11) For each t in M, and for $\varphi(r) = r^i$, $\mathfrak{M}(t; \varphi) = \dot{x}^i(t)$, $i = 1, 2, \cdots, \nu$ (we write $\mathfrak{M}(t; r^i) = \dot{x}^i(t)$).

(12) For each φ in Q, $\mathfrak{M}(t; \varphi)$ is measurable on M.

The closure of M is denoted by $\bar{M}$.

We denote a GC by

$$C^*: [x(t), \mathfrak{M}(t; \varphi), M].$$

The ordinary curve in $\nu + 1$ space defined by the functions $x^i(t)$, $t_0 \leqq t \leqq t_1$, $i = 1, 2, \cdots, \nu$ is called the track of C^*.

DEFINITION 2. Let $F(t, x, r)$ be a function defined and continuous for all x, r and for all t; $t_0 \leqq t \leqq t_1$. Let $C^*: [x(t), \mathfrak{M}(t; \varphi), M]$ be a GC. Then

$$J(C^*) = \int_{t_0}^{t_1} \mathfrak{M}(t; F(t, x(t), r))\, dt,$$

provided the integral exists. We adopt the convention that the function $\mathfrak{M}(t; F(t, x(t), r)) = 0$ on the complement of M.

REMARK: If one considers the space of absolutely continuous vector functions $x(t) = (x_1(t), \cdots, x_n(t))$ on $[t_0, t_1]$, with the uniform metric, then the only integrals

$$\int_{t_0}^{t_1} F(t, x(t), \dot{x}(t))\, dt$$

which are continuous are those whose integrand functions F are linear in $\dot{x}(t)$. With the introduction of generalized curves, one can make every integral continuous (definition 4) while still retaining compactness (Theorem 1). A generalized curve is essentially an absolutely continuous function, together with an "average value" $\mathfrak{M}$ computed at each value of $F(t, x(t), \dot{x}(t))$, with this value then being used to compute the integral of F along C^*. The following example will illustrate: Let x_n be a sequence of real valued continuous functions on [0, 1] defined as follows:

$$x_n(t) = \begin{cases} 0 & t = 0, \dfrac{2}{2n}, \dfrac{4}{2n}, \cdots, 1 \\ \dfrac{1}{2n} & t = \dfrac{1}{2n}, \dfrac{3}{2n}, \cdots, \dfrac{2n-1}{2n} \\ \text{linear on } \left[\dfrac{k}{2n}, \dfrac{k+1}{2n}\right] & k = 0, 1, \cdots, 2n-1. \end{cases}$$

Then $\dot{x}_n(t)$ is alternately $+1$ and -1. If $F(t, x, \dot{x})$ is any continuous integrand, then

$$\lim_{n\to\infty} \int_0^1 F(t, x_n(t), \dot{x}_n(t))\, dt = \frac{1}{2}\int_0^1 [F(t, 0, 1) + F(t, 0, -1)]\, dt.$$

For each function $x_n(t)$ we define

$$\mathfrak{M}_n(t, F(t, x_n(t), r)) = F(t, x_n(t), \dot{x}_n(t)), n = 1, 2, \cdots$$

that is, at each point t, the value of $\mathfrak{M}_n$ is simply the value of F with the derivative of x_n inserted in the third argument. [In case this is the rule for computing the functional $\mathfrak{M}$, the generalized curve will be called an ordinary curve (definition 8 and lemma 10)]. For the curve $x_0\ (t) = 0\ 0 \leqq t \leqq 1$, we define $\mathfrak{M}_0\ (t; F(t, x_0(t), r)) = \frac{1}{2}[F(t, x_0(t), 1) + F(t, x_0(t), -1)]$,

Then

$$\lim_{n\to\infty} \int_0^1 \mathfrak{M}_n(t; (Ft, x_n(t), r))\, dt = \int_0^1 \mathfrak{M}_0(t; F(t, x_0(t), r))dt$$

Let $|1, r|$ stand for $(1 + \sum_{i=1}^{\nu} r^{i^2})^{1/2}$, then

Definition 3. If $\mathfrak{M}(t; |1, r|)$ is summable, we say that C^* is rectifiable and its length is given by

$$L(C^*) = \int_{t_0}^{t_1} \mathfrak{M}(t; |\ 1, r\ |)\, dt$$

Lemma 3. *If C^* is rectifiable, then so is its track and*

$$\int_{t_0}^{t_1} |\ 1, \dot{x}(t)|\, dt \leqq \int_{t_0}^{t_1} \mathfrak{M}(t; |\ 1, r\ |)dt. \qquad [1, \text{p. } 376].$$

Lemma 4. *If C^* is a GC, and $\varphi(t, r)$ is defined and continuous for all $t \in [t_0, t_1]$ and all r, then $\mathfrak{M}(t; \varphi(t, r))$ is measurable on M.* [5; p. 517].

We assume henceforth that the determining sets for our class of GC's are uniformly bounded, that is,

(13) there is a bounded set $B \subset R^{\nu}$ such that for each C^*:$[x(t), \mathfrak{M}(t; \varphi), M]$ and for each t in M, B is a determining set for $\mathfrak{M}(t; \varphi)$.

This assumption implies that the derivative of the track of C^* is bounded on M, and that C^* is rectifiable.

Remark 1: The assumption (13) is not so restrictive as applied to control problems. For example, we shall see that if the control vector of the introduction is constrained to lie in a bounded set of R^m, the corresponding determining sets for the problem are automatically uniformly bounded.

Lemma 5. *If C^* is a* GC *and $F(t, x, r)$ is continuous on $[t_0, t_1] \times R^{\nu} \times R^{\nu}$, and* (13) *holds, then the integral*

$$J(C^*) = \int_{t_0}^{t_1} \mathfrak{M}(t; F(t, x(t), r))\, dt \text{ exists.} \qquad [1, \text{p. } 376].$$

REMARK 2: If (13) does not hold but it is supposed only that C^* is rectifiable, then the conclusion of Lemma 5 does not follow. For example consider the ordinary curve (in E^2) $x(t) = \sqrt{t(1-t)} + \tan^{-1}\sqrt{t/1-t}$ $0 \leqq t \leqq 1$, then $\int_0^1 \sqrt{1+\dot{x}^2(t)}\,dt = 2$, however $\int_0^1 (1+\dot{x}^2(t))$ does not exist.

We shall presently discuss families of GC's with the property that if $C_1^*:[x_1(t), \mathfrak{M}_1(t,\varphi), M_1]$, $C_2^*:[x_2(t), \mathfrak{M}_2(t,\varphi), M_2]$ are two members of the family, the intervals $\bar{M}_1$, $\bar{M}_2$ may be different. However, each family we shall discuss will have the property;

(14) there is an interval $[t_0^*, t_1^*]$ such that the set M for every member of the family is contained in $[t_0^*, t_1^*]$.

If $C^*:[x(t), \mathfrak{M}(t,\varphi), M]$, $\bar{M} = [t_0, t_1]$ is a GC and $[t_0^*, t_1^*]$ is any interval with the property $t_0^* \leqq t_0 \leqq t_1 \leqq t_1^*$, it will be convenient to extend the definition of C^* to the set $[t_0^*, t_0] \cup M \cup [t_1, t_1^*] = \tilde{M}$ in the following way: Define

$$\tilde{x}(t) \begin{cases} x(t) & t \in [t_0, t_1] \\ x(t_0) & t \in [t_0^*, t_0] \\ x(t_1) & t \in [t_1, t_1^*] \end{cases}$$

$$\tilde{\mathfrak{M}}(t,\varphi) = \begin{cases} \mathfrak{M}(t,\varphi) & t \in M \\ \varphi(0) & t \in \tilde{M} - M. \end{cases}$$

With this definition, the system $\tilde{C}^*:[\tilde{x}(t), \tilde{\mathfrak{M}}(t;\varphi), \tilde{M}]$ is also a GC.

We adopt the following definition of convergence of a sequence of generalized curves [5; p. 518].

DEFINITION 4: The GC $C_0^*:[x_0(t), \mathfrak{M}_0(t,\varphi), M_0]$, $\bar{M}_0 = [t_0, t_1]$ is the limit of the sequence $\{C_n^*:[x_n(t), \mathfrak{M}_n(t,\varphi), M_n]$, $\bar{M}_n = [t_{0n}, t_{1n}]\}$ of generalized curves if

(a) $$\lim_{n\to\infty} \{|t_{0n} - t_0| + |t_{1n} - t_1| + \sup_{t\in[t_0^*, t_1^*]} |\tilde{x}_n(t) - \tilde{x}_0(t)|\} = 0$$

(b) for every integrand $F(t, x, r)$ which is continuous on $[t_0^*, t_1^*] \times R^\nu \times R^\nu$

$$\lim_{n\to\infty} \int_{t_{0n}}^{t_{1n}} \mathfrak{M}_n(t; F(t, x_n(t), r))\,dt = \int_{t_0}^{t_1} \mathfrak{M}_0(t; F(t, x_0(t), r))\,dt.$$

DEFINITION 5: The generalized curves $C_1^*:[x_1(t), \mathfrak{M}_1(t;\varphi), M_1]$,

$$\bar{M}_1 = [t_{01}, t_{11}],\ C_2^*:[x_2(t), \mathfrak{M}_2(t;\varphi), M_2],\ \bar{M}_2 = [t_{02}, t_{12}]$$

are *identical* if

(a) $$|t_{01} - t_{02}| + |t_{11} - t_{12}| + \sup_{t\in[t_{01}, t_{02}]} |x_1(t) - x_2(t)| = 0$$

(b) for every integrand $F(t, x, r)$ which is continuous on $[t_{01}, t_{11}] \times R^\nu \times R^\nu$

$$\int_{t_{01}}^{t_{11}} \mathfrak{M}_1(t; F(t, x_1(t), r))\, dt = \int_{t_{02}}^{t_{12}} \mathfrak{M}_2(t; F(t, x_2(t), r))\, dt.$$

With the definition 4, if $F(t, x, r)$ is continuous, then the integral $J(C^*)$ considered as a functional on the space of generalized curves, is continuous.

The following lemma which will be useful in the proof of the compactness theorem is immediate.

LEMMA 6. *Referring to definition* 4, *if* (a) *is satisfied, and if*

$$\operatorname{limit}_{n\to\infty} \int_{t_0^*}^{t_1^*} \tilde{\mathfrak{M}}_n(t; F(t, \tilde{x}_n(t), r))\, dt = \int_{t_0^*}^{t_1^*} \tilde{\mathfrak{M}}_0(t, F(t, \tilde{x}_0(t), r)\, dt$$

for every function $F(t, x, r)$ *continuous on* $[t_0^*, t_1^*] \times R^\nu \times R^\nu$, *then* (b) *is satisfied.*

THEOREM 1. *Let* A *be a compact set in* R^ν, *and let* T *be a compact interval of the real line. The set of generalized curves whose tracks lie in* $T \times A$ *and which satisfy* (13) *is sequentially compact.*

Proof: Let $\{C_n^* : [x_n(t), \mathfrak{M}_n(t; \varphi), M_n], \bar{M}_n = [t_{0n}, t_{1n}]\}$ be a sequence of generalized curves satisfying the hypotheses of the theorem. Let $B \subset R^\nu$ be a *compact* set which serves as a determining set for each C_n^*. Let $T : \{t : t_0^* \leqq t \leqq t_1^*\}$. For convenience denote the set $T \times A \times B$ by W.

Since $T \times A$ is compact, we may select a subsequence of the C_n^*, retaining the notation of the original sequence, so that $\{t_{0n}, x_n(t_{0n})\}$ converges to a limit as $n \to \infty$.

Select a further subsequence, (retaining the notation) so that $\{t_{1n}, x_n(t_{1n})\}$ converges to a limit as $n \to \infty$.

Extend each C_n^* in this last subsequence to $\tilde{C}_n^*$. If we can extract a subsequence of the $\tilde{C}_n^*$ which converge to a limit GC $\tilde{C}_0^*$, then by Lemma 6, the corresponding C_n^* will converge to C_0^*. Because of the hypotheses of the theorem we need to consider only those integrands $F(t, x, r)$ which are defined and continuous on the compact set W. With this in mind, and with minor changes in notation, the remainder of the proof follows exactly as in [5, pp. 522–525], and will not be repeated here.

2. Existence theorems. Let K denote a specified class of generalized curves. Let $F(t, x, r)$ be a continuous integrand. Let m denote the g.l.b. of $J(C^*) = \int_{t_0}^{t_1} \mathfrak{M}(t; F(t, x(t), r))\, dt$ for $C^* \in K$.

DEFINITION 6: A "minimizing sequence" is a sequence of generalized curves C_n^*, $C_n^* \in K$ such that $\lim_{n\to\infty} J(C_n^*) = m$. The following assumption is basic for our purposes

(15) There exists a minimizing sequence $\{C_n^*\}$ satisfying (13), (14) and such that the track of every C_n^* lies in a compact set $[t_0^*, t_1^*] \times E$, $E \subset R^\nu$.

THEOREM 2. *Let K be a closed class of generalized curves. Let $F(t, x, r)$ be continuous on $[t_0^*, t_1^*] \times E \times R^\nu$. Let (15) hold. Then if K is not empty it contains a* GC C_0^* *such that $J(C_0^*) \leqq J(C^*)$ for all C^* in K.*

Proof: Let $\{C_n^*\}$ be a minimizing sequence. By Theorem 1, there is a subsequence (suppose it to be the whole sequence) which converges to a GC C_0^*. $C_0^* \in K$ since K is closed, and

$$\lim_{n\to\infty} J(C_n^*) = J(C_0) = m$$

since the integral is continuous. [5, p. 527]

REMARK 3: Let g be a function defined and continuous on a closed set G in $R^{2\nu+2}$. Let the hypotheses of Theorem 2 be satisfied. For each curve $C^*:[x(t), \mathfrak{M}(t; \varphi), M]$, $\bar{M} = [t_0, t_1]$, let $(t_0, x(t_0), t_1, x(t_1)) \in G$. Then if K is not empty, it contains a GC C_0^* for which the functional $J(C^*) + g(t_0, x(t_0), t_1, x(t_1))$ assumes its minimum value relative to the class K, since $t_0, x(t_0), t_1, x(t_1)$ are continuous functionals on the class K.

We investigate now some conditions under which the hypotheses of Theorem 2 are satisfied. If $G(t, x, r)$ is a given continuous integrand, let K be the class of all generalized curves satisfying

$$I(C^*) = \int_{t_0}^{t_1} \mathfrak{M}(t; G(t, x(t), r))\, dt = \gamma$$

where γ is a given constant (isoperimetric condition). By the continuity of every integral, if $I(C_n^*) = \gamma$ and $\lim_{n\to\infty} C_n^* = C_0^*$, then $I(C_0^*) = \gamma$, thus the class K is *closed*.

A differential equation may be written as an isoperimetric condition with the following definition.

DEFINITION 6: If $C^*:[x(t), \mathfrak{M}(t; \varphi), M]$ is a GC and $f(t, x, r)$ is continuous, then C^* satisfies the differential equation

$$f(t, x, r) = 0 \tag{16}$$

if

$$\int_{t_0}^{t_1} \mathfrak{M}(t; |f(t, x(t), r)|)\, dt = 0. \tag{17}$$

Thus, the class of generalized curves which satisfy (16) is a closed class.

REMARK 4: A differential inequality $h(t, x, r) \geqq 0$ can be written as a differential equation by the following device: Put $f(t, x, r) \equiv \min [0, h(t, x, r)] = 0$. Thus, any class of generalized curves which satisfies a differential inequality is a closed class. Since the intersection of any number

of closed classes of generalized curves is closed, it follows that the class of generalized curves which satisfies any collection of differential equations and inequalities is closed [5].

Assumption (15) will be taken up in the next section.

3. Vectors carried by a GC. The concept of "vectors carried" by a GC [5, p. 529] is useful in formulating existence theorems for optimal controls. The "vectors carried" are the points r used in computing the function $\mathfrak{M}(t; \varphi)$. To be precise:

DEFINITION 7: If $C^*:[x(t), \mathfrak{M}(t; \varphi), M]$ is a GC and $\bar{t} \in M$, the vector $\bar{r}$ is *carried* by C^* at $x(\bar{t})$ if for every non-negative function φ such that $\varphi(\bar{r}) > 0$ the inequality $\mathfrak{M}(\bar{t}, \varphi) > 0$ holds.

LEMMA 7. *The set of vectors carried at $x(\bar{t})$ is closed and bounded* [5, p. 530].

LEMMA 8. *For each $\bar{t} \in M$, the value of $\mathfrak{M}(\bar{t}; \varphi)$ is determined by the values of φ on the set of vectors carried by C^* at $x(\bar{t})$* [5, p. 530].

LEMMA 9. *The* GC $C^*:[x(t), \mathfrak{M}(t; \varphi), M]$ *satisfies the differential equation $f(t, x, r) = 0$ if and only if for almost all t in M the equation*

$$f(t, x(t), r) = 0 \tag{18}$$

is satisfied for all vectors r carried at $x(t)$ [5, p. 530–531].

4. Existence theorems for optimal controls. In this section, we first state conditions which assure the existence of a minimizing GC for problem II of the introduction. These conditions are then strengthened so that the minimizing GC for that problem is an ordinary curve or, in the language of problem I of the introduction, that an optimal control exists. The relationship between ordinary curves and a subset of generalized curves must first be established.

DEFINITION 8: A GC $C^*:[x(t), \mathfrak{M}(t; \varphi), M]$ is an *isomorph of an ordinary curve* if for almost all t in M, the value of the functional $\mathfrak{M}(t; \varphi)$ depends on the value of φ at a single point r_t. [5, p. 519].

LEMMA 10. *If the ordinary curves C and the isomorphs C^* of ordinary curves are put into correspondence by letting each C^* correspond to its own track the correspondence is one-one, and for every continuous integrand $F(t, x, r)$,*

$$\int_{t_0}^{t_1} \mathfrak{M}(t; F(t, x(t), r))\, dt = \int_{t_0}^{t_1} F(t, x(t), \dot{x}(t))\, dt. \tag{19}$$

Proof: If $\mathfrak{M}(t; \varphi)$ depends only on the value of $\varphi(r)$ at $r = r_t$, then by (10) $\mathfrak{M}(t, \varphi) = \varphi(r_t)$, and by (11), if $\varphi(r) = r^i$, then $r_t{}^i = \dot{x}^i(t)$. Hence, if $F(t, x, r)$ is a continuous integrand, we have

$$\mathfrak{M}(t; F(t, x(t), r)) = \mathfrak{M}(t; F(t, x(t), r^1, \cdots, r^\nu))$$
$$= F(t, x(t), r_t^1, \cdots, r_t^\nu) = F(t, x(t), \dot{x}^1(t), \cdots, \dot{x}^\nu(t))$$
$$= F(t, x(t), \dot{x}(t)).$$

This holds for almost all t in (t_0, t_1), so (19) follows by integration. Let two isomorphs C_1^*, C_2^* of generalized curves both correspond to $C: x = x(t)$, $[t_0 \leqq t \leqq t_1]$, then by (19) $J(C_1^*) = J(C_2^*)$ for every integrand F, hence by Definition 5, the generalized curves C_1^* and C_2^* are identical, and the lemma is proved [5, p. 519].

REMARK 5: If $C^*: [x(t), \mathfrak{M}(t; \varphi), M]$ is an isomorph of an ordinary curve $C: x = x(t)$, $t_0 \leqq t \leqq t_1$, and C^* satisfies the differential equation $f(t, x, r) = 0$, then so does C, i.e. $f(t, x(t), \dot{x}(t)) = 0$. That is, the track of C^* satisfies the differential equation in the usual sense. Henceforth we shall refer to "isomorphs of ordinary curves" as "ordinary curves."

Referring to problem II of the introduction, let x, r be $n + m$ vectors. Let K_0 be the set of generalized curves satisfying (2′), (3′), (4′). Condition (4′) means that the tracks of the generalized curves satisfying (2′), (3′) have endpoints in $\bar{E}_0$, $\bar{E}_1$ respectively. Let $[t_0^*, t_1^*]$ be the interval described in (14). We make the following assumptions:

(20) There is a compact set $E \subset [t_0^*, t_1^*] \times R^n \times R^m$ which contains the track of each GC in K_0.

(21) Let $\bar{E} = \{(t, x^1, \cdots, x^n) : (t, x^1, \cdots, x^n, x^{n+1}, \cdots, x^{n+m}) \in E\}$. Let $\Omega(t, x^1, \cdots, x^n) = \{(r^{n+1}, \cdots, r^{n+m}) : h^i(t, x^1, \cdots, x^n, r^{n+1}, \cdots, r^{n+m}) \geqq 0, i \in I\}$.
Assume that for each $(t, x^1, \cdots, x^n) \in \bar{E}$ the set $\Omega(t, x^1, \cdots, x^n)$ is compact, and that the union of the sets $\Omega(t, x^1, \cdots, x^n)$ over all $(t, x^1, \cdots, x^n) \in \bar{E}$ is compact.

(22) The functions h^i, g^i, f, f^0 are continuous in their arguments.

THEOREM 3. *If K_0 is not empty and* (20), (21), (22) *are satisfied, then there is a* GC C_0^* *in K_0 which gives the functional*

$$(1') \qquad I(C^*) = \int_{t_0}^{t_1} \mathfrak{M}(t; f(t, x^1(t), \cdots, x^n(t), r^{n+1}, \cdots, r^{n+m}))\, dt$$
$$+ f^0(t_0, x^1(t_0), \cdots, x^n(t_0), t_1, x^1(t_1), \cdots, x^n(t_1))$$

its least value with respect to all generalized curves in K_0.

Proof: By Remark 4, K_0 is closed. By (20), (21), (22), there is a compact set in R^n which contains every vector $(r^1, \cdots, r^n)$ satisfying

$$(2') \qquad r^j = g^j(t, x^1, \cdots, x^n, r^{n+1}, \cdots, r^{n+m}) \qquad j = 1, 2, \cdots, n.$$

Thus there is a compact set $E^* \subset R^{n+m}$ which contains all vectors

$(r^1, \cdots, r^n, r^{n+1}, \cdots, r^{n+m})$ satisfying (2′), (3′). By Lemma 9, E^* contains the set of vectors carried by each member of K_0, and by Lemma 8 E^* serves as a common determining set for K_0. Thus the hypotheses of Theorem 2 are fulfilled by K_0.

We can now strengthen the previous conditions in order to assure the existence of a minimizing curve which is an ordinary curve. In addition to (20), (21), (22), we assume also:

(23) for each $(t, x^1, \cdots, x^n) \in \bar{E}$, the set $\Omega(t, x^1, \cdots, x^n)$ is convex.

(24) The functions g^j of (2′) have the form

$$g^j(t, x^1, \cdots, x^n, r^{n+1}, \cdots, r^{n+m}) \equiv g_0{}^j(t, x^1, \cdots, x^n) + \sum_{i=1}^{n} g_j{}^i(t, x^1, \cdots, x^n) r^{n+i}, \quad j = 1, 2, \cdots, n.$$

(The control appears linearly.)

(25) for each $(t, x^1, \cdots, x^n) \in \bar{E}$ the function $f(t, x^1, \cdots, x^n, r^{n+1}, \cdots, r^{n+m})$ is convex on $\Omega(t, x^1, \cdots, x^n)$.

(26) there exists a minimizing sequence of ordinary curves in K_0.

THEOREM 4. *If* (20)–(26) *are satisfied, then there exists an ordinary curve in* K_0 *which gives the functional* (1′) *its least value with respect to the class* K_0.

Proof: Let m be the g.l.b. of (1′) with respect to K_0. Notice that m is finite. Let $\{C_n\}$ be a minimizing sequence of ordinary curves. By Theorems 1 and 3 there is a subsequence (retain the notation) converging to a GC: $C_n \to C_0^*:[x_0(t), \mathfrak{M}_0(t; \varphi), M]$ with the property that $C_0^* \in K_0$ and $I(C_0^*) = m$. We show first that the track of C_0^* is in K_0.

By (5), (6), (11), (24), and Lemmas 8, 9, 10, we see that for almost all t in M,

$$0 = \mathfrak{M}_0\left[(t: r^j - g_0{}^j(t, x_0{}^1(t), \cdots, x_0(t)) - \sum_{i=1}^{n} g_i{}^j(t, x_0{}^1(t), \cdots, x_0{}^n(t)) r^{n+i})\right] = \dot{x}_0{}^j(t) - g_0{}^j(t, x_0{}^1(t), \cdots, x_0{}^n(t)) - \sum_{i=1}^{n} g_i{}^j(t, x_0{}^1(t), \cdots, x_0{}^n(t)) \dot{x}_0{}^{n+i}(t), \quad j = 1, 2, \cdots, n.$$

Thus, the derivative of the track of C_0^* satisfies (2′). For each t in M, the set of vectors $(r^{n+1}, \cdots, r^{n+m})$ satisfying (3′) lies in the convex set $\Omega(t, x_0{}^1(t), \cdots, x_0{}^n(t))$. By Jensen's inequality [8] the vector

$$(\mathfrak{M}_0(t; r^{n+1}), \cdots, \mathfrak{M}_0(t; r^{n+m}))$$

also lies in $\Omega(t, x_0{}^1(t), \cdots, x_0{}^n(t))$.

Thus by (11), the vector $(\dot{x}_0^{n+1}(t), \cdots, \dot{x}_0^{n+m}(t))$ lies in $\Omega(t; x_0^1(t), \cdots, x_0^n(t))$ for almost all t. Hence the derivative of the track of C_0^* satisfies (2′), (3′), and so the track of C_0^* is in K_0. By Jensen's inequality [8] and (25) we have

$$\begin{aligned}(27)\qquad &\mathfrak{M}_0(t; f(t; x_0^1(t), \cdots, x_0^n(t), r^{n+1}, \cdots, r^{n+m}) \\ &\quad \geqq f(t; x_0^1(t), \cdots, x_0^n(t), \dot{x}_0^{n+1}(t), \cdots, \dot{x}_0^{n+m}(t)) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{for almost all } t.\end{aligned}$$

But the integral of the left side of (27) is equal to m which is the g.l.b. of $I(C^*)$ on K_0, hence the integral of the right side of (27) is equal to m, and the proof is completed [cf. 5, p. 532].

REMARK 6. Theorem 4 appears in only a slightly less general form in [4, pp. 38–42]. It is interesting to note that the method of proof given here yields the additional fact; in the language of the control problem the minimizing sequence of Theorem 4 has the form

$$\{C_n\}: \left\{x_n^1(t), \cdots, x_n^n(t), \int_{t_{on}}^t u_n^1(t), \cdots, \int_{t_{on}}^t u_n^m(t)\right\}$$

and a subsequence of $\{C_n\}$ (retain the notation) converges uniformly to the track of C_0^*. But as we have seen, the track of C_0^* yields the minimizing curve for the problem, hence the "optimal control" $u = (u^1, \cdots, u^m)$ is given by

$$u^k(t) = \frac{d}{dt} \lim_{n\to\infty} \int_{t_{on}}^t u_n^k(t)\, dt \qquad k = 1, 2, \cdots, m.$$

We shall see shortly that in some instances where conditions are imposed which assure the existence of an optimal control, that a minimizing sequence does not necessarily yield up the optimal control in the above sense.

For the next theorem [11, 13], we drop condition (23) and replace (24) by (24′) Referring to (2′), for each $(t, x^1, \cdots, x^n) \in \bar{E}$, the set of vectors

$$\begin{aligned}S = \{&(g^1(t, x^1, \cdots, x^n, r^{n+1}, \cdots, r^{n+m}), \cdots, \\ &g^n(t, x^1, \cdots, x^n, r^{n+1}, \cdots, r^{n+m})) : (r^{n+1}, \cdots, r^{n+m}) \\ &\qquad\qquad \in \Omega(t, x^1, \cdots, x^n)\}\end{aligned}$$

is a convex set in R^n.

THEOREM 5. *If* (20), (21), (22), (24′), (26), *are satisfied and if* $f \equiv 0$, *then there exists an ordinary curve in* K_0 *which gives the functional* (1′) *its least value with respect to* K_0.

Proof: Let m, $\{C_n\}$, C_0^* be as in Theorem 4.

As in Theorem 4 we have for almost all t in M and all i, $i = 1, 2, \cdots, n$.

$$0 = \mathfrak{M}_0(t; r^i - g^i(t, x_0^1(t), \cdots, x_0^n(t), r^{n+1}, \cdots, r^{n+m}))$$
$$= \dot{x}_0^i(t) - \mathfrak{M}_0(t; g^i(t, x_0^1(t), \cdots, x_0^n(t), r^{n+1}, \cdots, r^{n+m}).$$

By (24′) and Jensen's inequality [8], the vector

$$V = [\mathfrak{M}_0(t; g^1(t, x_0^1(t), \cdots, x_0^n(t), r^{n+1}, \cdots, r^{n+m}), \cdots,$$
$$\mathfrak{M}_0(t; g^n(t, x_0^1(t), \cdots, x_0^n(t), r^{n+1}, \cdots, r^{n+m})]$$

is in S, so there exists a vector (at least one for each t) $(r_t^{n+1}, \cdots, r_t^{n+m}) \in \Omega(t, x_0^1(t), \cdots, x_0^n(t))$ such that

$$V = [g^1(t, x_0^1(t), \cdots, x_0^n(t), r_t^{n+1}, \cdots, r_t^{n+m}), \cdots,$$
$$g^n(t, x_0^1(t), \cdots, x_0^n(t), r_t^{n+1}, \cdots, r_t^{n+m})]. \tag{28}$$

We cannot conclude here that $r_t^{n+j} = \dot{x}_0^{n+j}(t)$, however using a lemma of Fillipov [13, p. 78], we can construct a measurable vector valued function $u(t) = (u^1(t), \cdots, u^m(t))$ such that $u(t) = (r_t^{n+1}, \cdots, r_t^{n+m})$ for some vector $(r_t^{n+1}, \cdots, r_t^{n+m})$ satisfying (28). Since the components $x_0^{n+j}(t)$, $j = 1, 2, \cdots, m$ of the track of C_0^* do not appear explicitly in the problem, we may replace these by $\int_{t_0}^t u^j(t)\, dt$ without changing the value of f^0 along C_0^*. Thus, the ordinary curve $\bar{C} = (x_0^1(t), \cdots, x_0^n(t), \int_{t_0}^t u^1(t), \cdots, \int_{t_0}^t u^m(t))$ is in K_0, and is minimizing, and the proof is completed.

REMARK 7. If the integrand function $f \not\equiv 0$ but independent of u, then Theorem 5 holds, for the integral along C_0^* is unchanged by the construction of u. If $f \not\equiv 0$ and depends explicitly on u then the problem may be reduced to one in which $f \equiv 0$ by the familiar device of adding another differential equation to (2′) and another term (depending only on the endpoints) to f^0.

REMARK 8. Except for minor changes, Theorem 5 already appears in [11, 13]. The assumption (20) implies that the solutions of (2′) (in the ordinary sense) which satisfy (3′) (4′) are uniformly bounded. Further explicit assumptions on the form of the functions g^i, h^j could be made to assure that (20) holds [4, 11, 13].

Example 1. The construction of $u(t)$ in Theorem 5 indicates that the minimizing sequence $\{C_n\}$ may contain no subsequence which converges to the minimizing curve in the sense of Remark 6. This is indeed the case; let

$$\begin{cases} \dot{x}_1 = p_1u_1 + p_2u_2 \\ \dot{x}_2 = p_1u_3 + p_2u_4 \\ \dot{x}_3 = -1 \end{cases} \qquad \begin{cases} u_1^2 + u_3^2 = 1 \\ u_2^2 + u_4^2 = 1 \\ p_1 + p_2 = 1 \\ 0 \leqq p_1 \leqq 1, \qquad 0 \leqq p_2 \leqq 1 \end{cases}$$

$$I = \int_0^1 (x_1^2 + x_2^2)\, dt.$$

E_0 and E_1 are the points (0, 0, 0, 1), (1, 0, 0, 0) respectively.

Here $n = 3$, $m = 6$, the state vector is $x(t) = (x_1(t), x_2(t), x_3(t))$ and the control vector is $u(t) = (u_1(t), u_2(t), u_3(t), u_4(t), p_1(t), p_2(t))$. (Subscripts are used here to denote components of a vector.) Each of the equality constraints on the controls could obviously be replaced by two inequality constraints of the form (3′). Conditions (20), (21), (22), (24′) are satisfied, with $\Omega(t, x_1, x_2, x_3)$ independent of (t, x_1, x_2, x_3) and S is the convex set in R^3: $S = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 \leqq 1, x_3 = -1\}$. For this problem, $m = 0$, and the g.l.b. is attained by taking, for example, $u_1(t) \equiv u_3(t) = 1/\sqrt{2}$, $u_2(t) \equiv u_4(t) = -1/\sqrt{2}$, $p_1(t) \equiv p_2(t) = 1/2$. Consider the sequence of controls

$$\begin{cases} u_{1k}(t) \equiv u_{2k}(t) = \cos 2\pi kt \\ u_{3k}(t) \equiv u_{4k}(t) = \sin 2\pi kt \\ p_1(t),\ p_2(t) \text{ arbitrary.} \end{cases}$$

This sequence of controls yields a minimizing sequence for the problem, with

$$x_{1k}(t) = \frac{\sin 2\pi kt}{2\pi k}, \qquad x_{2k}(t) = \frac{1 - \cos 2\pi kt}{2\pi k}, \qquad x_{3k}(t) = 1 - t$$

$$x_{4k}(t) = x_{5k}(t) = \int_0^t u_{1k}(t) = x_{1k}(t), \quad x_{6k}(t) = x_{7k}(t) = \int_0^t u_{3k}(t) = x_{2k}(t),$$

$$I(C_k) = \frac{1}{2\pi^2 k^2}.$$

If we take the limit along any subsequence, there is obtained

$$\lim_{k\to\infty} x_{jk}(t) = 0 \qquad j = 1, 2, 4, 5, 6, 7,$$

and

$$\lim_{k\to\infty} I(C_k) = 0.$$

Thus, one cannot take the optimal control in the sense of Remark 6, for that control would not satisfy the constraints $u_1^2 + u_3^2 = u_2^2 + u_4^2 = 1$ [4, p. 43], [2].

REMARK 9. If the hypothesis of Theorem 3 are satisfied, and if there is no ordinary curve in K_0 which minimizes (1′), then the curve in $(n + 1)$ space defined by $x_0^1(t), \cdots, x_0^n(t), t_0 \leqq t \leqq t_1$, where the $x_0^j(t)$ are components of the track of C_0^* of Theorem 3 has been called an "optimal sliding state" by R. V. Gamkrelidze [2].

REFERENCES

[1] T. Botts, *Sufficient conditions for a generalized curve problem in the calculus of variations*, Duke Math. J., 11 (1944) pp. 373–403.

[2] R. V. Gamkrelidze, *Optimal sliding states*. Dok. Akad. Nauk. SSSR, 143 (6) (1962) pp. 1243–1245.

[3] J. P. LaSalle, *The time optimal control problem*. Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton Univ. Press (1960) pp. 1–24.

[4] E. B. Lee, and L. Markus, *Optimal Control for Nonlinear Processes*, Arch. Rational Mech. Anal., 8 (1961) pp. 36–58.

[5] E. J. McShane, *Generalized curves*, Duke Math. J. 6. (1940) pp. 513–536.

[6] E. J. McShane, *Necessary conditions in generalized curve problems in the calculus of variations*, Duke Math. J. 7, (1940) pp. 1–27.

[7] E. J. McShane, *Existence theorems for bolza problems in the calculus of variations*. Duke Math. J., 7, (1940) pp. 28–61.

[8] E. J. McShane, *Jensen's inequality*. Bull. Amer. Math. Soc. 43, (8) (1937) pp. 521–527.

[9] E. J. McShane, *A metric in the space of generalized curves*. Ann. of Math. 52 (2) (1950), pp. 328–349.

[10] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience (1962) Chap. III.

[11] E. Roxin, *The existence of optimal controls*, Michigan Math. J. 9 (1962) pp. 109–119.

[12] L. C. Young *Generalized curves and the Existence or an Attained Absolute Minimum in the Calculus of Variations*. Comptes Rendus de la Société des Sciences et des Lettres de Varsovie, Classe III, vol. 30 (1937), pp. 212–234.

[13] A. E. Filippov, *On certain questions in the theory of optimal control*. J. Soc. Indust. Appl. Math., Ser. A. 1, (1) (1962) pp. 76–84.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

ON THE EXISTENCE OF OPTIMAL FEEDBACK CONTROLS*

T. F. BRIDGLAND, JR.†

Introduction. Of the considerable number of papers dealing with problems of optimal control which has appeared in the recent past, most contain little more than passing mention of what has been called "the synthesis problem", i.e., the problem of expressing the optimal control as a function of the state of the system being controlled. LaSalle [1] comments briefly on the synthesis problem in connection with time-optimal control and Harvey [2] applies LaSalle's ideas. A few other studies of the synthesis problem in special cases exist, but the most detailed study of this problem to date is that contained in Berkovitz's comprehensive treatment [3] of the general problem of optimal control. However, in [3], as in other studies of optimal control, the synthesis problem is subordinated to the more general "open-loop" problem.

In a recent report [4], Kalman has resurrected a little known technique due to Carathéodory [5, pp. 198–201] which was originally devised for the purpose of obtaining sufficient conditions for the existence of extremal arcs in the simple problem of the calculus of variations. Formally stating this technique as a lemma, Kalman points out its applicability to the synthesis problem. The results of [4], in the main, constitute a straightforward extension of those of [5], the major point of departure being the introduction of differential side constraints of the form usually associated with problems of optimal control.

The major virtue of the Carathéodory technique is that it permits a direct approach to the solution of the synthesis problem; its major shortcoming, as manifested by the presentations in [4], [5], is the excessively onerous requirement of differentiability for the admissible controls and the functionals to be minimized. In this paper, we show that the Carathéodory technique is valid under much weaker restrictions than those assumed in [4] and [5], thereby largely eliminating the shortcoming of the technique while leaving its virtue untouched. Our principal results consist of necessary and sufficient conditions for the existence of a unique feedback control which extremizes a given criterion functional defined on the space of trajectories of a control system of general type.

Formulation of the optimal control problem. For each $t_0 \geqq 0$, we shall denote by $I(t_0)$ the half-line $\{t \mid t_0 \leqq t < \infty\}$ and by $\bar{I}(t_0)$, the one-point compactification of $I(t_0)$ obtained by adjoining the point at infinity.

* Received by the editors January 22, 1963 and in revised form May 12, 1963.
† The Martin Company, Denver, Colorado and University of South Carolina.

When $t_0 = 0$ we shall write briefly I, $\bar{I}$. R^k will be used to denote real Euclidean k-space.

Consider the differential equation

$$\dot{y} = g(t, y) \tag{1}$$

where y, g are vectors in R^n. An equation (1) will be said to be of class A if it possesses the following properties:

(A). for each $(t_0, y_0) \in I \times R^n$, (1) possesses a solution (in the sense of Carathéodory), $y(t; t_0, y_0)$, satisfying $y(t_0; t_0, y_0) = y_0$, which is unique, continuous in the pair (t_0, y_0), continuable to all of I and such that, for fixed (t, t_0), it satisfies a Lipschitz condition with respect to y_0 in each bounded region of R^n.

All of the differential equations considered in this paper will be assumed to be of class A. The results on optimal controls to be derived in the sequel may be established under less stringent conditions than those of (A) but to pursue these generalizations in detail would tend to obscure the central results of the paper.

We assume the existence of a *control set*, U, which comprises the totality of functions, $u(t)$, on I to R^m, $m \leqq n$, having values in a given bounded subset, Φ, of R^m and having components $u_i(t)$, $i = 1, \cdots, m$, which are measurable. The control system with which we shall be concerned is assumed to be described mathematically by an n-vector differential equation

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

which, for each $u \in U$, is of class A. We shall need to consider a (feedback) system related to (2),

$$\dot{x} = f(t, x, \varphi_0(t, x)) \tag{3}$$

which is also of class A. We shall consistently use the notation $x(t; t_0, x_0, u)$, $\bar{x}(t; t_0, x_0)$ to denote respectively the solutions of (2), (3) passing through the point (t_0, x_0).

In order to state the problem of optimal control with which we shall be concerned, we must introduce the following notation. Denote by C the set of all absolutely continuous functions on I to R^n and by S, the set of all bounded, measurable functions, on I to R^m, for which (2) is of class A. Evidently $U \subset S$. We assume the existence of functions $R(t; x, v)$ on $\bar{I} \times C \times S$ to R^1, and $t^*(x, v)$ on $C \times S$ to $\bar{I}$. We may then define a function $Q(x, v)$ on $C \times S$ to R^1 by

$$Q(x, v) = R(t^*(x, v); x, v).$$

For given $(t_0, x_0) \in I \times R^n$, the differential equation (2), together with the initial condition $x(t_0; t_0, x_0, u) = x_0$, determines a mapping of S into

C. Denoting by x_v the image, under this map, of $v \in S$, *the problem of optimal control is, given* $(t_0, x_0) \in I \times R^n$, *to determine the set of controls* $U^* \subset U$ *for which* $u_0 \in U^*$ *implies that*

$$Q(x_{u_0}, u_0) = \min_{u \in U} Q(x_u, u) \tag{4}$$

when this minimum exists. Such a u_0 will be called an *optimal control.*

In view of condition (A), the functionals Q, $t^\#$ depend only on t_0, x_0, v, while the functional R depends only on t, t_0, x_0, v, so that, henceforth, we shall write these functionals in the following forms, respectively: $Q(t_0, x_0, v)$, $t^\#(t_0, x_0, v)$, $R(t; t_0, x_0, v)$. In the sequel, Q will be styled a *criterion*; R, a *generator*, and $t^\#$, the *final time.* In addition, the range of $t^\#$ will be restricted to $\bar{I}(t_0)$.

It is important to recognize that, in most control problems, the differential equation (2) and the control set U are the fundamental entities which are known and that the subsequent selection of a generator manifests a desire to determine a control, u_0, which is optimal relative to the *entire* control set U and not just a subset thereof. This brings us to a discussion of the final time, $t^\#$.

Trivial examples of $t^\#$ are the following:

$$\begin{aligned} t^\#(t_0, x_0, u) &= T, \qquad t_0 < T \\ &= t_0, \qquad T \leqq t_0; \end{aligned} \tag{5}$$

and

$$t^\#(t_0, x_0, u) = t_0 + T, \tag{6}$$

where in both cases, T is a positive constant. Both these examples provide instances of optimization (i.e., determination of an optimal control) on the basis only of the behavior of the solutions of (2). Suppose, however, that one desires the solutions of (2) to lie in a set G (which may depend on t) at some point $t_1 > t_0$ and that one wishes to optimize relative to a criterion that depends on the value t_1. A natural impulse is to set $t^\#(t_0, x_0, u)$ equal to the least value t_1 for which $x(t_1; t_0, x_0, u) \in G$, but it is easy to construct examples in which such a t_1 does not exist for *any* $u \in U$, and others in which it exists for some u but not others. One is thus confronted with the problem of defining a $t^\#(t_0, x_0, v)$, on all of $I \times R^n \times S$, which depends meaningfully and nontrivially on the target set $G(t)$.

We next demonstrate the existence of a $t^\#$ which satisfies the requirements of our statement of the optimal control problem and which at the same time permits the formulation of a particular problem of optimal control which has been of engineering interest for several years but has

not yet received adequate mathematical treatment. The problem referred to is called *the minimum miss distance problem*, and in this problem, as we shall point out, it is possible to imbed several control problems of current interest.

To start with, let Ω^n denote the class of all nonempty compact subsets of R^n. If $\rho(a, b)$ denotes the distance between $a, b \in R^n$, where ρ is the metric induced by the Euclidean norm (denoted $\|\cdot\|$) in R^n, then the distance, $\partial(a, B)$, between a point $a \in R^n$ and a set $B \in R^n$, is defined by

$$\partial(a, B) = \min_{b \in B} \rho(a, b).$$

An s-neighborhood of a set $G \in \Omega^n$ may then be defined by

$$N_s(G) = \{x \in R^n \mid \partial(x, G) < s\};$$

as a consequence, Ω^n may be metrized by defining the (Hausdorff) distance, d, between $G, H \in \Omega^n$ as

$$d(G, H) = \inf \{s \mid G \subset N_s(H) \quad \text{and} \quad H \subset N_s(G)\}.$$

The set $\{\Omega^n; d\}$ is a metric space in which R^n is isometrically imbedded. A continuous function, $G(t)$, on I to $\{\Omega^n; d\}$ will be called a *target set* (cf. [6]).

LEMMA 1. *$\partial(a, B)$ is continuous in $R^n \times \Omega^n$.*

Proof. Let $q, r \in R^n$ and $G, H \in \Omega^n$ be such that $\rho(q, r) < \epsilon/2$ and $d(G, H) < \epsilon/2$; choose $\gamma \in G$ such that $\rho(q, \gamma) = \partial(q, G)$. Since there exists $\eta \in H$ such that $\rho(\gamma, \eta) \leqq d(G, H)$, it follows by the triangle law that

$$\partial(r, H) \leqq \rho(r, \eta) \leqq \rho(r, q) + \rho(q, \gamma) + \rho(\gamma, \eta) < \epsilon + \partial(q, G).$$

The conclusion follows by symmetry from this inequality.

By virtue of Lemma 1, the function $p(t; t_0, x_0, v)$, which is identical to $\partial(x(t; t_0, x_0, v), G(t))$ on $I(t_0)$ and whose value at $t = \infty$ is defined to be $\liminf_{t \to \infty} \partial(x(t; t_0, x_0, v), G(t))$, is lower semicontinuous on $\bar{I}(t_0)$. Consequently, it is meaningful to define a *miss distance*, $\delta(t_0, x_0, v)$, as

$$\delta(t_0, x_0, v) = \min_{t \in \bar{I}(t_0)} p(t; t_0, x_0, v).$$

Thus it is clear that the set

$$T = \{t \in \bar{I}(t_0) \mid p(t; t_0, x_0, v) = \delta(t_0, x_0, v)\}$$

is nonempty and bounded below (by $t = t_0$) so that it has an infimum, $t^* = t^*(t_0, x_0, v)$; it is a trivial matter to show that $t^* \in T$. (In fact, T is closed.) Since t^* is defined throughout $I \times R^n \times S$, it satisfies our requirements as a t^* candidate.

The minimum miss distance problem cited previously is now formulated by putting $t^* \equiv t^*$ and taking $R(t; t_0, x_0, v) \equiv p(t; t_0, x_0, v)$. It is interesting to observe that the set $U^{**} \subset U^*$, of optimal controls for which $p(t^*; t_0, x_0, u) = 0$, is coextensive with the set of controls u for which the corresponding solution of (2) lies in $G(t_1)$ at some time t_1. Consequently, if U^{**} contains an element u^{**} for which

$$t^*(t_0, x_0, u^{**}) = \min_{u \in U^{**}} t^*(t_0, x_0, u),$$

then this control u^{**} is evidently a solution to the *time optimal control problem* [1].

Some fundamental lemmas. In this section we state several fundamental results which will be essential in the establishment of our principal theorems. In the next lemma, which generalizes [7, Lemma 2.4], μ_0, μ_1, μ_2 will denote, respectively, Lebesgue measures in I, R^n and $I \times R^n$; μ_2 is the completion of the product measure $\mu_0 \times \mu_1$ in $I \times R^n$[8]. Before stating this lemma, we establish the following statement which will play a key role in proving it:

(α). *if $g(x)$ is a function from R^n onto itself which, in each bounded region of R^n, satisfies a Lipschitz condition, then a set $E \subset R^n$ satisfies $\mu_1(E) = 0$ only if $\mu_1[g(E)] = 0$.*

Suppose first that E is bounded. Since $\mu_1(E) = 0$, for every $\epsilon > 0$ there exists a sequence, $\{I_m^\epsilon\}$, of open n-cells such that E is contained in their union and $\sum_{m=1}^{\infty} \mu_1(I_m^\epsilon) < (2k\sqrt{n})^{-n}\epsilon$, where k is the Lipschitz constant [9, p. 107]. Fixing ϵ for the moment, we recall [9, p. 18] that for each m, I_m^ϵ is the union of a countable collection, $\{C_{mk}^\epsilon\}$, of disjoint, half-open n-cubes, each of diameter d_{mk}^ϵ; I_m^ϵ is also the union of the closures, $\bar{C}_{mk}^\epsilon$, of these n-cubes. Clearly, $\mu_1(I_m^\epsilon) = \sum_{j=1}^{\infty} \mu_1(C_{mj}^\epsilon)$, where $\mu_1(C_{mj}^\epsilon) = (d_{mj}^\epsilon/\sqrt{n})^n$.

Consider now the set $J_{mj}^\epsilon = g(\bar{C}_{mj}^\epsilon)$; J_{mj}^ϵ has a finite diameter ∂_{mj}^ϵ which, by virtue of the Lipschitz condition, satisfies $\partial_{mj}^\epsilon \leqq k d_{mj}^\epsilon$. Moreover, there is an open hypercube, K_{mj}^ϵ, with edge length $2\partial_{mj}^\epsilon$, such that $J_{mj}^\epsilon \subset K_{mj}^\epsilon$. We derive immediately

$$\mu_1(K_{mj}^\epsilon) = (2\partial_{mj}^\epsilon)^n \leqq (2kd_{mj}^\epsilon)^n = (2k\sqrt{n})^n \mu_1(C_{mj}^\epsilon)$$

and there follows

$$\sum_{m,j=1}^{\infty} \mu_1(K_{mj}^\epsilon) \leqq (2k\sqrt{n})^n \sum_{m=1}^{\infty} \mu_1(I_m^\epsilon) < \epsilon.$$

Since $g(E)$ is contained in the union of the K_{mj}^ϵ and ϵ is arbitrary, we conclude $\mu_1[g(E)] = 0$. If E is unbounded, it may be expressed as the union of a countable number of bounded subsets and the preceding result applied to each of these subsets. The assertion (α) is thus established.

REMARK. Although it appears that (α) should be well known, the most nearly similar assertion the author has found is that of Apostol [10, Thm. 10-8]. H. G. Hermes has pointed out to the author that, under slight additional restriction of g, a proof of (α) may be based on Lemmas 6.1, 6.2 of [11].

LEMMA 2. *Let $y(t; t_0, y_0)$ denote a solution of* (1) *with t_0 fixed; then a measurable set $M \subset I(t_0) \times R^n$ has (μ_2) measure zero if and only if, for almost all $\eta \in R^n$, the set*

$$H_M[\eta] = \{t \in I(t_0) | (t, y(t; t_0, \eta)) \in M\}$$

has (μ_0) measure zero.

Proof. We follow closely the proof of [7, Lemma 2.4]. Assuming $t_0 = 0$, without loss of generality, it follows from the earlier assumption that (1) is of class A, that

$$\Psi(t, \eta) = (t, y(t; 0, \eta))$$

is a homeomorphism of $I \times R^n$ onto itself, with inverse given by

$$\Psi^{-1}(t, \eta) = (t, y(0; t, \eta)).$$

Suppose first that $\mu_2(M) = 0$; then there is a Borel set $\tilde{M}$, of measure zero, containing M. Since Ψ is a homeomorphism, $\Psi^{-1}(\tilde{M})$ is a Borel set and, by [8, p. 39], the sets $\tilde{M}_t$, $(\Psi^{-1}(\tilde{M}))_t$, defined by

$$\tilde{M}_t = \{\eta \in R^n | (t, \eta) \in \tilde{M}\},$$

$$(\Psi^{-1}(\tilde{M}))_t = \{\eta \in R^n | (t, \eta) \in \Psi^{-1}(\tilde{M})\},$$

are (μ_1) measurable for almost all t and $\mu_1(\tilde{M}_t) = 0$ for almost all t. From these last equations it is easy to deduce that

$$(\Psi^{-1}(\tilde{M}))_t = \{\eta \in R^n \mid \eta = y(0; t, \eta_0);\ \eta_0 \in \tilde{M}_t\}. \tag{7}$$

For fixed $t \in I$, $\eta = y(0; t, \eta_0)$ is a homeomorphism of R^n onto itself with inverse given by $\eta_0 = y(t; 0, \eta)$. By virtue of condition (A), both the direct and inverse maps satisfy a Lipschitz condition in each bounded region of R^n; it is thus a consequence of (α) and (7) that $\mu_1(\tilde{M}_t) = 0$ if and only if $\mu_1((\Psi^{-1}(\tilde{M}))_t) = 0$. The fact that $\mu_1((\Psi^{-1}(\tilde{M}))_t) = 0$ for almost all $t \in I$ implies [8, p. 40] that $\mu_2(\Psi^{-1}(\tilde{M})) = 0$. This in turn implies $\mu_0((\Psi^{-1}(\tilde{M}))_\eta) = 0$ for almost all $\eta \in R^n$ and, since

$$(\Psi^{-1}(\tilde{M}))_\eta \supset (\Psi^{-1}(M))_\eta = \{t | (t, \eta) \in \Psi^{-1}(M)\}$$

$$= \{t \mid \Psi(t, \eta) \in M\} = H_M[\eta],$$

there results $\mu_0(H_M[\eta]) = 0$ for almost all $\eta \in R^n$. The converse is obtained, *mutatis mutandis*, by reversing the steps of this proof.

A function $V(t, x)$ on $I \times R^n$ to R^1 is said to be *locally lipschitzian* if to each $t^1 \in I$ and every $r > 0$ there correspond $\delta = \delta(t^1, r) > 0$, $K = K(t^1, r) > 0$ such that $| t - t^1 | < \delta$, $\| x_i \| \leqq r$, $i = 1, 2$, imply $| V(t, x_1) - V(t, x_2) | \leqq K \| x_1 - x_2 \|$. For a continuous, locally lipschitzian function $V(t, x)$, the *Yoshizawa* $(Y-)$ *derivate by virtue of* (2) is defined as

$$V^+(t, x; u) = \limsup_{h \to 0+} h^{-1}\{V(t + h, x + hf(t, x, u(t))) - V(t, x)\}.$$

Denoting by $D^+m(t)$ the upper right hand derivative of m with respect to t, we have almost trivially

LEMMA 3. *For almost all* $t \in I$,

$$D^+V(t, x(t; t_0, x_0, u)) = V^+(t, x(t; t_0, x_0, u); u).$$

A continuous, locally lipschitzian function $V(t, x)$ is said to be (locally) *absolutely continuous uniformly* (*acu*) at a point (t', x') if there exists $\rho = \rho(t', x') > 0$ and, for $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that on any finite set of disjoint intervals $(t_n, t_n + \delta_n)$ satisfying

$$\delta_n > 0, \qquad \sum_n \delta_n < \delta, \qquad t' - \rho \leqq t_n \leqq t' + \rho,$$

$$t' - \rho \leqq t_n + \delta_n \leqq t' + \rho$$

we have

$$\sum_n | V(t_n + \delta_n, x_n) - V(t_n, x_n) | < \epsilon$$

for every sequence $\{x_n\}$ satisfying $\| x_n - x' \| \leqq \rho$.

LEMMA 4. [12] *If* $V(t, x)$ *is acu, then* $V(t, x(t))$ *is an absolutely continuous function of* $t \in I$, *for every absolutely continuous* $x(t)$.

In the sequel, a continuous, locally lipschitzian, acu function $V(t, x)$ on $I \times R^n$ to R^1 will be called a *gauge function*.

The final concept which we wish to introduce in this section is that of determinacy. Let the function $L(t, x, \varphi)$ on $I \times R^n \times R^m$ to R^1 be such that, for every bounded, measurable $u(t)$ and every absolutely continuous $x(t)$, $L(t, x(t), u(t))$ is integrable in the sense of Lebesgue on every bounded, measurable subset of I. $L(t, x, \varphi)$ will be said to be *determinate* if it satisfies all the following conditions:

(i). there exists a function $\varphi_0(t, x)$ on $I \times R^n$ to R^m for which $\varphi_0(t, x(t))$ is bounded and measurable for every continuous $x(t)$ and such that $L(t, x, \varphi_0(t, x)) = 0$ almost everywhere in $I \times R^n$;

(ii). almost everywhere in $I \times R^n$, the conditions $u \in U$ and $u(t) \neq \varphi_0(t, x)$ together imply $L(t, x, u(t)) > 0$;

(iii). with this $\varphi_0(t, x)$, (3) is of class A.

Instead of requiring that (i), (ii) hold almost everywhere in $I \times R^n$, we may require that they hold almost everywhere in some subset $D \subset I \times R^n$, in which case we need require only that (3) satisfy condition (A) in the set D. In this event, we shall speak of "determinacy in D".

Let us now define the set B as

$$B = \{(t, x) \in I \times R^n \mid t^*(t, x, u) > t \quad \text{for some} \quad u \in U\}; \tag{8a}$$

with each $(t, x) \in B$ there is associated a set $U(t, x) \subset U$ for each member of which the inequality in the definition of B is satisfied. Then, corresponding to a given determinate function $L(t, x, \varphi)$, we may define the set $\tilde{B} \subset B$ as

$$\tilde{B} = \{(t_0, x_0) \in B \mid \varphi_0(t, \bar{x}(t; t_0, x_0)) \in U(t_0, x_0)\}. \tag{8b}$$

It is a consequence of the definition of t^* that, for fixed (t_0, x_0, u), $t^*(t, x(t; t_0, x_0, u), u) = t^*(t_0, x_0, u)$ for all $t \in [t_0, t^*(t_0, x_0, u))$; throughout the remainder of the paper, we shall assume that this property holds for arbitrary functions t^*.

Optimal feedback controls—sufficient conditions. Our first theorem is a formal statement of the generalization of the Carathéodory technique [5, p. 198].

THEOREM 1. *Let $L(t, x, \varphi)$ be determinate and let $R(t; t_0, x_0, v)$ be a generator for which the following conditions are satisfied:*

a) *for all $(t_0, x_0, v) \in I \times R^n \times S$, $R(t_0; t_0, x_0, v) = 0$;*

b) *for each fixed (t_0, x_0, v), $R(t; t_0, x_0, v)$ is an absolutely continuous function of t with derivative equal almost everywhere on $I(t_0)$ to $L(t, x(t; t_0, x_0, v), v(t))$.*

Then it follows that

c) *for each $t_0 \in I$ and almost all $x_0 \in R^n$, $R(t; t_0, x_0, \bar{v}) = 0$ on $I(t_0)$, where $\bar{v}(t) = \varphi_0(t, \bar{x}(t; t_0, x_0))$;*

d) *for each $(t_0, u) \in I \times U$ and almost all $x_0 \in R^n$, if*

$$u(t) \neq \varphi_0(t, x(t; t_0, x_0, u)) \tag{9}$$

on a set of positive measure contained in an interval $[t_1, t_2]$, then $R(t_2; t_0, x_0, u) > R(t_1; t_0, x_0, u)$. Hence, for each $t_0 \in I$ and almost all $x_0 \in R^n$ for which $(t_0, x_0) \in \tilde{B}$, the control

$$u_0(t; t_0, x_0) \equiv \varphi_0(t, \bar{x}(t; t_0, x_0)) \tag{10}$$

is the unique optimal control in $U(t_0, x_0)$ relative to the criterion (4).

Proof. Let M denote the set in $I \times R^n$ for which $L(t, x, \varphi_0(t, x)) \neq 0$ and denote by $H[x_0]$ the set $\{t \in I(t_0) | (t, \bar{x}(t; t_0, x_0)) \in M\}$. By determinacy condition (i), $\mu_2(M) = 0$ and, by determinacy condition (iii)

and Lemma 2, $\mu_0(H[x_0]) = 0$ for almost all $x_0 \in R^n$. Similarly, denoting by N the set in $I \times R^n$ for which $u(t) \neq \varphi_0(t, x)$ implies $L(t, x, u(t)) \leqq 0$ and, by $J_u[x_0]$, the set $\{t \in I(t_0) | (t, x(t; t_0, x_0, u)) \in N\}$, it is a consequence of Lemma 2, determinacy condition (ii) and the property (A) of (2) that $\mu_2(N) = 0$ and $\mu_0(J_u[x_0]) = 0$ for almost all $x_0 \in R^n$.

By b), we have

$$R(t_2; t_0, x_0, v) - R(t_1; t_0, x_0, v) = \int_{t_1}^{t_2} L(\tau, x(\tau; t_0, x_0, v), v(\tau))\, d\tau. \tag{11}$$

Furthermore, it is clear that the solutions $\bar{x}(t; t_0, x_0)$ of (3) and $x(t; t_0, x_0, \bar{v})$ of (2) coincide when $\bar{v}(t) = \varphi_0(t, \bar{x}(t; t_0, x_0))$. Hence, $L(t, \bar{x}(t; t_0, x_0), \varphi_0(t, \bar{x}(t; t_0, x_0))) = 0$ almost everywhere on $I(t_0)$, for all $x_0 \in R^n$ except perhaps those in a set of measure zero. From this and (11), we find by setting $t_1 = t_0$ that $R(t; t_0, x_0, \bar{v}) = 0$ on $I(t_0)$ for almost all $x_0 \in R^n$ and we have proved c). The conclusion d) follows in a similar way from (11).

We now assert that, for almost every $x_0 \in R^n$, if $u_1 \in U$ is a control for which $R(t; t_0, x_0, u_1) = 0$ on an interval $[t_0, t_1]$ then $u_1(t) = \varphi_0(t, \bar{x}(t; t_0, x_0))$ almost everywhere on that interval. To prove the assertion, we note first that u_1 must satisfy

$$u_1(t) = \varphi_0(t, x(t; t_0, x_0, u_1)) \tag{12}$$

almost everywhere on the interval, for otherwise the condition d) refutes the assumption that $R(t; t_0, x_0, u_1) = 0$. But since (12) holds, the solution of $\dot{x} = f(t, x, u_1)$ must coincide with the solution of (3); hence, in (12), $x(t; t_0, x_0, u_1)$ may be replaced by $\bar{x}(t; t_0, x_0)$ which yields the assertion. From this fact, the final statement of the theorem follows easily.

The next theorem is an immediate consequence of Lemmas 3, 4.

THEOREM 2. *If $V(t, x)$ is a gauge function with determinate Y-derivate $V^+(t, x; \varphi)$, then the generator $R(t; t_0, x_0, u)$, defined by*

$$R(t; t_0, x_0, u) = V(t, x(t; t_0, x_0, u)) - V(t_0, x_0), \tag{13}$$

satisfies the hypotheses of Theorem 1.

An even more general result is the following.

THEOREM 3. *Let $V(t, x)$ be a gauge function and let the function $L(t, x, \varphi)$ on $I \times R^n \times R^m$ to R^1 be such that $L(t, x(t), v(t))$ is integrable on every bounded, measurable subset of I for each bounded, measurable function $v(t)$ and each absolutely continuous $x(t)$. If $L^*(t, x, \varphi)$, defined by*

$$L^*(t, x, \varphi) = V^+(t, x; \varphi) + L(t, x, \varphi),$$

is determinate, if $\Pi(t; t_0, x_0, v)$ *is a generator satisfying*

f) $\Pi(t_0; t_0, x_0, v) = V(t_0, x_0)$, $v \in S$;

g) $\Pi(t; t_0, x_0, v)$ *is an absolutely continuous function of* t *with derivative equal almost everywhere to*

$$DV(t, x(t; t_0, x_0, v)) + L(t, x(t; t_0, x_0, v), v(t)) \quad \text{on} \quad I(t_0);$$

then the generator $R(t; t_0, x_0, v) = \Pi(t; t_0, x_0, v) - V(t_0, x_0)$ *satisfies the hypotheses of Theorem* 1.

Proof. It is a matter of direct verification, taking into account Lemmas 3, 4, to determine that the hypotheses of Theorem 1, with L replaced by L^*, are satisfied by R.

REMARK. If, in (2), we take $f(t, x, u(t)) \equiv u(t)$, then the preceding theorems are direct generalizations of Carathéodory's results; in particular, Theorem 3 is the extension of [5, Satz 2, p. 200]. If in Theorems 2, 3, $V(t, x)$ be taken to be continuously differentiable with respect to each of its arguments, then $V(t, x)$ is a gauge function. By following Carathéodory's line of thought, we may then require that V be a solution, in a certain general sense, of a Hamilton-Jacobi differential equation. We shall consider this point further in a later section.

Optimal feedback controls—necessary conditions. For the subsequent statement of the converses of Theorems 1, 2, 3, we shall require the following conditions.

I. The set B(*vide* (8)) has positive μ_2 measure and, for each $t_0 \in I$ and almost all $x_0 \in R^n$ such that $(t_0, x_0) \in B$, there is an optimal control $u_0 = u_0(t; t_0, x_0) \in U$, relative to the criterion $Q(t_0, x_0, u)$, which is uniquely defined on the interval $[t_0, t^*(t_0, x_0, u_0))$.

II. There is a function $\varphi_0(t, x)$ on $I \times R^n$ to R^m for which is satisfied $\varphi_0(t, x(t; t_0, x_0, u_0)) = u_0(t; t_0, x_0)$ almost everywhere on $[t_0, t^*(t_0, x_0, u_0))$ for all (t_0, x_0) for which u_0 exists and, for this φ_0, (3) is of class A.

III. Let t_0 be fixed and denote by B_{t_0} the set $\{x_0 \in R^n \mid (t_0, x_0) \in B\}$; then there exists $u^* \in U$ such that, for almost all $x_0 \in B_{t_0}$, $u^*(t) \neq u_0(t; t_0, x_0)$ on a subset of positive measure contained in $[t_0, t^*(t_0, x_0, u^*))$.

Of course, the criterion Q mentioned in I will depend on the particular generator chosen, which in turn will be governed by the theorem whose converse we are obtaining. Note that II implies that determinacy condition (iii) is satisfied in B. The function $u_0(t; t, x)$, provided it is uniquely defined almost everywhere in B, may serve as the function $\varphi_0(t, x)$, as can be readily verified (cf. [13, p. 136]). Finally, observe that for the set B_{t_0} of III, there is a set of points $t_0 \in I$, of positive μ_0 measure, for which

$\mu_1(B_{t_0}) > 0$; this is a direct consequence of the theorem of [8, p. 40], and the fact that $\mu_2(B) > 0$.

The next theorem is a partial converse of Theorem 1.

THEOREM 4. *If $Q(t_0, x_0, u)$ is the criterion corresponding to the generator of Theorem* 1, *if conditions* I, II, III *are satisfied and if $Q(t_0, x_0, u_0) = 0$ then $L(t, x, \varphi)$ is determinate in B.*

Proof. From (11) we have

$$Q(t_0, x_0, u) = \int_{t_0}^{t^*(t_0, x_0, u)} L(\tau, x(\tau; t_0, x_0, u), u(\tau))\, d\tau \tag{14}$$

and, since this is a criterion of "Markov type" [14, p. 54], the "principle of optimality" [14, p. 57] may be applied to (14) to yield, by virtue of I, II,

$$\int_t^{t^*(t_0, x_0, u_0)} L(\tau, x(\tau; t_0, x_0, u_0), u_0(\tau; t_0, x_0))\, d\tau = 0,$$

$$t \in [t_0, t^*(t_0, x_0, u_0)).$$

This in turn implies that

$$L(t, x(t; t_0, x_0, u_0), u_0(t; t_0, x_0)) = 0 \tag{15}$$

almost everywhere on $[t_0, t^*(t_0, x_0, u_0))$. Now suppose that, for points t', t'' satisfying $t_0 \leqq t' < t'' < t^*(t_0, x_0, u)$, the control $u(t) = u_0(t; t'', x'')$ on $[t'', t^*(t_0, x_0, u))$ but that $u(t) \neq u_0(t; t', x')$ on a subset, of positive measure, of $[t', t'')$, where $x' = x(t'; t_0, x_0, u)$ and $x'' = x(t''; t_0, x_0, u)$. We then have from I

$$\begin{aligned} &\int_{t'}^{t''} L(\tau, x(\tau; t_0, x_0, u), u(\tau))\, d\tau \\ &\quad = \int_{t'}^{t^*(t', x', u)} L(\tau, x(\tau; t_0, x_0, u), u(\tau))\, d\tau = Q(t', x', u) > 0. \end{aligned} \tag{16}$$

The statements (15), (16) hold for each $t_0 \in I$ and almost all $x_0 \in R^n$ for which $(t_0, x_0) \in B$.

Having established these facts, let us now suppose that determinacy condition (i) fails in B. This implies that there is a set $M \subset B$ for which $\mu_2(M) > 0$ and in which $L(t, x, \varphi_0(t, x)) \neq 0$. By virtue of Lemma 2, this in turn implies that there is a set $\mathfrak{M}$ of points $\eta \in R^n$ for which $\mu_1(\mathfrak{M}) > 0$ and such that $\eta \in \mathfrak{M}$ implies $\mu_0(H_M[\eta]) > 0$, where

$$H_M[\eta] = \{t \in I(t_0) | (t, \bar{x}(t; t_0, \eta)) \in M\}.$$

As a consequence, for $x_0 \in \mathfrak{M}$, $L(t, \bar{x}(t; t_0, x_0), \varphi_0(t, \bar{x}(t; t_0, x_0))) \neq 0$ on

a subset, of positive measure, of $[t_0, t^{\#}(t_0, x_0, u_0))$; since $\mathfrak{M}$ is of positive measure we have, by virtue of (15), a contradiction. Hence, determinacy condition (i) holds in B.

Suppose determinacy condition (ii) fails in B; then there is a set $N \subset B$ such that $\mu_2(N) > 0$ and such that $u(t) \neq \varphi_0(t, x)$ at a point $(t, x) \in N$ implies $L(t, x, u(t)) \leqq 0$. Let u be the u^* of condition III; then by Lemma 2, there exists $\mathfrak{N}_{u^*} \subset R^n$ such that $\mu_1(\mathfrak{N}_{u^*}) > 0$ and such that $\mu_0(J_N[\eta]) > 0$ when $\eta \in \mathfrak{N}_{u^*}$, where

$$J_N[\eta] = \{t \in I(t_0) | (t, x(t; t_0, \eta, u^*)) \in N\}.$$

In view of condition III, there is then a set, λ, of positive measure, contained in a subinterval $[t', t'') \subset [t_0, t^{\#}(t_0, x_0, u^*))$, on which $u^*(t) \neq \varphi_0(t, x(t; t_0, x_0, u^*))$ and on λ we then have

$$L(t, x(t; t_0, x_0, u^*), u^*(t)) \leqq 0. \tag{17}$$

On the complement of λ in $[t', t'')$, we have $u^*(t) = \varphi_0(t, x(t; t_0, x_0, u^*))$ almost everywhere. It then follows by determinacy condition (i) that (17) still holds almost everywhere on the complement of λ in $[t', t'')$, except perhaps for a set, Y, of points $x_0 \in \mathfrak{N}_{u^*}$ for which $\mu_1(Y) = 0$. Now $\mu_1(\mathfrak{N}_{u^*} - Y) > 0$ and for $x_0 \in (\mathfrak{N}_{u^*} - Y)$ we then know that (17) holds almost everywhere on $[t', t'')$. Integrating (17) over this interval yields

$$\int_{t'}^{t''} L(\tau, x(\tau; t_0, x_0, u^*), u^*(\tau))\, d\tau \leqq 0$$

which, in view of (16), yields a contradiction. Hence determinacy condition (ii) holds in B and, as has been indicated, determinacy condition (iii) holds in B so that $L(t, x, \varphi)$ is determinate in B.

A converse of Theorem 3 is obtained from Theorem 4 by replacing L by $L^* \equiv V^+(t, x; \varphi) + L(t, x, \varphi)$ in the latter theorem. From the converse of Theorem 3, a converse of Theorem 2 is then obtained by taking $L \equiv 0$ in the expression for L^*. The converses of Theorems 2 and 3 have statements which are sufficiently obvious that we omit them here. Note that if, in Theorem 1, we assume $B = \tilde{B}$ and that determinacy holds only on B, then the thus modified Theorem 1 and Theorem 4 are strict converses; a corresponding statement can be made for Theorems 2, 3 and their converses.

Final remarks. One of the most frequently encountered problems of optimal control is that in which it is required to obtain an optimal control relative to the criterion $Q(t_0, x_0, u)$ defined by

$$Q(t_0, x_0, u) = \int_{t_0}^{t^{\#}(t_0, x_0, u)} L(\tau, x(\tau; t_0, x_0, u), u(\tau))\, d\tau$$

where the functional $L(t, x(t), u(t))$ is required to be bounded and measurable for every absolutely continuous $x(t)$ and every bounded, measurable $u(t)$. If $L(t, x, \varphi)$ is determinate, then Theorem 1 is applicable. If it is not, then the problem may still admit a solution if one can find a gauge function, $V(t, x)$, for which $V^+(t, x; \varphi) + L(t, x, \varphi)$ is determinate and which satisfies the *transversality condition*

$$V(t, x) = 0, \ (t, x) \in C(t_0, x_0) \tag{18}$$

where the set $C(t_0, x_0)$ is defined by

$$C(t_0, x_0) = \{(t, x) \in I \times R^n \mid t = t^*(t_0, x_0, u), x = x(t^*; x_0, u) \quad \text{for all} \quad u \in U\}.$$

Under these conditions, it is an immediate consequence of Theorem 3 that

$$V(t_0, x_0) = \min_{u \in U} Q(t_0, x_0, u). \tag{19}$$

Conversely, if it is known that there is a gauge function for which (19) is satisfied, then it can be shown that the necessity of the transversality condition (18) and of the determinacy of $V^+ + L$ is a consequence of Theorem 4.

In the event that $V(t, x)$ is continuously differentiable with respect to each of its arguments, then it is a simple matter to show that the requirements of determinacy and transversality are together equivalent to the requirement that V satisfy the Hamilton-Jacobi equation (*vide* the remarks following Theorem 1)

$$V_t + H^0(t, x, V_x) = 0$$

with boundary condition (18), where

$$L(t, x, \varphi) + V_x \cdot f(t, x, \varphi) - H^0(t, x, V_x)$$

is determinate. A method of determining H^0 has been described by Kalman [4].

The effect of the transversality condition (18) is to make the value of

$$\int_{t_0}^{t^*(t_0, x_0, u)} V^+(t, x(t; t_0, x_0, u); u(t)) \, dt$$

independent of the control u. Thus, a more general approach to the determination of an optimal control relative to the criterion $Q(t_0, x_0, u)$ would be the discovery of a functional $L_1(t, x, \varphi)$ with the same properties as L and which satisfies the following conditions:

(a) $L(t, x, \varphi) + L_1(t, x, \varphi)$ is determinate;

$$(b)\quad \int_{t_0}^{t^*(t_0,x_0,u)} L_1(\tau, x(\tau; t_0, x_0, u), u(\tau))\, d\tau = -k(t_0, x_0)$$

for all $u \in U$.

Then if (a), (b) are satisfied, it follows from Theorem 1 that

$$k(t_0, x_0) = \min_{u \in U} Q(t_0, x_0, u).$$

The use of a particular functional L_1 satisfying (b) has been made the basis of the methods of Pontryagin [13] which, as is well known, stem from the results of Weierstrass in the classical calculus of variations.

REFERENCES

[1] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, Vol. V, Ann. Math. Studies, no. 45, (ed., L. Cesari, J. LaSalle, S. Lefschetz) Princeton, 1960, pp. 1–24.

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

[3] L. D. Berkovitz, *Variational methods in problems of control and programming*, ibid. *3* (1961), pp. 145–169.

[4] R. E. Kalman, *The theory of optimal control and the calculus of variations*, RIAS Tech. Rep. 61-3, 1961.

[5] C. Carathéodory, *Variationsrechnung und partielle Differentialgleichungen erster Ordnung*, Leipzig, 1935.

[6] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal. *8* (1961), pp. 36–58.

[7] J. L. Massera and J. J. Schäffer, *Linear differential equations and functional analysis III, Liapunov's second method in the case of conditional stability*, Ann. of Math. *69* (1959), pp. 535–574.

[8] A. C. Zaanen, *Linear Analysis*, Groningen, 1956.

[9] E. J. McShane, *Integration*, Princeton Press, Princeton, 1947.

[10] T. M. Apostol, *Mathematical Analysis*, Addison-Wesley, Boston, 1957.

[11] C. B. Morrey, Jr., *Functions of several variables and absolute continuity, II*, Duke Math. J. *6* (1940), pp. 187–215.

[12] T. Yoshizawa, *Liapunov's function and boundedness of solutions*, Funkcial. Ekvac. *2* (1959), pp. 95–142.

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

[14] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton Press, Princeton, 1961.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

RELAY TYPE CONTROL SYSTEMS WITH RETARDATION AND SWITCHING DELAY*

M. N. OĞUZTÖRELİ†

Abstract. In the present paper, we wish to investigate relay type control systems with retardation and switching delay. For this purpose, we extend some basic results on the continuation of solutions which are due to J. André and P. Seibert. Also, we extend some stability theorems of R. Bellman and K. L. Cooke, making use of their kernel function representation of the solutions. We also consider the dependence of the solutions upon switching delay.

1. Introduction. In recent publications, relay type control systems have attracted a great deal of attention from many authors. These control processes have been described mathematically by systems of ordinary differential equations with piecewise continuous right-hand sides.

In a previous paper [21] we investigated a time optimal control problem for a dynamical system described by a linear differential-difference equation, and have shown that the optimal control is "bang-bang." In the present paper, we wish to investigate relay type control systems with retardation and switching delay. For this purpose we shall give here a generalisation of some basic results, especially those of J. André and P. Seibert [2]. Since the solutions of a differential-difference equation can be continued only in the forward direction with respect to time, our method will be slightly different from that which is used for differential equations.

Let a control system be given by the piecewise linear differential-difference equation

$$(1.1)\quad x'(t + h_m) = \sum_{k=0}^{m} A_k(t)x(t + h_k) + B(t)\ \operatorname{sgn} s\{x[(t - \tau)]\},$$

where t is a real variable (time), the prime denotes differentiation with respect to t, τ (switching delay) and the "spans" $h_k (k = 0, 1, \cdots, m)$ are given constants such that

$$(1.2)\qquad 0 = h_0 < h_1 < h_2 < \cdots < h_m, \qquad \tau > 0.$$

In (1.1), $A_k(t) (k = 0, 1, \cdots, m)$ are given $n \times n$ matrix functions analytic for $t \geqq t_0$, $B(t)$ is a given $n \times r$ analytic matrix function for $t \geqq t_0$, $s(x)$ is a given r-dimensional analytic vector function in the n-dimensional Euclidean space R^n, $x(t)$ is an n-dimensional vector function

* Received by the editors October 26, 1962 and in revised form May 21, 1963.

† Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia.

giving the state of the control system at time t. The r-dimensional vector function sgn s is defined by

$$\operatorname{sgn} s = (\operatorname{sgn} s_1, \cdots, \operatorname{sgn} s_r), \tag{1.3}$$

where $s_1, \cdots, s_r$ are the components of s and

$$\operatorname{sgn} s_k = \begin{cases} +1 \text{ for } s_k > 0, \\ \text{undefined for } s_k = 0, \\ -1 \text{ for } s_k < 0. \end{cases} \tag{1.4}$$

We suppose that s_k and grad $s_k(x) (k = 1, 2, \cdots, r)$ do not vanish simultaneously at any point of the space R^n.

The corresponding *uncontrolled system* is described by the homogenous differential-difference equation

$$x'(t + h_m) = \sum_{k=0}^{m} A_k(t)x(t + h_k), \tag{1.5}$$

which has been extensively investigated in literature (see e.g. [5, 8, 10–14, 17, 19, 22–27]).

If, in particular, there is no retardation in (1.1) and if the matrices $A_0 = A$ and B are constants, we have

$$y'(t) = Ay(t) + B \operatorname{sgn} s\{y(t - \tau)\}. \tag{1.6}$$

This differential-difference equation has been investigated by J. André and P. Seibert [2] and an interesting problem related to (1.6) has been considered by R. Bass [4]. The differential equation (1.6) with $\tau = 0$ belongs to the class of differential equations with piecewise continuous right-hand sides, which is considered by many authors (see e.g. [1, 2, 3, 6, 9, 15, 17, 19, 20]).

Let us now consider the smooth hypersurfaces S_j in R^n defined by the equations

$$S_j = \{x \mid s_j(x) = 0\}, \qquad j = 1, 2, \cdots, r, \tag{1.7}$$

which are called *switching spaces*. The space R^n can be decomposed into domains D_j^+ and D_j^- in which $s_j(x) \gtrless 0$ respectively and S_j will be the common boundary of the domains D_j^+ and D_j^-, $j = 1, \cdots, r$. Furthermore, we denote

$$S = \bigcup_{j=1}^{r} S_j, \qquad D = R^n - S. \tag{1.8}$$

The right-hand side of (1.1) is discontinuous along the analytic hypersurface S.

Let $\mathfrak{M}$ be the set of all n-dimensional analytic vector functions in the *initial interval* $t_0 - \tau \leqq t \leqq t_0 + h_m$. The elements of the set $\mathfrak{M}$ will be called *initial functions* or *initial conditions.*

Following André-Seibert [2d], a continuous vector function $x(t)$ will be called a *solution* of the system (1.1) if

A. $x(t)$ satisfies (1.1) in D,

B. $x(t)$ has only isolated points in S.

The solution of the system (1.1) which satisfies the initial condition

$$x(t) = \phi(t), \qquad t_0 \leqq t \leqq t_0 + h_m, \tag{1.9}$$

with $\phi(t) \in \mathfrak{M}$, will be denoted by $x = x(t, \phi)$ or simply by $x = x(t)$.

Since, by our hypotheses, the matrix functions $B(t)$ and $A_k(t)(k = 0, 1, \cdots, m)$ are analytic for $t \geqq t_0$, and since $s(x)$ is also analytic, the differential-difference equation (1.1) has a unique continuous solution analytic from the right in the neighbourhood of $t = t_0 + h_m$, which satisfies (1.8), if the vector

$$\operatorname{sgn} \{s[\phi(t_0 + h_m' - \tau)]\}$$

is defined.

2. Continuation of a solution. Suppose that the value $\phi(t_0 + h_m - \tau)$ of the initial function $\phi(t) \in \mathfrak{M}$ is an interior point of one of the domains D, say D_j^+:

$$\phi(t_0 + h_m - \tau) \in D_j^+. \tag{2.1}$$

(The case $\phi(t_0 + h_m - \tau) \in D_j^-$ can be investigated similarly). Let $\mathfrak{M}_j^+$ be the set of all initial functions which possess this property. Put

$$e^+ = \operatorname{sgn} \{s[\phi(t_0 + h_m - \tau)]\}. \tag{2.2}$$

The vector e^+ has $+1$ as its jth component:

$$e_j^+ = \operatorname{sgn} \{s_j[\phi(t_0 + h_m - \tau)]\} = +1. \tag{2.3}$$

The solution of (1.1) which satisfies the initial condition (1.9) with $\phi(t) \in \mathfrak{M}_j^+$, satisfies the equation

$$x'(t + h_m) = \sum_{k=0}^{m} A_k(t)x(t + h_u) + B(t)e^+, \tag{2.4}$$

for $t \geqq t_0 + h_m$. Since this solution is analytic from the right in the neighbourhood of $t = t_0 + h_m$, it will remain in the domain D_j^+ for some further time interval. Let us denote this solution by $x^+(t, \phi)$ or simply by $x^+(t)$. Since the solution of a differential-difference equation with retarded argument can be continued in the forward direction with respect to time, we have to follow $x^+(t)$ for increasing t. Suppose the smallest $t(> t_0 + h_m)$

for which $x^+(t)$ reaches the switching space S_j is T, the point u that it reaches being given by

$$u = \lim_{t \to T-0} x^+(t, \phi). \tag{2.5}$$

Of course, T and u depend on the selection of the initial function $x = \phi(t)$ from the set $\mathfrak{M}_j^+$.

According to our definitions

$$\frac{d}{dt} x^+(t + h_m) = \sum_{k=0}^{m} A_k(t) x^+(t + h_m) + B(t)e^+, \tag{2.6}$$

and

$$x^+(t) = \phi(t) \quad \text{for} \quad t_0 \leqq t \leqq t_0 + h_m .$$

For the sake of simplicity we shall denote the right-hand side of (2.6) by $F(t, e^+)$:

$$F(t, e^+) = \sum_{k=0}^{m} A_k(t) x^+(t + h_k) + B(t)e^+. \tag{2.7}$$

The solution $x^+(t)$ can, with the help of the *kernel matrix* representation due to R. Bellman and K. L. Cooke [5c], be expressed as follows:

$$x^+(t + h_m) = X(t + h_m , \phi) + \int_{t_0}^{t} Y(s, t) B(s) e^+ \, ds, \tag{2.8}$$

$$(t + h_m < T + \tau).$$

Here $X(t, \phi)$ is that solution of the homogeneous (uncontrolled) system (1.5) which satisfies the initial condition (1.8), and $Y(s, t)$ is the kernel matrix. $Y(s, t)$ is the unique solution of the *adjoint equation*

$$\frac{\partial}{\partial s} Y(s, t) = \sum_{k=0}^{m} Y(s + h_m - h_k , t) A_k(s + h_m - h_k) \tag{2.9}$$

$$(t > t_0 \, ; \, t_0 < s < t - h_m \quad \text{and} \quad t - h_m < s < t)$$

satisfying the initial condition

$$Y(s, t) = \begin{cases} 0, & t < s < t + h_m , \\ I, & s = t, \end{cases} \tag{2.10}$$

where I is the identity matrix. $Y(s, t)$ is defined for $t > t_0$, $t_0 \leqq s \leqq t + h_m$ and is continuous for $t_0 \leqq s \leqq t$.

According to (2.4) and (2.8), we have

$$u = X(T, \phi) + \int_{t_0}^{T - h_m} Y(s, T - h_m) B(s) e^+ \, ds. \tag{2.11}$$

Since the function sgn $\{s[x^+(t-\tau)]\}$ is still continuous and equal to e^+ in the interval $T \leqq t < T+\tau$, our solution $x^+(t, \phi)$ will continue to satisfy the same differential-difference equation (2.6) in this interval. We put

$$
\begin{aligned}
u_\tau^* &= \lim_{t \to T+\tau-0} x^+(t, \phi) \\
&= X(T+\tau, \phi) + \int_{t_0}^{T+\tau-h_m} Y(s, T+\tau-h_m) B(s) e^+ \, ds.
\end{aligned} \tag{2.12}
$$

Of course, the point u_τ^* depends on the choice of the initial function $\phi(t)$ from the set $\mathfrak{M}_j^+$. We shall denote by $S_{j,\tau}^+$ the set of all points u_τ^*. Note that according to the analyticity in the interval $T \leqq t < T+\tau$ the solution $x^+(t, \phi)$ can intersect the switching surface S_j at most only at a finite number of points; that is, the intersections of $x^+(t, \phi)$ with S_j cannot have an accumulation point which is reached in the interval $T \leqq t \leqq T+\tau$.

In order to follow the further continuation of the solution $x^+(t, \phi)$ after $t = T+\tau$, we have to investigate the behaviour of the function $s_j[x^+(t, \phi)]$ in the neighborhood of $t = T$. For this purpose, we first assume that the point u belongs to only one switching space S_j. Therefore,

$$
s_k(u) = 0 \quad \text{for} \quad k = j, \qquad s_k(u) \neq 0 \quad \text{for} \quad k \neq j. \qquad (k = 1, 2, \cdots, r) \tag{2.13}
$$

Let e^- be the vector whose components are equal to those of e^+ except for for the jth, which is equal to -1.

We now consider the following differential-difference equations:

$$
x'(t+h_m) = F(t, e^{\pm}), \tag{2.14}
$$

and put

$$
\phi_1(t) = x^+(t, \phi) \quad \text{for} \quad T-h_m \leqq t \leqq T+\tau. \tag{2.15}
$$

This function is analytic in its interval of definition. Let $x^+(t, \phi_1)$ and $x^-(t, \phi_1)$ be the respective solutions of these equations which satisfy the same initial condition

$$
x(t) = \phi_1(t) \qquad T-h_m \leqq t \leqq T+\tau. \tag{2.16}
$$

Since all the conditions of the existence and uniqueness theorems for differential-difference equations are still satisfied, $x^+(t, \phi_1)$ and $x^-(t, \phi_1)$ are uniquely determined (analytic) functions. Obviously, we have

$$
x^+(T+\tau, \phi_1) = x^-(T+\tau, \phi_1) = u_\tau^*, \tag{2.17}
$$

and from (2.3), (2.15) and (2.16),

$$\lim_{t\to T-0} \operatorname{sgn}\{s_j[x^{\pm}(t,\phi_1)]\} = +1. \tag{2.18}$$

We can easily calculate the following limit

$$\gamma_j = \lim_{t\to T+0} \operatorname{sgn}\{s_j[x^{\pm}(t,\phi_1)]\}. \tag{2.19}$$

For this purpose, remembering the hypotheses which we laid on the functions $B(t)$, $A_k(t)(k = 0, 1, \cdots, m)$, $\phi(t)$ and $s(x)$, let us consider the Taylor expansion of $s_j[s^{\pm}(t,\phi_1)]$ at $t = T$:

$$\begin{aligned} s_j[x^{+}(t,\phi)] &= \frac{d}{d\alpha} s_j[x^{+}(\alpha,\phi)]_{\alpha=T}(t-T) \\ &\quad + \frac{1}{2}\frac{d^2}{d\alpha^2} s_j[x^{+}(\alpha,\phi)]_{\alpha=T'}(t-T)^2 \\ &= c_1(u)(t-T) + \tfrac{1}{2}c_2(T',u)(t-T)^2, \end{aligned} \tag{2.20}$$

where T' is an intermediate value between t and T, and where

$$c_i(t,u) = \frac{d^i}{dt^i} s_j[x^{+}(t,\phi)], \qquad c_i(u) = c_i(T,u), \qquad (i = 1, 2). \tag{2.21}$$

It can be verified easily that

$$c_1(u) = F(T - h_m, e^{+}) \left(\frac{d}{dx} s_j(x)\right)_{x=u}, \tag{2.22}$$

$$\begin{aligned} c_2(u) &= F^{*}(T - h_m, e^{+}) \left(\frac{d^2}{dx^2} s_j(x)\right)_{x=u} F(T - h_m, e^{+}) \\ &\quad + \left(\frac{d}{dt} F(t, e^{+})\right)_{t=T-h_m} \left(\frac{d}{dx} s_j(x)\right)_{x=u}. \end{aligned} \tag{2.23}$$

Here by * we denote the operation of transposition, while the operator d/dx is the gradient when applied to a scalar and the Jacobian matrix when applied to a vector. The operatior d^2/dx^2, applied to a scalar, denotes the Jacobian matrix of the gradient.

Since u is given by the formula (2.11), the quantities $c_1(u)$ and $c_2(u)$ are well defined and they determine the character of the switching at the point $x = u$. As we noted above, the switching operation actually happens when the trajectory reaches the point $x = u_\tau^{*} \in S_{j,\tau}^{+}$ at time $t = T_{+c}$, because of the switching delay τ.

If $c_1(u) \neq 0$, on account of (2.18), we have $c_1(u) < 0$. Therefore, $s_j < 0$ for $t > T$. Thus, we have the differential-difference equation

$$x'(t + h_m) = F(t, e^{-}) \quad \text{for} \quad t \geqq T + \tau. \tag{2.24}$$

If $c_1(u) = 0$, $c_2(u) \not\equiv 0$ on account of (2.18), we have $c_2(u) > 0$. Thus, we have the differential-difference equation

$$x'(t + h_m) = F(t, e^+) \quad \text{for} \quad t \geqq T + \tau. \tag{2.25}$$

If $c_1(u) = c_2(u) = 0$, the sign to be taken will be

$$\lim_{T' \to T+0} \operatorname{sgn} [c_2(T', u)].$$

Let $x(t, \phi_1)$ be the continuation of $x(t, \phi)$ obtained by this manner. $x(t, \phi_1)$ will clearly satisfy one of the equations (2.24) and (2.25) for $t \geqq T + \tau$ and the initial condition $x(t, \phi_1) = x(t, \phi)$ for $T - h_m - \tau$ $T \leqq T + \tau$. Since $x(t, \phi)$ is analytic for $t - h_m - \tau \leqq t \leqq T + \tau$, the solution $x(t, \phi_1)$ is also analytic for some time interval $T + \tau \leqq t \leqq T + \tau + \rho$, where ρ is a positive number. In this interval the solution $x(t, \phi_1)$ may intersect the switching space S at most only at a finite number of points because of the analyticity of the vector functions $s(x)$ and $x(t, \phi_1)$. Therefore there is no accumulation point in this interval of the intersections of $x(t, \phi_1)$ with S.

Let us now assume that the point u belongs to more than one switching space S_k. To investigate this case generally, consider a domain D whose boundary S consists of some parts of the switching spaces $S_{k_1}, S_{k_2}, \cdots, S_{k_i} (1 \leqq i \leqq r)$. Let us assume that $\phi(t_0 + h_m - \tau) \in D$ and denote by $\mathfrak{M}_D$ the set of all analytic functions $\phi(t)$ in $t_0 - \tau \leqq t \leqq t_0 + h_m$ having this property. In this case, the vector functions sgn $\{s[\phi(t)]\}$ is piecewise defined for $t_0 - \tau \leqq t \leqq t_0 + h_m$. Put

$$\sigma = \operatorname{sgn} \{s[\phi(t_0 + h_m - \tau)]\} \quad \text{for} \quad t_0 - \tau \leqq t \leqq t_0 + h_m. \tag{2.26}$$

Each component $\sigma_j (j = 1, 2, \cdots, r)$ of σ has one of the values ± 1, as before, and we can apply the same method as we used at the beginning of this section. Let $x(t, \phi)$ be the solution of (1.1) which satisfies the initial condition (1.9) so that now $x(t, \phi)$ will take the place of $x^+(t, \phi)$. Let $T (> t_0 + h_m)$ again denote the first time at which $x(t, \phi)$ reaches the switching space S (at point u, say) and suppose that u belongs to the switching spaces $S_{k_1}, S_{k_2}, \cdots, S_{k_\nu} (\nu \leqq i \leqq r)$. Let us now calculate the limits $\gamma_j (j = k_1, k_2, \cdots, k_\nu)$ by the formulae (2.20) to (2.23) and with the help of these limits calculate the components of the vector

$$\tilde{\sigma} = \lim_{t \to T+0} \operatorname{sgn} \{s[x(t, \phi)]\}. \tag{2.27}$$

Therefore, the continuation $x(t, \phi_1)$ of the solution $x(t, \phi)$ for $t \geqq T + \tau$ will satisfy the differential-difference equation

$$x'(t + h_m) = F(t, \tilde{\sigma}) \tag{2.28}$$

under the initial condition

$$x = \phi_1(t) = x(t, \phi) \quad \text{for} \quad T - h_m \leqq t \leqq T + \tau. \tag{2.29}$$

Since all the conditions of the existence and uniqueness theorems are still satisfied, $x(t, \phi_1)$ is uniquely determined.

From the analyticity of $x(t, \phi_1)$ and $s(x)$, the function $x(t, \phi_1)$ cannot intersect the switching space S at an infinite number of points in the interval of analyticity of $x(t, \phi_1)$. Consequently:

THEOREM 1. *Each solution $x(t, \phi)$ of the differential-difference equation (1.1) which satisfies the initial condition (1.9) with $\phi(t) \in \mathfrak{M}_D$ can be continued indefinitely into the future, i.e., they have no end-points.*

3. Stability properties of solutions. Stability properties of many kinds of systems with time lags have been investigated in literature (see, e.g. [5a, 11, 12, 13, 14, 17, 19, 22, 26]; for a complete bibliography on this subject see [7] and [19]). We shall now extend some stability theorems due to R. Bellman and K. L. Cooke [5c].

Let us suppose that all continuous solution $X(t, \phi)$ of the uncontrolled system (1.5) are bounded as $t \to +\infty$:

$$\| X(t, \phi) \| < c_1 < +\infty, \qquad t \geqq t_0, \tag{3.1}$$

where c_1 is a constant. Let the kernel function $Y(s, t)$ and the matrix function $B(t)$ be of the form

$$\| Y(s, t) \| < c_2 < +\infty, \qquad t_0 \leqq s \leqq t, \tag{3.2}$$

and

$$\int_{t_0}^{+\infty} \| B(t) \| \, dt < +\infty, \tag{3.3}$$

where c_2 is a constant. Furthermore, let $D_k(t)$ be continuous $n \times n$ matrix functions for $t \geqq t_0$ such that

$$\int_{t_0}^{+\infty} \| D_k(t) \| \, dt < +\infty, \qquad k = 0, 1, \cdots, m. \tag{3.4}$$

In (3.1) to (3.4), $\| Y(s, t) \|$, $\| X(t, \phi) \|$, $\| B(t) \|$ and $\| D_k(t) \|$ denote respectively the sums of the absolute values of the elements of Y, X, B and D_k, $k = 0, 1, \cdots, m$.

Consider now the differential-difference equation

$$x'(t + h_m) = \sum_{k=0}^{m} [A_k(t) + D_k(t)]x(t + h_k) + B(t) \operatorname{sgn} \{s[x(t - \tau)]\}. \tag{3.5}$$

THEOREM 2. *Under the hypotheses* (3.1) *to* (3.4) *all continuous solutions of the differential-difference equation* (3.5) *are uniformly bounded as* $t \to +\infty$.

Proof. To prove this, following Bellman and Cooke, we put

$$W(t) = \sum_{k=0}^{m} D_k(t)x(t + h_k) + B(t) \operatorname{sgn} \{s[x(t - \tau)]\}. \tag{3.6}$$

Continuous solutions of (3.5) can be obtained by the method of previous section. For a continuous solution $\tilde{x}(t)$ of (3.5), $W(t)$ is continuous in every domain D bounded by the switching spaces S_k and is integrable everywhere. Therefore, by the kernel function representation (2.8) we have, for $t \geqq t_0$,

$$\begin{aligned} \tilde{x}(t + h_m) = X(t) + \sum_{k=0}^{m} \int_{t_0}^{t} Y(s, t)D_k(s)\tilde{x}(s + h_k)\, ds \\ + \int_{t_0}^{t} Y(s, t)B(s) \operatorname{sgn} \{s[\tilde{x}(s - \tau)]\}\, ds. \end{aligned} \tag{3.7}$$

From (3.1) and (3.2) we have

$$\begin{aligned} &\| \tilde{x}(t + h_m) \| \leqq c_1 \\ &+ c_2 \sum_{k=0}^{m} \int_{t_0}^{t} \| D_k(s) \| \, \| \tilde{x}(s + h_k) \| \, ds \quad + c_2 \int_{t_0}^{t} \| B(s) \| \, ds \\ &\leqq c_1 + c_2 \sum_{k=0}^{m} \int_{t_0}^{t+h_k-h_m} \| D_k(s + h_m - h_k) \| \, \| x(s + h_m) \| \, ds, \end{aligned} \tag{3.8}$$

and using a well-known lemma (see e.g. [5a], p. 35), we find

$$\| \tilde{x}(t + h_m) \| \leqq c \exp\left\{c_2 \sum_{k=0}^{m} \int_{t_0}^{t+h_k-h_m} \| D_k(s + h_m - h_k) \| \, ds\right\}, \tag{3.9}$$

where

$$c = c_1 + c_2 \int_{t_0}^{+\infty} \| B(s) \| \, ds. \tag{3.10}$$

It follows, from (3.4) and (3.9), that $\tilde{x}(t)$ is uniformly bounded as $t \to +\infty$. Let us suppose that

$$\| Y(s, t) \| \leqq \Psi(t) \quad \text{for} \quad t \geqq t_0, \tag{3.11}$$

where $\Psi(t)$ is a monotonic decreasing function such that

$$\lim_{t \to +\infty} \Psi(t) = 0. \tag{3.12}$$

In this case, we have the following result:

THEOREM 3. *If* (3.3), (3.11) *and* (3.12) *are satisfied, then the solution* $x(t, \phi)$ *of the system* (1.1) *is asymptotic to the solution* $X(t, \phi)$ *of the uncontrolled system* (1.5), i.e.,

$$\lim_{t \to +\infty} \| x(t, \phi) - X(t, \phi) \| = 0. \tag{3.13}$$

Proof. From the kernel function representation (2.8) and from the hypothesis (3.11), we have

$$\| x(t, \phi) - X(t, \phi) \| \leqq 2\psi(t) \int_{t_0}^{t} \| B(s) \| \, ds.$$

If we now make use of the hypothesis (3.3), we obtain

$$\| x(t, \phi) - X(t, \phi) \| < \epsilon$$

for sufficiently large t. This proves the theorem.

Consider now two solutions of the uncontrolled system (1.5), namely $X(t, \phi)$ and $\tilde{X}(t, \tilde{\phi})$, corresponding to the continuous initial functions $\phi(t)$ and $\tilde{\phi}(t)$ in the same initial interval $t_0 - \tau \leqq t \leqq t_0 + h_m$. Supposing that the system (1.5) has a Lipschitz constant L, that is, that

$$\sum_{k=0}^{m} \| A_k(t) \| \leqq L \quad \text{for} \quad t \geqq t_0. \tag{3.14}$$

Then, as is well known (see [17]),

$$\| X(t, \phi) - \tilde{X}(t, \tilde{\phi}) \| \leqq e^{L(t-t_0)} \| \phi - \tilde{\phi} \|^{(*)}, \tag{3.15}$$

where

$$\| z \|^{(*)} = \operatorname*{Sup}_{t_0 - \tau \leqq t \leqq t_0 + h_m} \| z(t) \|. \tag{3.16}$$

THEOREM 4. *If* (3.3), (3.11), (3.12) *and* (3.14) *are satisfied and if it is possible to choose a neighbourhood of* $\phi(t)$,

$$N_\delta(\phi) = \{ \| \phi - \tilde{\phi} \| < \delta \}, \tag{3.17}$$

such that

$$\| X(t, \phi) - \tilde{X}(t, \tilde{\phi}) \| < \frac{\epsilon}{2}, \qquad \tilde{\phi} \in N_\delta(\phi), \tag{3.18}$$

holds for all sufficiently large t, *where* ϵ *is an arbitrarily small positive number and* δ *depends on* ϵ, *then*

$$\| x(t, \phi) - \tilde{x}(t, \tilde{\phi}) \| < \epsilon \quad \text{and} \quad \| \tilde{x}(t, \tilde{\phi}) - X(t, \phi) \| < 2\epsilon$$

for $\tilde{\phi} \in N_\delta(\phi)$ *and for* t *large enough.*

Proof. Let $x(t, \phi)$, $\tilde{x}(t, \tilde{\phi})$, $X(t, \phi)$ and $\tilde{X}(t, \tilde{\phi})$ be previously defined

solutions of (1.1) and (1.5). With the help of the kernel function representation (2.8) of the solutions $x(t, \phi)$ and $\tilde{x}(t, \tilde{\phi})$ and also of (3.11) we can write

$$\| x(t, \phi) - \tilde{x}(t, \tilde{\phi}) \| \leqq \| X(t, \phi) - \tilde{X}(t, \tilde{\phi}) \| + 2\psi(t) \int_{t_0}^{t} \| B(s) \| \, ds.$$

From the boundedness of the last integral, by (3.3), and also from the hypotheses (3.12) and (3.17), we have, for sufficiently large t and $\tilde{\phi} \in N_\delta(\phi)$,

$$\| X(t, \phi) - \tilde{X}(t, \tilde{\phi}) \| < \frac{\epsilon}{2}, \qquad 2\psi(t) \int_{t_0}^{t} \| B(s) \| \, ds < \frac{\epsilon}{2}.$$

Therefore,

$$\| x(t, \phi) - \tilde{x}(t, \tilde{\phi}) \| < \epsilon \tag{3.19}$$

for $\tilde{\phi} \in N_\delta(\phi)$ and for t large enough. To prove the last part of the theorem, observe that

$$\| \tilde{x}(t, \tilde{\phi}) - X(t, \phi) \| \leqq \| x(t, \phi) - \tilde{x}(t, \tilde{\phi}) \| + \| x(t, \phi) - X(t, \phi) \|. \tag{3.20}$$

From (3.13) and (3.19), we have

$$\| \tilde{x}(t, \tilde{\phi}) - X(t, \phi) \| < 2\epsilon \quad \text{for} \quad \tilde{\phi} \in N_\delta(\phi), \tag{3.21}$$

for $t \to +\infty$, which proves the theorem.

We note that the above results can be extended immediately to the case where the system (1.1) is affected by small disturbing forces. In this case the control system is of the form

$$x'(t + h_m) = \sum_{k=0}^{m} A_k(t) x(t + h_k) + B(t) \operatorname{sgn} \{s[x(t - \tau)]\} + r(t, x(t)), \tag{3.22}$$

where $r(t, x(t))$ is a summable function such that

$$\| r(t, x(t)) \| < \eta \tag{3.23}$$

for all $t \geqq t_0$ and x and η a small positive constant.

4. Limits of the solutions of (1.1) as the retardations and switching delay approach zero. Let D be any domain bounded by the switching spaces S_k and let $\phi(t) \in \mathfrak{M}_D$, where $\mathfrak{M}_D$ is the set of all analytic functions in D for $t_0 \leqq t \leqq t_0 + h_m$. Consider now the distance between the points u and u_τ^* defined by (2.11) and (2.12). We have

$$
\begin{aligned}
\| u_\tau^* - u \| &\leqq \| X(T+\tau, \phi) - X(T, \phi) \| \\
&+ \left\| \int_{t_0}^{T-h_m} [Y(x, T+\tau-h_m) - Y(s, T-h_m)]B(s)e^+ \, ds \right\| \\
&+ \left\| \int_{T-h_m}^{T+\tau-h_m} Y(s, T+\tau-h_m)B(s)e^+ \, ds \right\| .
\end{aligned} \tag{4.1}
$$

Let ϵ be an arbitrarily small positive number. From the continuity of $X(t, \phi)$, $Y(s, t)$ and $B(t)$ we can write

$$\| X(T+\tau, \phi) - X(T, \phi) \| \leqq \frac{\epsilon}{3},$$

and

$$\| Y(s, T+\tau-h_m) - Y(s, T-h_m) \| < \frac{\epsilon}{3(T-h_m-t_0)c_3}$$

for $\tau < \delta$, $\delta = \delta(\epsilon, \phi) > 0$, and

$$\left\| \int_{T-h_m}^{T+\tau-h_m} Y(s, T-h_m)B(s)e^+ \, ds \right\| < \frac{\epsilon}{3}$$

for $\tau < \dfrac{\epsilon}{3c_2 c_3}$, where c_2 is given by (3.2) and

$$c_3 = \sup_{t_0 \leqq t \leqq T} \| B(t) \|.$$

Therefore, if

$$\delta < \frac{\epsilon}{3c_2 c_3}, \tag{4.2}$$

we have

$$\| u_\tau^* - u \| < \epsilon \quad \text{for} \quad \tau < \delta. \tag{4.3}$$

By the same argument we see that

$$\lim_{\tau \to 0} x_\tau(t, \phi) = x_0(t, \phi), \tag{4.4}$$

where $x_\tau(t, \phi)$ denotes the solution of the system (1.1) with the switching delay τ. Obviously these results are also true for any continuation of $x_\tau(t, \phi)$ according to the method given in section 2.

R. Bellman and K. L. Cooke showed [5d] that

$$\lim_{h_m \to 0} x(t, \phi) = y(t), \tag{4.5}$$

where $y(t)$ is the solution of the equation

$$
\begin{aligned}
y'(t) &= A(t)y(t) + B(t)\ \operatorname{sgn}\{s[y(t-\tau)]\}, \\
y(t_0) &= \phi(t_0), \qquad A(t) = \sum_{k=0}^{m} A_k(t),
\end{aligned}
\tag{4.6}
$$

and the convergence is uniform for some interval $t_0 \leqq t \leqq T_0$, provided we stay in a domain D bounded by switching spaces. The equation (4.6) is of the form (1.6), which has been investigated by J. André and P. Seibert [2]. Clearly, this limiting process may be accomplished for any continuation of $x(t, \phi)$ beyond the domain D by the method of section 2. The solution of (4.6) gives an approximation to the limiting behaviour of the solutions of (1.1) for sufficiently small time lags.

5. Acknowledgment. The author wishes to express his appreciation to Dr. P. Seibert for his helpful comments and to the referee for his valuable remarks.

REFERENCES

[1] Y. I. Alimov, and E. A. Barbashin. *On the theory of differential equations for relay systems*, Reports from Institutes of Higher Education, Math., Kazan, No. 1 (26) 1962, pp. 3–13. (Russian).

[2a] J. André and P. Seibert. *Über stückweise lineare differentialgleichungen, die bei regelungsproblemen auftreten I–II*, Archiv der Mathematik, 7 (1956) pp. 148–156, 157–164.

[2b] J. André and P. Seibert. *The local theory of piecewise continuous differential equations, I. Ideal systems, Contributions to the Theory of Nonlinear Oscillations*, Vol. V, Princeton Press 1960, pp. 225–255.

[2c] J. André and P. Seibert. *After end-point motions of general discontinuous control systems and their stability properties*, Automatic and Remote Control, 1960, pp. 919–922.

[2d] J. André and P. Seibert. *Piecewise continuous differential equations*, Bol. Soc. Mat. Mexicana, (1961) pp. 242–245.

[3a] M. A. Aizerman and F. R. Gantmakher. *On the stability of the equilibrium positions for discontinuous systems*, Prikl. Mat. Meh. 24 (1960) pp. 283–293 (Russian); English translation: Journal of Applied Mathematics and Mechanics, 24 (1960) pp. 406–421.

[3b] M. A. Aizerman and F. R. Gantmakher. *Some aspects of the theory of a nonlinear automatic control with discontinuous characteristics*, Automatic and Remote Control, (1960) pp. 1399–1403.

[3c] M. A. Aizerman and A. I. Lurie. *Method for constructing of periodic motions in piecewise linear systems*, International Union of Theoretical and Applied Mechanics, Symposium on Non-linear Vibrations, Kiev, September 1961.

[4] R. Bass. *A generalization of the functional relation $Y(s+t) = Y(s)\cdot Y(t)$ to piecewise linear difference differential equations*, Quart. Appl. Math., 14 (1956) pp. 415–417.

[5a] R. Bellman. *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1953.

[5b] R. Bellman. *On the existence and boundedness of solutions of nonlinear differential-difference equations*, Ann. of Math., 50 (1949) pp. 347–355.

[5c] R. Bellman and K. L. Cooke. *Stability theory and adjoint operators for linear differential-difference equations*, Trans. Amer. Math. Soc., 92 (1959) pp. 470–500.

[5d] R. Bellman and K. L. Cooke. *On the limit of solutions of differential-difference equations as the retardation approaches zero*, Proc. Nat. Acad. Sci. U.S.A. 45 (1959) pp. 1026–1028.

[5e] R. Bellman and J. M. Danskin. *A survey of the mathematical theory of time-lag, retarded argument, and hereditary processes*, The RAND Corporation, Report R-256, 1956.

[6] D. Bushaw. *Optimal discontinuous forcing terms, contributions to the theory of nonlinear oscillations*, Princeton Press, Vol. IV, 1958, pp. 29–52.

[7] N. H. Choksy. *Time-lag systems—a bibliography*, I.R.E. Trans. Prof. Group on Automatic Control, 5AC (1960) pp. 66–70.

[8] L. E. El'sgol'ts. *Differential equations*, Hindustan Publ. Corp. (India), Delhi, 1961, Chap. V.

[9] A. F. Filippov. *Differential equations with a discontinuous right-hand side*, Mat. Sbornik, 51 (93), 1960, pp. 99–128.

[10] J. Franklin. *On the existence of the solutions of functional differential equations*, Proc. Amer. Math. Soc., 5 (1954) pp. 363–369.

[11] V. E. Germaidze. *On asymptotic stability of systems with retarded argument*, Uspehi Mat. Nauk, 14 (1950) 4 (88), pp. 149–156.

[12] W. Hahn. *Zur stabilität der lösungen von linearen differential-differenzengleichungen mit konstanten koeffizienten*, Math. Annalen, Bd. 131 (1956) pp. 151–166.

[13] A. Halanay. *Asymptotic stability and small perturbations of periodic systems of differential equations with retarded argument*, Uspehi Mat. Nauk, 17 (1962) No. 1 (103), pp. 231–233.

[14] J. K. Hale. *Asymptotic behaviour of the solutions of differential-difference equations*, RIAS Technical Reports: 61-10, 1961.

[15] R. E. Kalman. *The theory of optimal control and calculus of variations*, RIAS Technical Reports: 61-3, 1961.

[16] M. Z. Kolosovskii. *On Conditions for existence of periodic solutions of systems of differential equations with discontinuous right-hand sides containing a small parameter*, Prikl. Mat. Meh. 24 (1960) pp. 738–745 (Russian); English translation: Journal of Applied Mathematics and Mechanics, 24 (1960) pp. 1105–1118.

[17] N. Krasovskii. *Some Problems in the Theory of Stability of Motion*, Moscow, 1959.

[18] I. V. Livartovskii. *The stability of a solution of a system of differential equations with discontinuous right-hand sides*, Prikl. Mat. Meh., 23 (1959), pp. 598–603; English translation: Journal of Applied Mathematics and Mechanics, 23 (1959) pp. 850–859.

[19] A. D. Myskis, S. N. Shimanov, and L. E. El'sgol'ts. *Stability and oscillations of systems with time lag*, International Union of Theoretical and Applied Mechanics, Symposium on Nonlinear Vibrations, Kiev, September 1961.

[20] Yu. I. Neimark. *Method of point mappings in the theory of nonlinear oscillations*, International Union of Theoretical and Applied Mechanics, Symposium on Nonlinear Vibrations, Kiev, September 1961.

[21] M. N. Oğuztöreli. *On a time optimal control problem*, J. Soc. Indust. Appl. Math. Ser. A. On Control. 1 (1963) pp.

[22] S. N. Shimanov. *On the instability of the motion of systems with retardation*, Prikl. Mat. Meh., 24 (1960) pp. 55–63 (Russian); English translation. Journal of Applied Mathematics and Mechanics, 24 (1960) pp. 70–81.

[23] S. Sugiyama. *Existence theorems on difference-differential equations*, Proc. Japan Acad., 38, (1962) pp. 145–149.

[24] E. M. Wright. *The linear difference-differential equations*, Proc. Roy. Soc. Edinburgh, Sect. A, 62, (1949) pp. 387–393.

[25] W. I. Zubov. *On the theory of linear stationary systems with lagging argument*, Reports from Institutes of Higher Education, Math., 1958, No. 6 (7), pp. 86–95 (Russian).

[26] A. M. Zverkin. *Dependence of the stability of solutions of linear differential equations with lagging argument upon the choice of initial moment*, Vestnik Moskov Univ., Ser. Mat. Meh. Astr. Fiz. Him., No: 5, pp. 15–20 (Russian).

[27] E. Pinney. *Ordinary Difference-Differential Equations*, University of California Press, Berkeley, 1959.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

A TIME OPTIMAL CONTROL PROBLEM FOR SYSTEMS DESCRIBED BY DIFFERENTIAL DIFFERENCE EQUATIONS*

M. N. OĞUZTÖRELI†

Abstract. The aim of this paper is to establish the solution of an optimal time control problem for a physical system whose state is described by a linear differential-difference equation with retarded argument. We have obtained here a generalisation of the results of Bellman and his collaborators Glicksberg, Gross and Kalaba, and of LaSalle and Neustadt by using a technique due to LaSalle, with the help of the kernel matrix representation of Bellman and Cooke and also a new integral representation for the solutions of linear differential-difference equations.

1. Introduction. We consider a control system given by a linear differential-difference equation with retarded argument of the form

$$(1.1)\quad x'(t + h_m) = \sum_{k=0}^{m} A_k(t)x(t + h_k) + B(t)u(t) + f(t),$$

where t is a real variable (time), $' = \frac{d}{dt}$, h_k $(k = 0, 1, \cdots, m)$ are given constants such that

$$0 = h_0 < h_1 < h_2 < \cdots < h_m,$$

$A_k(t)$ $(k = 0, 1, \cdots, m)$ are given $n \times n$ continuous matrix functions, $B(t)$ is a given continuous $n \times r$ matrix function, $x(t)$ and $f(t)$ are n-dimensional continuous vector functions, $x(t)$ giving the state of the control system at time t, and $u(t)$ is an r-dimensional vector function. We suppose that to control our system we are free to choose the "steering" function u. We assume that the functions $u(t)$ are "admissible" if they are piecewise continuous (or measurable) and have components less than 1 in absolute value: $| u_k(t) | \leqq 1, k = 1, \cdots, r$.

We denote by U the set of all r-dimensional vector functions $u(t)$ measurable on each finite interval $[t_0, t]$ and $| u_k(t) | \leqq 1$, $k = 1, \cdots, r$. Let U^0 be the set of all functions $u^0(t)$ in U with $| u^0(t) | = 1$, $k = 1, \cdots, r$. The set U is the set of allowable steering functions for our control problem, and U^0 is the set of "bang-bang" steering functions.

We denote by S a compact closed (bounded) convex subset of the set C of all real-valued n-dimensional continuous vector functions in the interval $t_0 \leqq t \leqq t_0 + h_m$.

* Received by the editors August 8, 1962 and in revised form July 31, 1963.

† Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia.

A solution $x(t)$ of the system (1.1) which satisfies the initial condition

$$x(t) = \phi(t) \qquad t_0 \leqq t \leqq t_0 + h_m , \tag{1.2}$$

eivdently depends on the choice of the 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 (1.2).

Now let us consider a moving particle $z(t)$. $z(t)$ is an n-dimensional continuous vector function.

Our control problem for the system (1.1) is to hit the particle $z(t)$ with the system $x(t, \phi, u)$ in minimum time. We say that an admissible steering function u and an initial function $\phi(t) \in S$ are optimal if $x(t, \phi, u) = z(t)$ for some $T > t_0 + h_m$ and if $x(t, \phi, u) \neq z(t)$ for $t_0 + h_m < t < T$ and all $u \in U, \phi \in S$.

This optimal time control depends not only on the choice of the steering function $u(t)$ but also on the choice of the initial function $x(t) = \phi(t)$ in the interval $[t_0 , t_0 + h_m]$. One should therefore consider, together with the problem of determining the optimal steering function, that of choosing the optimal initial function.

For $m = 0$, $A_0(t) = A(t)$, the system (1.1) reduces to

$$x'(t) = A(t)x(t) + B(t)u(t) + f(t), \tag{1.3}$$

for which the time control problem has been extensively described in the literature (see for example: [6, 8, 16, 17, 19]). The optimal steering function $u(t)$ of the system (1.3) has the bang-bang property. In this case, the right-hand side of (1.3) is piecewise continuous. This kind of differential equations has been investigated by many authors (see for example: [1, 11]).

The linear differential-difference equation (1.1) has been investigated by Bellman and Cooke [4].

Bellman [3] and his collaborators Glicksberg and Gross [5] and Kalaba [7] have treated terminal control problems involving linear systems with retardation. Recently Haratišvili [15] investigated a time optimal control problem involving delay. In all these contributions the initial functions $\phi(t)$ are fixed. We shall suppose here that the functions $\phi(t)$ range over the set S defined above. This assumption gives rise to new features.

2. Reduction of the problem to an integral equation. Let us consider now the time optimal control problem, described in the introduction, for the system

$$x'(t + h_m) + \sum_{k=0}^{m} A_k(t)x(t + h_k) = B(t)u(t) + f(t), \tag{2.1}$$

which is equivalent to (1.1). Under the hypotheses on the functions $A_k(t)$,

$B(t)$, $f(t)$, $u(t)$ and $\phi(t)$, which are stated in the Introduction, it is well known that (Cf. [4], pp. 475-477)

i) There is a unique continuous solution of (2.1) for $t > t_0$ which satisfies the initial condition

$$x(t, \phi, u) = \phi(t) \qquad t_0 \leqq t \leqq t_0 + h_m ; \tag{2.2}$$

ii) The kernel matrix $Y(s, t)$ of the equation (2.1) defined for $t > t_0$, $t_0 \leqq s \leqq t + h_m$, is the unique continuous solution for $t_0 \leqq s \leqq t$ of the adjoint equation of (2.1)

$$\frac{\partial}{\partial s} Y(s, t) = \sum_{k=0}^{m} Y(s + h_m - h_k, t) A_k(s + h_m - h_k) \tag{2.3}$$

$$(t > t_0 ; t_0 < s < t - h_m \quad \text{and} \quad t - h_m < s < t)$$

which satisfies the initial condition

$$Y(s, t) = \begin{cases} 0, & t < s \leqq t + h_m , \\ I, & t = s, \end{cases} \tag{2.4}$$

where I is the identity matrix.

iii) The unique continuous solution of (2.1) for $t > t_0$ satisfying the initial condition

$$x(t) = 0 \qquad t_0 \leqq t \leqq t_0 + h_m , \tag{2.5}$$

is given by the integral formula

$$x(t + h_m) = x(t + h_m, 0, u) = \int_{t_0}^{t} Y(s, t) v(s)\, ds \tag{2.6}$$

$$(t > t_0)$$

where

$$v(t) = B(t)u(t) + f(t). \tag{2.7}$$

iv) The solution of the system (2.1) satisfying the initial condition (2.2) has the form

$$x(t + h_m, \phi, u) = X(t + h_m, \phi) + \int_{t_0}^{t} Y(s, t) v(s)\, ds, \tag{2.8}$$

where $X(t, \phi)$ is the solution of the corresponding homogeneous system

$$x'(t + h_m) + \sum_{k=0}^{m} A_k(t) x(t + h_k) = 0, \tag{2.9}$$

which satisfies the initial condition (2.2).

Consequently, the state $x(t, \phi, u)$ of the control system (2.1) at time t

is known by the formula (2.8). As described in section 1, for the time optimal control problem, the particle $z(t)$ must be reached by the system $x(t, \phi, u)$ in minimum time. We want, therefore, at some time t to have

$$x(t + h_m, \phi, u) = z(t + h_m), \tag{2.10}$$

that is, to have

$$\int_{t_0}^{t} Y(s, t)B(s)u(s)\, ds = w(t, \phi), \tag{2.11}$$

where

$$w(t, \phi) = z(t + h_m) - X(t + h_m, \phi) - \int_{t_0}^{t} Y(s, t)f(s)\, ds. \tag{2.12}$$

On account of the continuity of the functions $A_k(t)$, $B(t)$, $\phi(t)$ and $z(t)$, the integral equation (2.11) has quite the same form as an integral equation considered by LaSalle [17] and Neustadt [19]; they differ only in the functions $X(t, \phi)$ and $Y(s, t)$; in fact, here the function $X(t, \phi)$ depends also on the initial function $\phi(t)$ selected from the set S, and the function $Y(s, t)$ depends on both variables s and t; but, in the case considered by LaSalle, the function X depends only on the variable t and the function Y solely on s, but not on t.

3. An integral representation for the solution of the homogeneous equation (2.9). Consider the homogeneous differential-difference equation (2.9) corresponding to (2.1), and its solution $X(t, \phi)$ which satisfies (2.2). Evidently $X(t, \phi)$ is a linear functional in ϕ over the set C. By a well-known theorem of functional analysis (see for example [20], p. 62), we can express this functional as follows:

$$X(t + h_m, \phi) = \int_{t_0}^{t_0+h_m} K(s, t)\phi(s)\, ds, \tag{3.1}$$

where the kernel $K(s, t)$ is defined uniquely by the functional $X(t, \phi)$. From (2.2) and (3.1) we have

$$\int_{t_0}^{t_0+h_m} K(s, t)\phi(s)\, ds = \phi(t), \qquad t_0 \leqq t + h_m \leqq t_0 + h_m, \tag{3.2}$$

for all $\phi(t) \in C$. By differentiation in (3.1) and using (2.9) and (3.1), we obtain

$$\int_{t_0}^{t_0+h_m} \left\{ \frac{\partial K(s, t)}{\partial t} + \sum_{k=0}^{m} A_k(t - h_m)K(s, t + h_k - h_m) \right\} \phi(s)\, ds = 0. \tag{3.3}$$

This equation holds for every $\phi(t) \in C$. Hence

$$\frac{\partial K(s,t)}{\partial t} + \sum_{k=0}^{m} A_k(t-h_m)K(s,t+h_k-h_m) = 0. \tag{3.4}$$

Therefore the kernel $K(s, t)$ is the matrix solution of the homogeneous differential-difference equation (2.9) satisfying the condition (3.2) for all $\phi(t) \in C$. With the help of this kernel function the solution $X(t, \phi)$ of (2.9) which satisfies the initial condition (2.2) is given by the integral formula (3.1).

4. The functional $\Omega(t, \phi, u)$. Let us now consider the functional

$$\Omega(t,\phi,u) = \int_a^b K(s,t)\phi(s)\,ds + \int_a^t Z(s,t)u(s)\,ds, \tag{4.1}$$

where

$$Z(s,t) = Y(s,t)B(s) \tag{4.2}$$

and $a = t_0$, $b = t_0 + h_m$.

From the formulas (2.11), (2.12), (3.1) and (4.1) we can write

$$\Omega(T,\phi,u) = C(T), \tag{4.3}$$

where

$$C(T) = z(T+h_m) - \int_a^T Y(s,T)f(s)\,ds, \tag{4.4}$$

and T is the time at which the control system $x(t, \phi, u)$, defined by (2.1), first hits the particle $z(t)$. T is the first root of the equation (2.10) which is greater than $b = t_0 + h_m$. Obviously $T = T(\phi, u)$. By its definition it is single-valued. It is also continuous in both ϕ and u, because $C(t)$ is a continuous function and $\Omega(t, \phi, u)$ is a continuous functional in ϕ and u.

Consider now the set

$$E(t) = \{\Omega(t,\phi,u);\ \phi \in S,\ u \in U\}. \tag{4.5}$$

We shall now extend some theorems due to LaSalle [17].

THEOREM 1. *$E(t)$ is convex.*

Proof. Let E^1 and E^2 be two elements of the set $E(t)$:

$$E^i = \Omega(t,\phi^i,u^i), \qquad \phi^i \in S, \qquad u^i \in U, \qquad i = 1,2.$$

Let α and β be two non-negative numbers such that $\alpha + \beta = 1$. Consider the functions

$$\phi^0 = \alpha\phi^1 + \beta\phi^2, \qquad u^0 = \alpha u^1 + \beta u^2.$$

Clearly $\phi^0 \in S$, $u^0 \in U$. Put $E^0 = \Omega(t, \phi^0, u^0)$. Since $E^0 = \alpha E^1 + \beta E^2$ and $\phi^0 \in S$, $u^0 \in U$, we have $E^0 \in E(t)$.

THEOREM 2. *$E(t)$ is closed.*

Proof. Consider a sequence of points $E^i \in E(t)$, with $E^i \to E^0$, $i \to \infty$:

$$E^i = \Omega(t, \phi^i, u^i) = \int_a^b K(s, t)\phi^i(s)\, ds + \int_a^t Z(s, t)u^i(s)\, ds,$$

$\phi^i \in S$, $u^i \in U$, $i = 1, 2, 3, \cdots$. We shall prove that $E^0 \in E(t)$. Since $u^i(t)$ are admissible, their components $u_j{}^i(t)$ are uniformly bounded in absolute value. Therefore, there exists a subsequence $\{u^{i_k}(t)\}$ of $\{u^i(t)\}$, and measurable functions $u_j{}^0(t)$ in $L^2(a, t)$ such that for $j = 1, \cdots, r$,

$$u_j^{i_k}(t) \xrightarrow{\text{Weakly}} u_j{}^0(t)$$

in $L^2(a, t)$, and $u^0(t) = (u_1{}^0(t), \cdots, u_r{}^0(t))$ may be chosen to be admissible.

Let us consider now the sequence $\{\phi^{i_k}(t)\}$. According to the definition of the set S, it is closed and compact. Therefore, we can select a subsequence

$$\{\phi^{i_{k'}}(t)\},$$

which tends uniformly to a limite function $\phi^0(t) \in S$. The subsequence $\{u^{i_{k'}}(t)\}$ tends to the limit function $u^0(t)$, obtained before. Without loss of generality we may assume that

$$\phi_j{}^i(t) \xrightarrow[i\to\infty]{} \phi_j{}^0(t) \qquad (j = 1, 2, \cdots, n)$$

uniformly, and

$$u_j{}^i(t) \xrightarrow[i\to\infty]{} u_j{}^0(t) \qquad (j = 1, 2, \cdots, r)$$

weakly. Therefore, by the definition of weak convergence and also uniform convergence, $E^0 = \Omega(t, \phi^0, u^0) \in E(t)$:

$$E^i \xrightarrow[i\to\infty]{} E^0 = \int_a^b K(s, t)\phi^0(s)\, ds + \int_a^t Z(s, t)u^0(s)\, ds.$$

Consequently, the limite of a convergent sequence in $E(t)$ belongs to $E(t)$. This proves the theorem.

Consider now the set

$$E^0 = \{\Omega(t, \phi, u^0); \phi \in S, u^0 \in U^0\}, \tag{4.6}$$

where U^0 is the set of all bang-bang functions.

THEOREM 3. *$E(t) = E^0(t)$.*

Proof. Define, for every fixed $t > a$, the subsets

$$M(t) = \left\{\int_a^t Z(s, t)u(s)\, ds;\ u \in U\right\}, \tag{4.7}$$

and

$$M^0(t) = \left\{ \int_a^t Z(s, t)u^0(s)\, ds;\ u^0 \in U^0 \right\}, \tag{4.8}$$

where $Z(s, t)$ is defined by (4.2). The $n \times r$ matrix function $Z(s, t)$ is continuous for $a \leqq s \leqq t$, because the kernel matrix $Y(s, t)$ is continuous for $a \leqq s \leqq t$ by section 2 and $B(s)$ is continuous by hypothesis.

Let $z^j(s, t)$ be the jth ($j = 1, \cdots, r$) column vector in $Z(s, t)$. We define

$$M_j(t) = \left\{ \int_a^t z^j(s, t)u_j(s)\, ds;\ u \in U \right\},$$

and

$$M_j{}^0(t) = \left\{ \int_a^t z^j(s, t)u_j{}^0(s)\, ds;\ u^0 \in U^0 \right\}.$$

Thus we have

$$\int_a^t Z(s, t)u(s)\, ds = \sum_{j=1}^r \int_a^t z^j(s, t)u_j(s)\, ds,$$

and

$$\int_a^t Z(s, t)u^0(s)\, ds = \sum_{j=1}^r \int_a^t z^j(s, t)u_j{}^0(s)\, ds,$$

and also

$$M(t) = \sum_{j=1}^r M_j(t), \qquad M^0(t) = \sum_{j=1}^r M_j{}^0(t).$$

For all fixed t ($> a$) the n-dimensional vectors $z^j(s, t)$, $j = 1, 2, \cdots, r$, are continuous for $a \leqq s \leqq t$, and, therefore, satisfy the conditions of LaSalle's first lemma [17]. Hence $M_j{}^0(t) = M_j(t)$ for each $j = 1, \cdots, r$, and therefore $M^0(t) = M(t)$. The proof of Theorem 3 is an immediate consequence of this extended LaSalle's lemma.

By the above theorem, anything that can be accomplished in time t by by $u(t) \in U$ can also be accomplished in time t by $u^0(t) \in U^0$. Consequently, we can state the following theorems due to LaSalle [17] for our control system (2.1):

THEOREM 4. *If of all $u^0(t) \in U^0$ there is an optimal one relative to U^0, then it is optimal relative to U.*

THEOREM 5. *If there is an optimal control function, then there is always a bang-bang control that is optimal.*

5. Existence of optimal functions. We shall prove now the following existence theorem, which is an extension of that due to LaSalle [17].

THEOREM 6. *If there is a pair of functions* $\phi \in S$, $u \in U$, *such that*

$$x(t, \phi, u) = z(t),$$

then there exists a pair of functions $\phi^0 \in S$, $u^0 \in U$, *which are optimal.*

Proof. By hypothesis the set

$$\mathfrak{M} = \{T,\ x(T, \phi, u) = z(T);\ \phi \in S,\ u \in U\} \tag{5.1}$$

is not empty. Let T^0 be the greatest lower bound of all T $(> b)$ which belong to the set $\mathfrak{M}$. By sections 2 and 4, we have, for $T \in \mathfrak{M}$,

$$C(T) = \Omega(T, \phi, u) \in E(T).$$

Let $T_i \in \mathfrak{M}$ be such that $T_i \to T^0$ for $i \to \infty$. Consider the sequences

$$E_i = \Omega(T_i, \phi^i, u^i) = \int_a^b K(s, T_i)\phi^i(s)\, ds + \int_a^{T_i} Z(s, T_i)u^i(s)\, ds \tag{5.2}$$

and

$$E_i^0 = \Omega(T^0, \phi^i, u^i) = \int_a^b K(s, T^0)\phi^i(s)\, ds + \int_a^{T^0} Z(s, T^0)u^i(s)\, ds. \tag{5.3}$$

Obviously $E_i \in E(T_i)$, $E_i^0 \in E(T^0)$. Clearly

$$\begin{aligned} E^i - E_i^0 &= \int_a^b \{K(s, T_i) - K(s, T^0)\}\phi^i(s)\, ds \\ &\quad + \int_a^{T^0} \{Z(s, T_i) - Z(s, T^0)\}u^i(s)\, ds + \int_{T^0}^{T_i} Z(s, T_i)u^i(s)\, ds. \end{aligned} \tag{5.4}$$

Let us denote by $\| h \|$ the norm of h. Since $Z(s, t)$ is continuous for $a \leqq s \leqq t$, there exists a positive number m_1 such that

$$\| Z(s, t) \| \leqq m_1 \tag{5.5}$$

for $a \leqq s \leqq t$, where t is sufficiently large. Therefore, for

$$T^i - T^0 < \frac{\epsilon}{3m_1},$$

we have

$$\left\| \int_a^b Z(s, T_i)u^i(s)\, ds \right\| < \frac{\epsilon}{3}. \tag{5.6}$$

From the continuity of the matrices $K(s, t)$ and $Z(s, t)$, we may write, for

$$T_i - T^0 < \frac{\epsilon}{3m_2},$$

that

$$\| K(s, T_i) - K(s, T^0) \| < \frac{\epsilon}{3h_m m_2}, \qquad (h_m = b - a), \tag{5.7}$$

and

$$\| Z(s, T_i) - Z(s, T^0) \| < \frac{\epsilon}{3n(T^0 - a)}, \tag{5.8}$$

where

$$m_2 = \operatorname*{Sup}_{\phi \in S} \| \phi \|, \tag{5.9}$$

which is a finite number by the definition of the set S.

From (5.3)–(5.9) we obtain

$$\| \Omega(T_i, \phi^i, u^i) - \Omega(T^0, \phi^i, u^i) \| < \epsilon \tag{5.10}$$

for

$$T_i - T^0 < \operatorname{Min}\left\{ \frac{\epsilon}{3m_i}, \frac{\epsilon}{3h_m m_2} \right\}. \tag{5.11}$$

Since the set $E(T^0)$ is compact and closed, we can select subsequences $\{\phi^{i_k}(t)\}$ and $\{u^{i_k}(t)\}$ of the sequences $\{\phi^i(t)\}$ and $\{u^i(t)\}$ so that they converge respectively to the functions $\phi^0(t) \in S$ and $u^0(t) \in U$. Therefore $C(T^0) = \Omega(T^0, \phi^0, u^0) \in E(T^0)$, that is $x(T^0, \phi^0, u^0) = C(T^0)$, where T^0 is the optimal control time. This proves the existence of optimal functions $\phi^0(t)$ and $u^0(t)$.

6. Properties of $\Omega(t, \phi, u)$. We shall establish, in this section, some further properties of the functional $\Omega(t, \phi, u)$.

THEOREM 7. *If $\phi(t) \in S$ and $u(t) \in U$ are in some sufficiently small neighborhood $N(\bar{\phi})$ and $N(\bar{u})$ of $\bar{\phi}(t) \in S$ and $\bar{u}(t) \in U$, then corresponding to each $\epsilon > 0$ there is a $\delta > 0$ such that*

$$\| \Omega(\bar{t}, \bar{\phi}, \bar{u}) - \Omega(t, \phi, u) \| < \epsilon \tag{6.1}$$

for each $\Omega(\bar{t}, \bar{\phi}, \bar{u})$ and all $\bar{t} - \delta < t < \bar{t}$ and all $\phi \in N(\bar{\phi})$, $u \in N(\bar{u})$.

Proof. Suppose $t < \bar{t}$ and consider

$$\begin{aligned} \Omega(\bar{t}, \bar{\phi}, \bar{u}) - \Omega(t, \phi, u) &= \int_t^{\bar{t}} Z(s, \bar{t})\bar{u}(s)\, ds \\ &+ \int_a^b \{K(s, \bar{t})\bar{\phi}(s) - K(s, t)\phi(s)\}\, ds \\ &+ \int_a^t \{Z(s, \bar{t})\bar{u}(s) - Z(s, t)u(s)\}\, ds. \end{aligned} \tag{6.2}$$

From the continuity of the matrix function $Z(s, t)$, we have

$$\left\| \int_t^{\bar{t}} Z(s, \bar{t})\bar{u}(s)\, ds \right\| < m_1(\bar{t} - t), \tag{6.3}$$

where m_1 is defined by (5.5). Using the mean value theorem, we obtain

$$K(s, t) = K(s, \bar{t}) + (t - \bar{t})K_t(s, t'), \qquad t < t' < \bar{t},$$

and

$$Z(s, t) = Z(s, \bar{t}) + (t - \bar{t})Z_t(s, t''), \qquad t < t'' < \bar{t},$$

where t' and t'' may be different for different elements of the respective matrices $K_t(s, t) = \partial K(s, t)/\partial t$ and $Z_t(s, t) = \partial Z(s, t)/\partial t$. Hence

$$K(s, \bar{t})\bar{\phi}(s) - K(s, t)\phi(s) = K(s, \bar{t})[\bar{\phi}(s) - \phi(s)] + (\bar{t} - t)K_t(s, t')\phi(s),$$

and

$$Z(s, \bar{t})\bar{u}(s) - Z(s, t)u(s) = Z(s, \bar{t})[\bar{u}(s) - u(s)] + (\bar{t} - t)Z_t(s, t'')u(s).$$

Therefore

$$\left\| \int_a^b \{K(s, \bar{t})\bar{\phi}(s) - K(s, t)\phi(s)\}\, ds \right\| < h_m[m_3 \| \bar{\phi} - \phi \| + (\bar{t} - t)m_2 m_4], \tag{6.4}$$

and

$$\left\| \int_a^t \{Z(s, t)\bar{u}(s) - Z(s, t)u(s)\}\, ds \right\| < (t - a)[m_1 \| \bar{u} - u \| + nm_5(\bar{t} - t)] \tag{6.5}$$

where

$$m_3 = \operatorname{Sup} \| K(s, \bar{t}) \|, \quad m_4 = \operatorname{Sup} \| K_t(s, \bar{t}) \|, \quad m_5 = \operatorname{Sup} \| Z_t(s, \bar{t}) \|$$

Consequently

$$\| \Omega(\bar{t}, \bar{\phi}, \bar{u}) - \Omega(t, \phi, u) \| < h_m m_3 \| \bar{\phi} - \phi \| + (t - a)m_1 \| \bar{u} - u \| + (\bar{t} - t)(h_m m_2 m_4 + nm_5).$$

Let us consider the neighborhoods $N(\bar{\phi})$ and $N(\bar{u})$ defined as follows

$$N(\bar{\phi}): \left\{ \| \bar{\phi} - \phi \| < \frac{\epsilon}{3h_m m_3}, \phi \in S \right\},$$

(6.6)

$$N(\bar{u}): \left\{ \| \bar{u} - u \| < \frac{\epsilon}{3m_1(\bar{t} - a)}, \quad u \in U \right\}.$$

In this case we have

$$\| \Omega(\bar{t}, \bar{\phi}, \bar{u}) - \Omega(t, \phi, u) \| < \epsilon$$

for $\bar{t} - \delta < t < \bar{t}$, and for all $\phi \in N(\bar{\phi})$, $u \in N(\bar{u})$, where

$$\delta = \frac{\epsilon}{3(h_m m_2 m_4 + n m_5)}. \tag{6.7}$$

THEOREM 8.[1] *If Ω is an interior point of $E(t)$ then there exists an $\epsilon > 0$ such that $N_\epsilon(\Omega) \subset E(\tau)$ for all τ in $(t - \epsilon, t]$.*

Proof. Let Ω be in the interior of $E(t)$ and let $N_{\epsilon_1}(\Omega)$ be a neighborhood of Ω of radius $\epsilon_1 > 0$ contained in the interior of $E(t)$. Let $\epsilon_2 = \frac{1}{2}\epsilon_1$ and let $N_{\epsilon_2}(\Omega)$ be the neighborhood of Ω of radius ϵ_2. Let $\delta = \delta(\epsilon_2) > 0$ be chosen to satisfy theorem 7, i.e. (6.7) with $\epsilon = \epsilon_2$. Consider $\Omega^* \in N_{\epsilon_2}(\Omega)$. Suppose for some t_1 satisfying $t - \delta < t_1 < t$ that Ω^* is not in the interior of $E(t_1)$. Then since $E(t_1)$ is convex, there exists a support plane P_{t_1} such that there are no points of $E(t_1)$ on one side of P_{t_2}. Because the neighborhood $N_{\epsilon_2}(\Omega^*) \subset N_{\epsilon_1}(\Omega) \subset E(t)$ we see that there is a point $p \in E(t)$ such that $\| p - E(t_1) \| \geqq \epsilon_2$. But this contradicts theorem 7. Therefore $N_{\epsilon_2}(\Omega) \subset E(\tau)$ for all $\tau \in (t - \delta, t]$. Now set $\epsilon = \text{Min}\,(\epsilon_2, \delta)$. Then $\epsilon > 0$ and $N_\epsilon(\Omega) \subset N_{\epsilon_2}(\Omega)$ and $(t - \epsilon, t] \subset (t - \delta, t]$. Thus $N_\epsilon(\Omega) \subset E(\tau)$ for all $\tau \in (t - \epsilon, t]$.

THEOREM 9. *Let $C(t)$ be given by (4.4). Then $C(T^0)$ is a boundary point of $E(T^0)$, where T^0 is the optimal time.*

Proof. Suppose, on the contrary, that $C(T^0) = \Omega(T^0, \phi^0, u^0)$ is an interior point of $E(T^0)$. Then from the theorem 8 there exists an $\epsilon > 0$ such that $N_\epsilon(C(T^0)) \subset E(t)$ for all $T^0 - \epsilon < t < T^0$. The continuity of $C(t)$ at T implies that there exists a $\delta > 0$ such that $C(t) \in N_\epsilon(C(T^0))$ for all $T^0 - \delta < t < T^0$. Let $2\gamma = \text{Min}\,(\delta, \epsilon)$. Then $C(T^0 - \gamma) \in N(C(T^0)) \subset E(T^0 - \gamma)$. But this is a contradiction since $T^0 - \gamma < T^0$ and T^0 is, by hypothesis, the minimum value of t such that $C(t) \in E(t)$.

THEOREM 10. *There exists a vector $q^0 = (q_1^0, \cdots, q_n^0)$ of the n-dimensional Euclidean space, such that*

$$q^0 \cdot \Omega(T^0, \phi, u) \leqq q^0 \cdot \Omega(T^0, \phi^0, u^0) \tag{6.10}$$

for all $\Omega(T^0, \phi, u)$.

Proof. Since $E(T^0)$ is a convex set and since $\Omega(T^0, \phi^0, u^0)$ is, by Theorem 10, a boundary point of the set $E(T^0)$, there is a support plane through $\Omega^0 = \Omega(T^0, \phi^0, u^0)$. Let q^0 be a non-zero vector orthogonal to this support plane and directed to the side which contains no points of the set $E(T^0)$. Therefore, for each $\Omega(T^0, \phi, u) \in E(T^0)$ and for all $\phi \in S$, $u \in U$, we have

$$(q^0, \Omega(T^0, \phi, u) - \Omega(T^0, \phi^0, u^0)) \leqq 0.$$

This inequality is equivalent to (6.10).

[1] The author is particularly indebted to the referee for theorems 8 and 9.

7. Optimal control functions. As in the case which LaSalle considered we can prove the following theorem for the general form of the optimal control function:

THEOREM 11. *All optimal control functions* $u^0(t)$ *are of the form*

$$u^0(t) = \operatorname{sgn}[q^0 Y(t, T^0)B(t)], \qquad t_0 \leqq t \leqq T^0, \tag{7.1}$$

where q^0 *is some non-zero n-dimensional row vector depending on* $\phi^0(t)$ *and* T^0 *is the optimal time.*

The proof of this theorem is very similar to LaSalle's proof for the system (1.3).

As LaSalle proved, if the system (2.1) is normal, that is, if no components of $[q^0 Y(t, T^0)B(t)]$ vanishes on any interval, no matter what the vector $q \neq 0$, the optimal control function is bang-bang, unique and given by (7.1).

The signum in the formula (7.1) is taken by components: i.e., if b is an r-vector, then $a = \operatorname{sgn} b$, means that $a_j = 1$ when $b_j > 0$ and $a_j = -1$ when $b_j < 0$ and a_j is indeterminate for $b_j = 0$.

8. Optimal initial functions. Let us consider the equation (2.10) with the optimal functions:

$$x(T^0, \phi^0, u^0) = z(T^0), \tag{8.1}$$

where we know the existence of T^0, ϕ^0 and u^0 and also the general form of the optimal steering function $u^0(t)$ by the formula (7.1). We write this equation in the following form:

$$\int_a^b K(s, T^0)\phi^0(s)\, ds = G(T^0), \tag{8.2}$$

where

$$G(T^0) = z(T^0 + h_m) - \int_a^{T^0} Z(s, T^0)u^0(s)\, ds - \int_a^{T^0} y(s, T^0)f(s)\, ds \tag{8.3}$$

We wish to establish the optimal initial functions $\phi^0(t)$. As all the quantities in $G(T^0)$ are known, except the vector q, (8.2) is a Fredholm integral equation of the first kind. Writing t instead of T^0, $\phi(t)$ instead of $\phi^0(t)$ and putting

$$K\phi = \int_a^b K(s, t)\phi(s)\, ds, \tag{8.4}$$

we obtain

$$K\phi = G(t). \tag{8.5}$$

Let $K^*(s, t)$ be the transposed matrix of $K(s, t)$. Put

$$K\psi = \int_a^b K(s, t)\psi(t)\, dt, \tag{8.6}$$

and

$$(\phi, \psi) = \int_a^b \phi(s)\psi(s)\, ds. \tag{8.7}$$

It is well known that

$$(K\phi, \psi) = (\phi, K^*\psi). \tag{8.8}$$

Let $\{\lambda_k\}$ be the set of eigenvalues of the following homogeneous Fredholm integral equations of the second kind:

$$K\phi = \lambda\phi, \qquad K^*\psi = \lambda\psi. \tag{8.9}$$

Let $\{\phi_k(t)\}$ and $\{\psi_k(t)\}$ be the complete sets of the principal eigenfunctions of the kernels $K(s, t)$ and $K^*(s, t)$, which are biorthonormalized:

$$(\phi_j\,, \psi_k) = \delta_{jk} = \begin{cases} 0, j \neq k, \\ 1, j = k, \end{cases} \tag{8.10}$$

where

$$K\phi_k = \lambda_k\phi_k\,, \qquad K^*\psi_k = \lambda_k\psi_k\,. \tag{8.11}$$

Let $\{\phi_{0i}(t)\}$ and $\{\psi_{0i}(t)\}$ be the non-trivial solutions of the integral equations

$$K\phi = 0, \qquad K^*\psi = 0. \tag{8.12}$$

Assume that the solvability condition

$$(\psi_{0i}\,, G) = 0 \tag{8.13}$$

is satisfied for each $\psi_{0i}(t)$. Consider now the expansion of the function $G(t)$ with respect to the biorthonormal system $\{\phi_k\,, \psi_k\}$:

$$G(t) = \sum_{k=1}^{\infty} c_k\phi_k(t), \tag{8.14}$$

where

$$c_k = (\psi_k\,, G). \tag{8.15}$$

Let us now determine the constants a_k, such that the series

$$\phi(t) = \sum_{k=1}^{\infty} a_k\phi_k(t) + \phi_0(t), \tag{8.16}$$

where $\phi_0(t)$ is a solution of the integral equation $K\phi = 0$, should be a solution of the integral equation (8.5). For this purpose, multiply both sides of (8.5) by $\psi_k(t)$ and integrate with respect to t from a to b; hence, making use of the preceding formulas, we obtain

$$a_k = \frac{c_k}{\lambda_k}, \qquad k = 1, 2, \cdots\cdots. \tag{8.17}$$

Therefore, the optimal initial function is given by the formula

$$\phi^0(t) = \sum_{k=1}^{\infty} \frac{c_k}{\lambda_k} \phi_k(t) + \phi_0(t). \tag{8.18}$$

We note that, the above deductions base upon the continuity of the kernel function $K(s, t)$.

We summarize this result as follows:

THEOREM 12. *Let* $\{\lambda_k\}$ *be the set of eigenvalues of the integral equations* (8.9) *and* $\{\phi_k(t)\}$, $\{\psi_k(t)\}$ *be the complete sets of eigenfunctions of the kernels* $K(s, t)$ *and its transpose* $K^*(s, t)$. *Let* $\{\phi_{0i}(t)\}$ *and* $\{\psi_{0i}(t)\}$ *be the sets of non-trivial solutions of the integral equations* (8.12). *If the solvability condition* (8.13) *is satisfied for each* $\psi_{0i}(t)$, *then the optimal initial condition* $\phi^0(t)$ *is given by the formula* (8.18), *where the coefficients* c_k *are defined by* (8.15) *and* $\phi_0(t)$ *is a linear combination of the functions* $\phi_{0i}(t)$.

9. A necessary condition for optimality. We have established in section 6, Theorem 10, the existence of a vector q^0 such that $q^0 \cdot C(T^0)$ maximizes the function $q^0 \cdot \Omega$ for $\Omega \in E(T^0)$.

Consider now the vector function $u_q(t)$ defined by

$$u_q(t) = \operatorname{sgn} [qY(t, T)B(t)], \qquad t_0 \leqq t \leqq T,$$

where q is an n-dimensional unit vector: $\| q \| = 1$. Obviously, more than one q may determine the same $u_q(t)$. Let $\phi_q(t)$ be the initial condition corresponding to the control function $u_q(t)$, obtained by a similar method to that of section 8, using $u_q(t)$ and T instead of $u^0(t)$ and T^0 respectively. It is clear that $u_{q^0}(t) = u^0(t)$ and $\phi_{q^0}(t) = \phi^0(t)$.

By the definitions of $u_q(t)$ and $\phi_q(t)$ and the unique maximum condition, we have

$$q \cdot \Omega(T, \phi, u) \leqq q \cdot \Omega(T, \phi_q, u_q)$$

for all $\Omega(T, \phi, u) \in E(T)$, $\Omega(T, \phi, u) \neq \Omega(T, \phi_q, u_q)$.

Let us consider the vectors q for which

$$q[\Omega(b, \phi, u) - C(b)] < 0, \tag{9.1}$$

and define the function

$$V(t, q) = q \cdot L(t, q), \tag{9.2}$$

where

$$L(t, q) = \Omega(t, \phi_q, u_q) - C(t), \tag{9.3}$$

ϕ_q and u_q being the extremal functions defined above.

Let H_0 be the convex set of vectors q which have the property $q\cdot\Omega \leqq q\cdot C(T^0)$ for $\Omega \in E(T^0)$. Clearly, for $q \in H_0$, $V(T^0, q) = 0$.

Suppose now that $V(t, q)$ is strictly increasing at $t = T^0$ for every $q \in H_0$. If $q \notin H_0$, $V(T^0, q) > 0$ regarding to the definition of the set H_0. On the other hand, we have $V(b, q) < 0$ by (9.1). Therefore, there exists some unique $T(q)$ such that

$$V(T(q), q) = 0, \tag{9.4}$$

for t in a neighborhood of T^0 and q in a neighborhood of H_0. If $q \in H_0$, $T(q) = T^0$, and if $q \notin H_0$, $T(q) < T^0$. Therefore, we obtain the following extension of a local maximum principle due to Neustadt [19] which gives a necessary condition for the optimality for our control problem:

THEOREM 13. *Let T^0 be the maximum time at which any $x(t, \phi, u)$ given by a normal system* (2.1) *and satisfying the initial condition* (2.2), *can reach the particle $z(t)$; if for every $q \in H_0$ the function $V(t, q)$ given by* (9.2) *is strictly increasing with t at $t = T^0$, then for q in a neighborhood of H_0, and t in a neighborhood of T^0, the vectors $q \in H_0$ maximize the time for which $q\cdot\Omega(t, \phi^0, u^0) = q\cdot C(t)$.*

This theorem, which is very close to Pontryagin's maximum principle ([8], [15]), reduces the optimal time problem to finding the maximum of the function $T(q)$. To find this maximum, we shall show first that the function $t = T(q)$ possesses continuous partial derivatives. For this purpose, since $t = T(q)$ is defined implicitly by the equation $V(t, q) = 0$, we shall show that $\partial V/\partial t$ and $\partial V/\partial q_i$, $i = 1, 2, \cdots, n$, exist and are continuous, and if $\partial V/\partial t \neq 0$:

$$\frac{\partial T}{\partial q_i} = -\frac{\partial V}{\partial q_i}\Big/\frac{\partial V}{\partial t} \qquad (i = 1, \cdots, n). \tag{9.5}$$

Put

$$z(t, q) = \int_a^t \{q\cdot Z(s, t)\}\cdot\{\operatorname{sgn} q\cdot Z(s, t)\}\, ds. \tag{9.6}$$

Denote again the components of $Z(s, t) = Y(s, t)B(s)$ by $z_{ij}(s, t)$ and the components of q by q_k ($i = 1, \cdots, n; j = 1, \cdots, r; k = 1, \cdots, n$). Hence

$$z(t, q) = z(t, q_1, \cdots, q_n) = \sum_{j=1}^{r} \int_a^t \left| \sum_{k=1}^{k} q_k z_{kj}(s, t) \right| ds. \tag{9.7}$$

As Neustadt showed [19], we have

$$\frac{\partial z}{\partial q_i} = e_i^*\cdot\int_a^t Z(s, t) \operatorname{sgn} [q\cdot Z(s, t)]\, ds, \tag{9.8}$$

where e_i^* is a row-vector which is defined by $e_i^* = (\delta_{i1}, \cdots, \delta_{in})$, δ_{ij}

being the Kronecker's symbol:

$$\delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

On the other hand we have

$$\frac{\partial}{\partial q_i}\left\{ \int_a^b q \cdot K(s,t)\phi^0(s)\, ds \right\} = e_i^* \cdot \int_a^b K(s,t)\phi^0(s)\, ds. \tag{9.9}$$

Therefore

$$\frac{\partial V}{\partial q_i} = e_i^* \cdot L(t,q). \tag{9.10}$$

Similarly, we have

$$\frac{\partial}{\partial t}\left\{ \int_a^b q \cdot K(s,t)\phi^0(s)\, ds \right\} = q \cdot \int_a^b \frac{\partial K(s,t)}{\partial t}\, \phi^0(s)\, ds. \tag{9.11}$$

Let us now calculate $\partial z/\partial t$; we put

$$\sigma_j(s,t,q) = \sum_{k=1}^{n} q_k z_{kj}(s,t), S_j(t,q) = \int_a^t |\, \sigma_j(s,t,q)\,|\, dt \tag{9.12}$$

Clearly

$$z(t,q) = \sum_{j=1}^{r} S_j(t,q). \tag{9.13}$$

Let

$$\begin{aligned} \Delta S_j &= S_j(t+\Delta t, q) - S_j(t,q) \\ &= \int_a^{t+\Delta t} |\, \sigma_j(s, t+\Delta t, q)\,|\, ds - \int_a^t |\, \sigma_j(s,t,q)\,|\, ds \\ &= \int_a^t \{|\, \sigma_j(s, t+\Delta t, q)\,| - |\, \sigma_j(s,t,q)\,|\}\, ds \\ &\quad + \int_t^{t+\Delta t} |\, \sigma_j(s, t+\Delta t, q)\,|\, ds \\ &= \Delta t\, |\, \sigma_j(t+\theta\Delta t, t+\Delta t, q)\,| \\ &\quad + \int_a^t \left\{ \left| \sigma_j(s,t,q) + \frac{\partial \sigma_j(s,t,q)}{\partial t}\Delta t \right| - |\, \sigma_j(s,t,q)\,| \right\} ds, \end{aligned} \tag{9.14}$$

where $0 < \theta < 1$. Since our system is normal the set $\{s : \sigma_j(s,t,q) = 0, s < t\}$ has measure zero for every q. Hence for any $\epsilon > 0$, there is a positive δ such that the set $A_0 = \{s : |\, \sigma_j(s,t,q)\,| < \delta, s < t\}$ has measure less than ϵ. Let

$A_+ = \{s : \sigma_j(s, t, q) \geqq \delta,\ s < t\}$, and $A_- = \{s : \sigma_j(s, t, q) \leqq -\delta,\ s < t\}$. Since $\sigma_j(s, t, q)$ is continuous and continuously differentiable,

$$M = \operatorname*{Max}_{a \leqq s \leqq t} \left\{ |\, \sigma_j(s, t, q)\,| \,, \left| \frac{\partial \sigma_j(s, t, q)}{\partial t} \right| \right\} < \infty. \tag{9.15}$$

Choose Δt such that $|\,\Delta t\,| < \delta/M$. Then

$$\begin{aligned} \Delta S_j &= \Delta t\,|\,\sigma_j(t + \theta\Delta t, t + \Delta t, q)\,| \\ &\quad + \Delta t \int_{A_+ \cup A_-} \frac{\partial \sigma_j(s, t, q)}{\partial t} \operatorname{sgn}[\sigma_j(s, t, q)]\, ds \\ &\quad + \int_{A_0} \left\{ \left| \sigma_j(s, t, q) + \frac{\partial \sigma_j(s, t, q,)}{\partial t} \Delta t \right| - |\,\sigma_j(s, t, q)\,| \right\} ds. \end{aligned} \tag{9.16}$$

Since the last integral is less, in absolute value, than $|\,\Delta t\,| M\epsilon$, we have

$$\left| \frac{\Delta S_j}{\Delta t} - |\,\sigma_j(t + \theta\Delta t, t + \Delta t, q)\,| - \int_{A_+ \cup A_-} \frac{\partial \sigma_j(s, t, q)}{\partial t} \operatorname{sgn}[\sigma_j(s, t, q)]\, ds \right| < 2M\epsilon,$$

therefore

$$\frac{\partial S_j}{\partial t} = |\,\sigma_j(t, t, q)\,| + \int_a^t \frac{\partial \sigma_j(s, t, q)}{\partial t} \operatorname{sgn}[\sigma_j(s, t, q)]\, ds \tag{9.17}$$

and, since $Z(s, t) = Y(s, t)B(s)$ and $Y(t, t) = I$ (identity matrix), we obtain

$$\begin{aligned} \frac{\partial z(t, q)}{\partial t} = \sum_{j=1}^{r} \frac{\partial S_j}{\partial t} &= q \cdot B(t) \operatorname{sgn}[q \cdot B(t)] \\ &\quad + \int_a^t \frac{\partial Z(s, t)}{\partial t} \operatorname{sgn}[q \cdot Z(s, t)]\, ds. \end{aligned} \tag{9.18}$$

Consequently

$$\begin{aligned} \frac{\partial V}{\partial t} &= \{q \cdot B(t)\} \cdot \{\operatorname{sgn}[q \cdot B(t)]\} - q \cdot C'(t) \\ &\quad + \int_a^b q \cdot \frac{\partial K(s, t)}{\partial t} \phi^0(s)\, ds + \int_a^t q \cdot \frac{\partial Z(s, t)}{\partial t} \operatorname{sgn}[q \cdot Z(s, t)]\, ds \\ &= P(t, q). \end{aligned} \tag{9.19}$$

If $P(T(q), q) = Q(q) \neq 0$, from (9.10) and (9.19), we obtain

$$\frac{\partial T}{\partial q_i} = -\frac{e_i^* \cdot L(t, q)}{Q(q)}, \tag{9.20}$$

and

$$\nabla T = -\frac{L(T(q)\cdot q)}{Q(q)}. \tag{9.21}$$

Clearly the vector function $L(t, q)$ is continuous in both t and q. We suppose that $Q(q) > 0$.

Let r be a parameter and consider the differential equation

$$\frac{dq}{dr} = Q(q)\cdot\nabla T = -L(T(q), q). \tag{9.22}$$

Since $L(T(q), q)$ is continuous, this equation has a solution satisfying a given initial condition. If $Q(q) \neq 0$, $dT(q)/dr$ exists and

$$\frac{dT(q)}{dr} = \nabla T\cdot\frac{dq}{dr} = Q(q)\cdot(\nabla T\cdot\nabla T) > 0. \tag{9.23}$$

If $q \notin H_0$, $\nabla T \neq 0$ and therefore $dT(q)/dr > 0$.

If $Q(q) = 0$, $dT(q)/dr$ does not exist. But, by the formula (9.10), ∇V, for fixed t, is defined, and

$$\frac{\partial V(t, q(t))}{\partial r} = -\left(\frac{dq}{dr}\right)^2 < 0 \qquad \text{for} \quad q \notin H_0. \tag{9.24}$$

Hence, $V(t, q(r))$ is monotonically decreasing with r, or $T(q(r))$ is increasing with r, so that for all $q \notin H_0$, $T(q(r))$ is increasing with r.

Let us denote by $\mathfrak{D}$ the domain of the function $T(q)$. This function is defined for all q for which the inequality (9.1) is satisfied. Consider the solution $q(r)$ of the differential equation (9.22) which satisfies the initial condition $q(r_0) = q_0$, where $q_0 \in \mathfrak{D}$. The function $q(r) \in \mathfrak{D}$ for all values of $r > r_0$. In fact, if $q(r)$ leaves $\mathfrak{D}$, then for some $r' > r_0$,

$$q(r')[\Omega(b, \phi, u) - C(b)] = 0.$$

Hence

$$V(T(q(r')), q(r')) = q(r')\cdot[\Omega(T(q(r')), \phi, u) - C(T(q(r'))] = 0.$$

Therefore, by (9.4), $T(q(r')) = b$. But this is impossible, because $T(q(r')) > T(q(r_0)) \geqq b$ if $r' > r_0$.

It can be easily seen that the norm $\| q \|$ of the solution of (9.22) is constant.

Let $q(r)$ be the solution of (9.22) such that $q(r_0) = q_0 \in \mathfrak{D}$. If $q(r)$ approaches a limit as $r \to \infty$, then this limit is in H_0. Thus $T(q) = b$ on the boundary of $\mathfrak{D}$ and $T(q)$ attains its maximum.

10. Construction of the kernel $K(s, t)$. The kernel matrix function

$K(s, t)$ is defined, in Section 3, as the solution of differential-difference equation

$$\frac{\partial K(s, t + h_m)}{\partial t} + \sum_{k=0}^{m} A_k(t) K(s, t + h_k) = 0, \tag{10.1}$$

which satisfies for all $\phi(t) \in C$, the condition

$$\int_a^b K(s, t)\phi(s)\, ds = \phi(t), \qquad a \leqq t \leqq b, \tag{10.2}$$

where $a = b_0$, $b = t_0 + h_m$.

Let $\delta(s - t)$ be the Dirac's delta-function:

$$\delta(s - t) = \begin{cases} 0, & s \neq t, \\ +\infty, & s = t. \end{cases} \tag{10.3}$$

It is well known (Cf. [21]) that

$$\int_a^b \delta(s - t)\phi(s)\, ds = \phi(t), \qquad a \leqq t \leqq b. \tag{10.4}$$

Obviously the solution of the differential-difference equation (10.1) which satisfies the initial condition

$$K(s, t) = \delta(s - t)I, \qquad a \leqq s, t \leqq b, \tag{10.5}$$

where I is the identity matrix, verifies the required condition (10.2). We may establish this solution. For this purpose let us consider the solution $K_\epsilon(s, t)$ of (10.1) satisfying the initial condition

$$K(s, t) = \delta_\epsilon(s - t)I, \qquad a \leqq s, t \leqq b, \tag{10.6}$$

where $\delta_\epsilon(s - t)$ is defined by

$$\delta_\epsilon(s - t) = \frac{\epsilon}{\pi[(s - t)^2 + \epsilon^2]}, \tag{10.7}$$

which gives a continuous approximation of the delta-function:

$$\lim_{\epsilon \to 0} \delta_\epsilon(s - t) = \delta(s - t). \tag{10.8}$$

$K_\epsilon(s, t)$ is uniquely determined for $t \geqq b$ (Cf. for example [12]), and, is continuously dependent on ϵ, and it possesses the following properties:

$$\lim_{\epsilon \to 0} K_\epsilon(s, t) = K(s, t), \tag{10.9}$$

for $a \leqq s \leqq b$, $t \geqq a$, and

$$\lim_{\epsilon \to 0} K_\epsilon(s, t) = \lim_{\epsilon \to 0} \delta_\epsilon(s - t)I = \delta(s - t)I \tag{10.10}$$

for $a \leqq s \leqq b$, $a \leqq t \leqq b$. Therefore, to construct the kernel matrix function $K(s, t)$ it suffices to establish the approximating kernel $K_\epsilon(s, t)$ and pass to the limite $\epsilon = 0$.

11. Acknowledgment. The author wishes to express his appreciation to Dr. P. Seibert and Prof. W. Hahn for their helpful comments and to the referee for his very valuable remarks.

REFERENCES

[1] J. André and P. Seibert, *The local theory of piecewise continuous differential equations*, Contributions To the Theory of Nonlinear Oscillations, Vol. V, Princeton Press (1960) pp. 225–255.

[2] L. Alaoğlu, *Weak topologies of normed spaces*, Ann. of Math., 41 (1940) pp. 252–267.

[3] R. Bellman, *Terminal Control, time lags, and dynamic programming*, Proc. Nat. Acad. Sci. U.S.A. 43 (1957) pp. 927–930.

[4] R. Bellman, and K. L. Cookke, *Stability theory and adjoint operators for linear differential-difference equations*, Trans. Amer. Math. Soc., 92 (1959) pp. 470–500.

[5] R. Bellman, I. Glicksberg, and O. Gross, *On some variational problems occurring in the theory of dynamic programming*, Rend. Circ. Mat. Palermo, 3 (1954) pp. 367–397.

[6] R. Bellman, I. Glicksberg, and O. Gross, *On the "bang-bang" control problem*, Quart. Appl. Math., 14 (1956), pp. 11–18.

[7] R. Bellman, and R. Kalaba, *Reduction of dimensionality, dynamic programming, and control processes*, J. Basic Engrg. (Trans. A.S.M.E.), 83, (1961), 82–84.

[8] V. G. Boltyanskii, R. V. Gamkrelidze, E. F. Mishcenko, and L. S. Pontryagin, *The maximum principle in the theory of optimal processes of control*, Automatic and Remote Control, (1960) pp. 1004–1008.

[9] D. Bushaw, *Optimal discontinuous forcing terms*, Contributions To The Theory of Nonlinear Oscillations, Vol. IV, Princeton Press (1958), 29–52.

[10] H. G. Eggleston, *Convexity*, Cambridge Tracts in Math. and Math. Phys., 47 (1958) Cambridge.

[11] A. F. Filippov, *Differential equations with a discontinuous right-hand side*, Mat. Sb., 51(93), (1960), 99–128.

[12] J. Franklin, *On the existence of solutions of systems of functional differential equations*, Proc. Amer. Math. Soc., 5, pp. 363–369.

[13] M. Frechet, *Généralisation d'un théorème de Weierstrass*, Comptes rendus Acad. Sci. Paris, 139, (1907) p. 848.

[14] P. R. Halmos, *The range of a vector measure*, Bull. Amer. Math. Soc., 54 (1958) pp. 416–421.

[15] G. L. Haratisvili, *The maximum principle in the theory of optimal processes involving delay*, Dokl. Akad. Nauk SSSR., 136 (1961) pp. 39–42 (Russian); translated as Soviet Math. Doklady, 2, pp. 28–32.

[16] N. N. Krasovskii, *Concerning the theory of optimal control*, Avtomatika i Telemeh., 18 (1957) pp. 960–970.

[17] J. P. LaSalle, *The time optimal control problem*, Contributions to the theory of nonlinear oscillations, Vol. V, Princeton Press (1960) pp. 1–24.

[18] A. LIAPUNOV, *Sur les fonctions-vecteurs completement additives*, Izv. Akad. Nauk SSSR., Ser. Mat. 4 (1940) pp. 465–478.

[19] L. NEUSTADT, *Synthesizing time optimal control systems*, J. Math. Anal. Appl. 1 (1960) pp. 484–493.

[20] M. H. STONE, *Linear transformations in Hilbert spaces and their applications to analysis*, Amer. Math. Soc. Collq. Publ., Vol. XV, 1932.

[21] B. VAN DER POL, AND H. BREMMER, *Operational Calculus Based On The Two-Sided Laplace Integral*, Cambridge, University Press, 1959, Chap. V.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

STABILITY CRITERIA FOR NONLINEAR ORDINARY DIFFERENTIAL EQUATIONS*

O. L. MANGASARIAN†

Abstract. The main results of this work are three sufficient conditions for the (1) stability, (2) uniform asymptotic stability in the large and (3) instability, of the equilibrium point $x = 0$ of the system of differential equations: $\dot{x} = f(t, x), f(t, 0) = 0$. Stated roughly these conditions are: The point $x = 0$ is (1) stable if $x'f(t, x)$ is a concave function of x, (2) uniformly asymptotically stable in the large if $x'f(t, x)$ is a strictly concave function of x, and (3) unstable if $x'f(t, x)$ is a strictly convex function of x. These results are obtained by using the stability and instability criteria of Liapunov and properties of concave and convex functions.

1. Introduction. We shall be concerned with the system of nonlinear ordinary differential equations

$$\dot{x} = f(t, x) \tag{1.1}$$

where x and f are n-dimensional vectors and $0 \leqq t < \infty$. We shall assume that $f(t, x)$ is piecewise continuous in the (x, t) space, the discontinuities lying on sufficiently smooth manifolds, and that to any given (x_0, t_0) there corresponds at least one function of t, $x = y(t, x_0, t_0)$, defined, continuous and with piecewise continuous derivative with respect to t for all $t \geqq t_0$, which satisfies the system (1.1) except at the points of discontinuity. According to Massera [9], under these conditions, the stability criteria of Liapunov and various modifications thereof hold.

We shall assume that $f(t, 0) = 0$ for all t, so that $x = 0$ is an equilibrium point of the system (1.1). We will be concerned with the stability or instability of this equilibrium point.

Our main results are given in Theorems 1, 2 and 3 in Section 3. These theorems give sufficient conditions for the stability, uniform asymptotic stability in the large, and instability of the equilibrium point $x = 0$. These results are based essentially on the concavity, strict concavity, and strict convexity, respectively, of the scalar function $x'f(t, x)$ with respect to x.

We use the commonly accepted definition [1, p. 18], [3] that $\varphi(x)$ is convex if, for $0 \leqq \lambda \leqq 1$.

$$(1 - \lambda)\varphi(x^1) + \lambda\varphi(x^2) \geqq \varphi[(1 - \lambda)x^1 + \lambda x^2] \tag{1.2}$$

for all vectors x^1 and x^2 in the convex region of definition of $\varphi(x)$. The function $\varphi(x)$ is concave if the inequality sign in (1.2) is reversed. For strictly convex (concave) functions the equality sign in (1.2) holds only for $\lambda = 0$,

* Received by the editors September 9, 1962 and in revised form March 24, 1963.
† Shell Development Company, Emeryville, California.

$\lambda = 1$ or $x^1 = x^2$. Note that convexity and concavity imply continuity in the interior of the convex region of definition but not necessarily differentiability. However if $\varphi(x)$ is twice continuously differentiable then sufficient conditions for convexity, concavity, strict convexity and strict concavity of $\varphi(x)$ are respectively that the symmetric matrices of second partial derivatives $\partial^2\varphi/\partial x_i \partial x_j$ be positive semidefinite, negative semidefinite, positive definite and negative definite for all values of x in the region of definition of φ [1, p. 18], [3].

Among previously obtained criteria for asymptotic stability are those of Krasovskii [7] and Hartman [4, Lemma 1′]. Both of these results are restricted to autonomous systems (that is $f = f(x)$) and require that f have continuous first partial derivatives. Our results for stability (Theorem 1) and uniform asymptotic stability in the large (Theorem 2) are not only valid for *nonautonomous* systems but do not require differentiability of f.

Prior to proving the stability theorems of Section 3, we establish certain properties of convex and concave functions. We do this in Section 2 where we also state certain stability theorems of Liapunov and Massera that are needed subsequently.

Vector notation is used throughout. In general, Latin capitals denote matrices, small Latin letters denote column vectors, and small Greek letters denote scalars. Exceptionally, $V(t, x)$ will denote a scalar function of the scalar t and vector x. A prime $'$ will denote the transpose. The Euclidean norm $(x'x)^{1/2}$ of a vector x will be denoted by $\| x \|$.

The scalar function $V(t, x)$ is *positive definite*[1] if for $0 \leqq t < \infty$, $V(t, x) > 0$ for $x \neq 0$, $V(t, 0) = 0$, and $\liminf_{t\to\infty} V(t, x) > 0$ for $x \neq 0$. The function $V(t, x)$ is *negative definite*[1] if for $0 \leqq t < \infty$, $V(t, x) < 0$ for $x \neq 0$, $V(t, 0) = 0$ and $\limsup_{t\to\infty} V(t, x) < 0$ for $x \neq 0$.[2]

The function $V(t, x)$ is said to have an *infinitely small upper bound* if, given $\epsilon > 0$, there exists a $\delta > 0$ such that $| V | < \epsilon$ for all $t \geqq 0$ and $\| x \| < \delta$. When $V(t, x)$ is of class C^1 in t and x, we have along solutions of (1.1)

$$\dot{V}(t, x) = \dot{x}'\nabla V + \frac{\partial V}{\partial t} = f'\nabla V + \frac{\partial V}{\partial t},$$

where

$$\nabla V \equiv \left[\frac{\partial V}{\partial x_1} \quad \frac{\partial V}{\partial x_2} \cdots \frac{\partial V}{\partial x_n}\right]'.$$

[1] These definitions of positive and negative definite functions differ somewhat from the conventional ones [2], [8], [9] which are implied if we assume in addition that $V(t, x)$ is continuous and has continuous first partial derivatives.

[2] $\liminf_{t\to\infty}$ and $\limsup_{t\to\infty}$ denote respectively the limit inferior and limit superior of $V(t, x)$ as t tends to infinity.

2. Preliminary results. We start by establishing some lemmas which are used in proving the main results of Section 3. In what follows t is a scalar and f and x are n-by-1 vectors.

LEMMA 1. *Let $f(t, x)$ be continuous in x at $x = 0$ for $0 \leqq t < \infty$ and let $f(t, 0) = 0$ for $0 \leqq t < \infty$. If $x'f(t, x)$ is a concave function of x for $0 \leqq t < \infty$, then $x'f(t, x) \leqq 0$ for $0 \leqq t < \infty$.*

Proof. Let $\varphi(t, x) \equiv x'f(t, x)$ and $x^0 \equiv 0$. By the assumed concavity of $\varphi(t, x)$ we have for $0 \leqq t < \infty$ and $0 < \lambda \leqq 1$

$$(1 - \lambda)\varphi(t, x^0) + \lambda\varphi(t, x) \leqq \varphi[t, (1 - \lambda)x^0 + \lambda x]$$

or

$$\varphi(t, x) \leqq \frac{-(1 - \lambda)\varphi(t, x^0) + \varphi[t, x^0 + \lambda(x - x^0)]}{\lambda}.$$

Now since $x^0 = 0$ and $\varphi(t, x^0) = x^{0\prime}f(t, x^0) = 0$, we have for $0 < \lambda \leqq 1$ and $0 \leqq t < \infty$

$$\varphi(t, x) \leqq \frac{\varphi(t, \lambda x)}{\lambda} = x'f(t, \lambda x).$$

Hence for $0 \leqq t < \infty$

$$x'f(t, x) \equiv \varphi(t, x) \leqq \lim_{\lambda \to 0} x'f(t, \lambda x) = x'\lim_{\lambda \to 0} f(t, \lambda x).$$

But since $f(t, x)$ is continuous in x at $x = 0$ for $0 \leqq t < \infty$ it follows that $\lim_{\lambda \to 0} f(t, \lambda x) = 0$ and thus $x'f(t, x) \leqq 0$ for $0 \leqq t < \infty$.

Lemma 1 is used to prove Theorem 1. The following lemma is needed in the proof of Theorem 2.

LEMMA 2.[3] *Let $f(t, x)$ be continuous in x at $x = 0$ for $0 \leqq t < \infty$, let $f(t, 0) = 0$ for $0 \leqq t < \infty$ and let $\psi(x) \equiv \limsup_{t \to \infty} x'f(t, x)$. If $x'f(t, x)$ is a strictly concave function of x for $0 \leqq t < \infty$ and if either (I) $\psi(x) < 0$ for $x \neq 0$, or (II) $\psi(x)$ is strictly concave in x, then $x'f(t, x)$ is negative definite.*

Proof. Let $\varphi(t, x) \equiv x'f(t, x)$. By the strict concavity of $\varphi(t, x)$ we have for $0 \leqq t < \infty$, $0 < \lambda < 1$ and $x \neq 0$

$$(1 - \lambda)\varphi(t, 0) + \lambda\varphi(t, x) < \varphi(t, \lambda x),$$

and since $\varphi(t, 0) = 0$,

$$\varphi(t, x) < \frac{\varphi(t, \lambda x)}{\lambda} \quad \text{for} \quad 0 \leqq t < \infty, 0 < \lambda < 1, x \neq 0. \tag{2.1}$$

[3] I am indebted to the referee for the removal of a redundant hypothesis from this lemma.

Now by Lemma 1, $\varphi(t, x) \equiv x'f(t, x) \leqq 0$ for $0 \leqq t < \infty$. However, if $\varphi(t, x) = 0$ for some t in $0 \leqq t < \infty$ and some $x \neq 0$ then it follows from (2.1) that $\varphi(t, \lambda x) > 0$ for $0 < \lambda < 1$ and $x \neq 0$, which contradicts the assertion of Lemma 1 that $\varphi(t, x) \leqq 0$ for $0 \leqq t < \infty$. Hence $\varphi(t, x) \equiv x'f(t, x) < 0$ for $0 \leqq t < \infty$ and $x \neq 0$.

Since it is obvious that $\varphi(t, 0) = 0$, it only remains to show that $\psi(x) < 0$ for $x \neq 0$ in order to prove the Lemma. We have two cases.

Case I: $\psi(x) < 0$ for $x \neq 0$, by assumption.

Case II: If $\psi(x)$ is strictly concave in x, we have for $0 < \lambda < 1$ and $x \neq 0$

$$(1 - \lambda)\psi(0) + \lambda\psi(x) < \psi(\lambda x),$$

and since $\psi(0) = 0$,

$$\psi(x) < \frac{\psi(\lambda x)}{\lambda} \qquad \text{for} \quad 0 < \lambda < 1, \quad x \neq 0. \tag{2.2}$$

Now, for $x \neq 0$

$$\psi(x) = \limsup_{t \to \infty} x'f(t, x) \leqq 0$$

where the last inequality follows from the fact (proven above) that $x'f(t, x) < 0$ for $0 \leqq t < \infty$ and $x \neq 0$. Hence

$$\psi(x) \leqq 0 \qquad \text{for } x \neq 0. \tag{2.3}$$

Now if the equality sign in (2.3) is ever satisfied for some $x \neq 0$, then it follows from (2.2) that $\psi(\lambda x) > 0$ for $0 < \lambda < 1$ and $x \neq 0$, which contradicts (2.3). Hence $\psi(x) < 0$ for $x \neq 0$ and thus $x'f(t, x)$ is negative definite.

A lemma similar to the preceding one will now be given which will be used in proving Theorem 3. Lemma 3 follows from Lemma 2 by essentially noting that the negative of a strictly concave function is a strictly convex function.

LEMMA 3. *Let $f(t, x)$ be continuous in x at $x = 0$ for $0 \leqq t < \infty$, let $f(t, 0) = 0$ for $0 \leqq t < \infty$, and let $\psi(x) \equiv \liminf_{t \to \infty} x'f(t, x)$. If $x'f(t, x)$ is a strictly convex function of x for $0 \leqq t < \infty$ and if either (I) $\psi(x) > 0$ for $x \neq 0$, or (II) $\psi(x)$ is strictly convex in x, then $x'f(t, x)$ is positive definite.*

We will also need the following theorems of Liapunov and Massera.

LIAPUNOV'S STABILITY THEOREM [2, p. 109], [8, p. 37], [9, p. 707]. *If a positive definite scalar function $V(t, x)$ of class C^1 exists for which $\dot{V} \leqq 0$, then the point $x = 0$ is a stable equilibrium point of the system* (1.1).

MASSERA'S THEOREM [10, p. 200]. *If a scalar function $V(t, x)$ of class C^1 exists which is positive definite, tends to infinity with $\| x \|$, has an infinitely small upper bound, and is such that $\dot{V}(t, x)$ is negative definite, then $x = 0$ is a uniformly, asymptotically stable point in the large of the system* (1.1).

LIAPUNOV'S INSTABILITY THEOREM [2, p. 110], [8, p. 38]. *If a function $V(t, x)$ of class C^1 exists which is positive definite, has an infinitely small upper bound and is such that $\dot{V}$ is positive definite, then $x = 0$ is unstable.*

3. Stability theorems.

THEOREM 1. (*Stability*) *Let $f(t, x)$ be continuous in x at $x = 0$ for $0 \leqq t < \infty$ and let $f(t, 0) = 0$ for $0 \leqq t < \infty$. If $x'f(t, x)$ is a concave function of x for $0 \leqq t < \infty$, then the point $x = 0$ is a stable equilibrium point of the system* (1.1).

Proof. Consider the Liapunov function $V(x, t) = x'x$, which is obviously positive definite. It follows then that for $0 \leqq t < \infty$, $\dot{V} = 2x'\dot{x} = 2x'f(t, x) \leqq 0$, where the last inequality holds because of Lemma 1. Hence by Liapunov's Stability Theorem, the point $x = 0$ is a stable equilibrium point.

The following corollary to Theorem 1 follows from the fact [1, p. 18], [3] that for a twice continuously differentiable concave function, the matrix of second partial derivatives is negative semi-definite.

COROLLARY 2. *If $f(t, 0) = 0$ for $0 \leqq t < \infty$ and if the function $f(t, x)$ is a twice continuously differentiable of x for $0 \leqq t < \infty$ and $\| x \| < \infty$, then the point $x = 0$ is a stable equilibrium point of* (1.1) *provided that the matrix $H_{ij} = [\partial^2 x'f(t, x)/\partial x_i \partial x_j]$ is negative semidefinite for $0 \leqq t < \infty$ and $\| x \| < \infty$.*

THEOREM 2. (*Uniform Asymptotic Stability in the Large*) *Let $f(t, x)$ be continuous in x at $x = 0$ for $0 \leqq t < \infty$, let $f(t, 0) = 0$ for $0 \leqq t < \infty$ and let $\psi(x) \equiv \limsup_{t\to\infty} x'f(t, x)$. If $x'f(t, x)$ is a strictly concave function of x for $0 \leqq t < \infty$ and if either (I) $\psi(x) < 0$ for $x \neq 0$, or (II) $\psi(x)$ is strictly concave in x, then $x = 0$ is a uniformly, asymptotically stable point in the large of the system* (1.1).

Proof. Consider the Liapunov function $V(t, x) = x'x$, which is obviously positive definite, tends to infinity with $\| x \|$, and has an infinitely small upper bound. Now $\dot{V} = 2x'\dot{x} = 2x'f(t, x)$, which is negative definite by Lemma 2. Hence it follows from Massera's Theorem that $x = 0$ is a uniformly asymptotically stable point in the large.

For autonomous systems, $\dot{x} = f(x)$, $f(0) = 0$, it is sufficient for Theorem 2 to hold to require only that $f(x)$ be continuous at $x = 0$ and $x'f(x)$ be strictly concave in x.

An example where neither condition (I) nor (II) of Theorem 2 are satisfied, is, $\dot{x} = -e^{-t}x$, for which $\psi(x) = 0$.[4] It is easy to verify that for this example $x = 0$ is not an asymptotically stable point. Both conditons (I) and (II) are satisfied by the system $\dot{x} = -(1 + e^{-t})x$, for which $x = 0$ is an asymptotically stable point in the large.

Note that Theorem 2 differs from Krasovskii's sufficient conditions for

[4] I am indebted to J. P. LaSalle for this example.

asymptotic stability [7], [6, Theorem 4] in a number of ways: (a) Krasovskii's result is for autonomous systems, Theorem 2 holds for autonomous and non-autonomous systems. (b) In Krasovskii's theorem, $f'f$ is taken as a Liapunov function, whereas in Theorem 2, $x'x$ is a Liapunov function. This is the reason why Krasovskii cannot handle nonautonomous systems for $d/dt(f'f) = f'(J + J')f + 2f'\partial f/\partial t$ (where J is the Jacobian matrix of f with respect to x). Hence unless $\partial f/\partial t = 0$, as is the case for autonomous systems, nothing in general can be said about the system. (c) Krasovskii requires that f be differentiable, Theorem 2 requires f be continuous at $x = 0$. Similarly the sufficient conditions obtained by Hartman [4] and Hartman and Olech [5] require the differentiability of f.

If the function f is twice continuously differentiable in x, then the following corollary to Theorem 2 holds.

COROLLARY 2. *Let* $f(t, 0) = 0$ *for* $0 \leqq t < \infty$, *let* $\psi(x) \equiv \limsup_{t\to\infty} x'f(t, x)$, *and let* $f(t, x)$ *be a twice continuously differentiable function of* x *for* $0 \leqq t < \infty$. *If the matrix* $H_{ij} = \partial^2(x'f(t, x))/\partial x_i\partial x_j$ *is negative definite for* $0 \leqq t < \infty$ *and* $\| x \| < \infty$, *and if either* (*I*) $\psi(x) < 0$ *for* $x \neq 0$, *or* (*II*) $\psi(x)$ *is twice continuously differentiable, and the matrix* $K_{ij} = \partial^2\psi(x)/\partial x_i\partial x_j$ *is negative definite for* $\| x \| < \infty$, *then the point* $x = 0$ *is a uniformly, asymptotically stable point in the large of the system* (1.1).

THEOREM 3. (*Instability*) *Let* $f(t, x)$ *be continuous in* x *at* $x = 0$ *for* $0 \leqq t < \infty$, *let* $f(t, 0) = 0$ *for* $0 \leqq t < \infty$ *and let* $\psi(x) \equiv \liminf_{t\to\infty} x'f(t, x)$. *If* $x'f(t, x)$ *is a strictly convex function of* x *for* $0 \leqq t < \infty$ *and if either* (*I*) $\psi(x) > 0$ *for* $x \neq 0$, *or* (*II*) $\psi(x)$ *is strictly convex in* x, *then* $x = 0$ *is an unstable equilibrium point of the system* (1.1).

Proof. Consider the function $V(t, x) = x'x$, which is obviously positive definite and has an infinitesimally small upper bound. Now $\dot{V} = 2x'\dot{x} = 2x'f(t, x)$, which is positive definite by Lemma 3. Hence it follows from Liapunov's Instability Theorem that $x = 0$ is an unstable point of (1.1).

For autonomous systems, $\dot{x} = f(x)$, $f(0) = 0$, it is sufficient for $x = 0$ to be unstable that $f(x)$ be continuous at $x = 0$ and $x'f(x)$ be strictly convex in x.

If the function f is twice continuously differentiable in x, then the following corollary to Theorem 3 holds.

COROLLARY 3. *Let* $f(t, 0) = 0$ *for* $0 \leqq t < \infty$, *let* $\psi(x) \equiv \liminf_{t\to\infty} x'f(t, x)$, *and let* $f(t, x)$ *be a twice continuously differentiable function of* x *for* $0 \leqq t < \infty$. *If the matrix* $H_{ij} \equiv \partial^2(x'f(t, x))/\partial x_i\partial x_j$ *is positive definite for* $0 \leqq t < \infty$ *and* $\| x \| < \infty$, *and if either* (*I*) $\psi(x) > 0$ *for* $x \neq 0$, *or* (*II*) $\psi(x)$ *is twice continuously differentiable, and the matrix* $K_{ij} = \partial^2\psi(x)/\partial x_i\partial x_j$ *is positive definite for* $\| x \| < \infty$, *then* $x = 0$ *is an unstable equilibrium point of* (1.1).

4. Examples and remarks.

Example 1:

$$\ddot{x}_1 + b(t)\dot{x}_1 + c(t)\dot{x}_1^3 + x_1 = 0, \quad b(t) \geqq 0; c(t) \geqq 0.$$

This equation may be interpreted as the movement of a unit point mass under a unit spring force x_1 and under a nonlinear damping force $b(t)\dot{x}_1 + c(t)\dot{x}_1^3$. The equation may be rewritten as the system

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

$$\dot{x}_2 = -x_1 - b(t)x_2 - c(t)x_2^3.$$

The scalar function $x'f(t, x) = -b(t)x_2^2 - c(t)x_2^4$ is concave (but not strictly concave). Hence by Theorem 1 or Corollary 1 the point $x_1 = 0$, $x_2 = 0$ is stable. Krasovskii's theorem does not apply for this non-autonomous system.

Example 2:

$$\dot{x}_1 = \begin{cases} -x_1 - x_1x_2 & \text{for } x_2 \leqq 1 \\ -x_1 - x_1x_2^3 & \text{for } x_2 > 1 \end{cases}$$

$$\dot{x}_2 = \begin{cases} -x_2 + x_1^2 & \text{for } x_2 \leqq 1 \\ -x_2 + x_1^2x_2^2 & \text{for } x_2 > 1 \end{cases}$$

For the above system $f(0) = 0$ and $x'f(x) = -x_1^2 - x_2^2$ is strictly concave. Hence by Theorem 2 the point $x_1 = 0$, $x_2 = 0$ is uniformly asymptotically stable in the large. Neither Krasovskii's nor Hartman's result can be applied here, because f is not differentiable.

Example 3:

$$\dot{x}_1 = \begin{cases} (1 + \gamma \sin t)(-x_1 + x_1x_2^2) & \text{for } x_1^2 + x_2^2 \leqq 1 \\ (1 + \gamma \sin t)(-x_1 - x_1x_2^2) & \text{for } x_1^2 + x_2^2 > 1 \end{cases}$$

$$\dot{x}_2 = \begin{cases} (1 + \gamma \sin t)(-x_2 - x_1^2x_2) & \text{for } x_1^2 + x_2^2 \leqq 1 \\ (1 + \gamma \sin t)(-x_2 + x_1^2x_2) & \text{for } x_1^2 + x_2^2 > 1 \end{cases}$$

where $0 \leqq \gamma \leqq 0.9$. Note that for this example $f(t, 0) = 0$, and that $f(t, x)$ has discontinuities on $x_1^2 + x_2^2 = 1$. However the scalar function $x'f(t, x) = -(1 + \gamma \sin t)(x_1^2 + x_2^2)$ is strictly concave in x for $0 \leqq t < \infty$, and $\psi(x) = \limsup_{t\to\infty} -(1 + \gamma \sin t)(x_1^2 + x_2^2) \leqq -0.1(x_1^2 + x_2^2) < 0$ for $x \neq 0$. Hence by Theorem 2, the point $x_1 = 0$, $x_2 = 0$, is a uniformly, asymptotically stable point in the large.

As a final remark it should be mentioned that the present results may be used in deriving sufficient conditions for the controllability or non-controllability of systems of the type $\dot{x} = f(t, x, u(t, x))$ where u is a m-dimensional control vector to be selected from a compact convex set Ω

of the m-dimensional real number space. This is being investigated now and will be reported on later.

Acknowledgments. I am indebted to J. B. Rosen for his paper on ρ-stability [11] which stimulated this work. I am also indebted to M. H. Protter for his suggestions and comments, especially in connection with the proof of Lemma 2, and to J. P. LaSalle for a helpful private communication.

REFERENCES

[1] T. Bonnesen and W. Fenchel, *Theorie der convexen Körper*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Vol. 3, Part I, J. Springer, Berlin, 1934. Reprinted Chelsea Publishing Co., New York, 1948.

[2] L. Cesari, *Asymptotic behavior and stability problems in ordinary differential equations*, Ergebnisse der Mathematik und ihrer Grenzegebiete, New Series, Vol. 16, J. Springer, Berlin, 1959.

[3] H. G. Eggleston, *Convexity*, Cambridge University Press, 1958, Chap. 3.

[4] P. Hartman, *On stability in the large for systems of ordinary differential equations*, Canad. J. Math., 13 (1961), pp. 480–492.

[5] P. Hartman and C. Olech, *On global asymptotic stability of solutions of differential equations*, The Johns Hopkins University, 1962.

[6] R. E. Kalman and J. E. Bertram, *Control system analysis and design via the second method of Liapunov, Part I continuous-time systems*, Trans. A.S.M.E., Ser. D. J. Basic Engrg., 82 (1960), pp. 371–393.

[7] N. N. Krasovskii, *On the stability in the large of a system of nonlinear differential equations* (Russian), Prikl. Mat. Mekh., 18 (1954), pp. 735–737.

[8] J. LaSalle and S. Lefschetz, *Stability by Liapunov's direct method with applications*, Academic Press, New York, 1961.

[9] J. L. Massera, *On Liapunov's conditions of stability*, Ann. of Math., 50 (1949), pp. 705–721.

[10] J. L. Massera, *Contributions to stability theory*, Ann. of Math., 64 (1956), pp. 182–205.

[11] J. B. Rosen, *Controllable stability and equivalent nonlinear programming problem*, Contributions to Differential Equations, Academic Press, New York, 1962, pp. 366–376.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

ON COMPUTING OPTIMAL CONTROL WITH INEQUALITY CONSTRAINTS*

YU-CHI HO† AND PIERO B. BRENTANI‡

1. Introduction. The growth of interest in the problem of optimal control has reached exponential rate in the past few years. The volume of papers and reports even in the special case of the bang-bang control problem, is almost large enough to fill a five foot shelf. Furthermore, most of these problems seem to require treatment by techniques not ordinarily associated with the familiar Laplace transform theory or other tools commonly used by practicing engineers. This is undesirable from the viewpoint of quickly reducing the theory to practice. It is the understood purpose of this symposium and this paper to accomplish the following: to present in a self-contained manner a body of recently developed techniques in the control field; to derive useful results using a minimum of advanced mathematics but a maximum of physical and geometrical intuition; and to show concrete working applications of the theoretical results.

To achieve these ends, we shall adopt a four step procedure in presenting the material:

(1) The basic motivation and approach to the solution of a general optimal control problem will be given immediately after its statement. This will serve as an outline and guidepost for the detailed development which follows.

(2) The method of solution will be presented for a special class (linear) of optimal control problems to illustrate the steps involved.

(3) The necessary modifications for extension to the general (nonlinear) case will be shown.

(4) Finally, practical applications and experiences will be discussed.

It is well to point out here certain basic assumptions we have adopted in writing this paper. We have assumed that the readers to whom this paper is addressed are not overly concerned with the matter of rigor in the technical developments. For example, any function used will be assumed to possess the necessary degree of smoothness to give rise to continuous derivatives of such order as may be required. Moreover, we shall not hesitate to drop the use of any terse, but precise, technical terms in favor of a more pic-

* Received by the editors July 5, 1962 and in revised form April 16, 1963. Presented at the Symposium on Multivariable System Theory, SIAM, November 1, 1962 at Cambridge, Mass.

† Harvard University, Cambridge, Mass.

‡ Deceased. Formerly with the Minneapolis-Honeywell Regulator Co., Boston, Mass.

turesque but perhaps less scientific description if the latter motivates our intuition better. On the other hand, we shall attempt to be accurate in the conceptual description of the various methods of solution, e.g. the distinction between the necessity and sufficiency of a condition will be emphasized instead of the pathological case for which such a condition does not apply.

2. Notations and terminology. Since this paper concerns methods applicable to general systems, it is necessary to manipulate formulas and equations possessing an arbitrary but finite number of variables. Thus, the use of vector matrix notation is almost imperative. The transpose of a vector or matrix is denoted by the prime $'$. The inner or vector product of a vector with itself is called the norm and is denoted by $x'x = \| x \|^2 = \sum_{i=1}^{n} x_i^2$. The generalized norm is then given by $x'Ax = \| x \|_A^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j a_{ij}$ provided that A is positive definite. Components of a vector or matrix are indicated by subscripts. Occasionally, when additional clarity is desired or possible, we shall employ the subscript notation in conjunction with the vector-matrix notation.

Finally, it is assumed that the readers are familiar with the *state* description of a dynamic system at least on the conceptual level. The detailed description and definition of terms such as state variables, trajectories, etc. can be found in [2] and [8].

3. Problem statement. To state the control problem properly, it is necessary to define some terms:

(A) *Plant*—Throughout the paper, it is assumed that the dynamic system, or plant, under consideration is governed by

$$\begin{aligned} &\dot{x} = f(x; u; t); x(t_0) = c \\ &\dot{x}_i = f_i(x_1, \cdots, x_n; u_1, \cdots, u_r; t); x_i(t_0) = c_i, \quad \text{for } i = 1, \cdots, n \end{aligned} \tag{1}$$

where x_i's are the generalized state variables such as position, velocity, energy, mass, heat, etc. Equation (1) further implies that the future behaviour of the system is completely determined by the specification of the state of the system at any one instant (a finite set of numbers) and the input $u(t)$ from that instant on.

(B) *Performance criterion*—The performance of the system is measured by

$$J = \lambda(x(t_1)) + \int_{t_0}^{t_1} L(x(t), u(t), t)\, dt, \tag{2}$$

where the interval $t_1 - t_0$ is called the control interval. If we define

$$\dot{x}_0 = L(x; u; t) = f_0(x; u; t); x_0(t_0) = 0, \tag{3}$$

then

$$J = \lambda(x(t_1)) + x_0(t_1) = \phi(x(t_1)). \tag{4}$$

Thus, there is no loss of generality if only the minimization of $\phi(x(t_1))$ is considered by adjoining (3) to (1) and the addition of an extra state variable x_0. Hence $i = 0, \cdots, n$.

(C) *Constraints*—In addition to (1), it is often required that the behaviour of the dynamic system satisfies certain constraints. There are generally two types:

(i) *terminal constraints* are represented by

$$\psi(x(t_1), t_1) = d \tag{5}$$

or

$$\psi_i(x_1(t_1), \cdots x_n(t_1), t_1) = d_i, \quad \text{for } i = 1, \cdots, m \leqq n$$

(ii) the so-called *in-flight constraints* are imposed on the system throughout the control interval. Typically,

$$|x(t)| \leqq \beta(t) \tag{6}$$

$$|u(t)| \leqq \gamma(t) \tag{7}$$

Note that these are inequality constraints rather than equality constraints (the latter when imposed would simply reduce the number of state variables). By redefining state variables it can also be shown easily that more involved functional constraints of the inequality type can be handled by (6) and (7).

It is now possible to formulate two problems of interest:

Optimal control problem: Given (A), determine u as a *function of time* such that (B) is minimized subject to (C).

Optimal feedback control problem: Given (A), determine u as a *function of the instantaneous state* x such that (B) is minimized subject to (C).

The problems[1] as stated above are quite general. The determination of the thrust program of a rocket such that it arrives at its destination with maximum payload is an example. The determination of the start-up sequence of a chemical process such that it reaches the operating point with minimum cost is another example. The optimal control problem is essentially open loop since the solution is only optimal with respect to the given initial conditions of the system. On the other hand, solution of the feedback problem permits closed loop operation which is important in many applications where noise and random disturbances are present. However, solution

[1] It is assumed that solutions to these problems exist, i.e., the problems are well posed.

to the closed loop problem proves to be very difficult. Only in the case where: (A) is linear, (B) is quadratic, and (C) are simple terminal constraints does one presently have a closed form solution. On the other hand, the open loop optimal control problem is nevertheless important in its own right for the following reasons: (i). The open loop solution will permit us to examine the ultimate performance of the system. This knowledge is necessary when one wishes to evaluate any suboptimal or approximately optimal control scheme for feedback operation; (ii). In many applications, an open loop solution is all that is required, e.g. in trajectory analysis of space missions; and (iii). There are cases, e.g. 24-hour satellite control, where the open loop solution can be carried out on a sufficiently fast time scale compared with the dynamics of the system so that one effectively has instantaneous solutions. Thus, feedback operation is again possible.

In this paper, we shall concentrate on the development of methods for the solution of the general optimal control problem. For a discussion of the closed loop problem, readers are referred to [13] and the discussion following [8].

The optimal open loop control problem as stated above is known as the Bolza problem in the calculus of variations. Effective numerical methods for solving the general problem were not known until the recent work of Bryson and Kelley [1, 3].[2] The purpose of this paper is to treat the problem from a somewhat different viewpoint by emphasizing the roles played by the constraints. This approach was motivated by the results obtained in the original solution of the well-known bang-bang control problem [8.] Briefly, we consider the following steps to the solution of the problem:

(i). *Discretization of the problem:* Since a closed form solution to the general problem is not available at present or in the foreseeable future, we consider a numerical solution using a digital computer. This implies that the problem must be discretized in some manner either by replacing the differential equations with difference equations or assuming, as we have done, that the control u is piecewise constant.

$$u(t) = u(t_0 + kT), \qquad t_0 + kT \leqq t < t_0 + (k+1)T, \quad k = 0, \cdots, K-1, \tag{8}$$

[2] There have been other proposed numerical methods for solving the general two point boundary value problem. Almost all these methods involve the repeated integration of the system as well as the adjoint set of differential equations in one direction. Except in special cases, one set of these differential equations is always unstable which leads to numerical difficulties concerning convergence [see W. Kipiniak *Dynamic Optimization and Control* M.I.T. Press, 1961 p. 42]. On the other hand, this difficulty is not inherent in the approach to be described. For the more specific class of variational problems, e.g. time optimal linear solutions, there are various special techniques (see Lee [15] and Neustadt [16]). Kalaba [17] introduces another numerical method, which however does not handle inequality constraints.

where $T = (t_1 - t_0)/K$; i.e., instead of determining a continuous time function, the values of the function at discrete instants of time are determined. The control interval is divided into many steps.[3] Thus, the problem can be regarded simply as a problem of minimization of a function[4] of many variables complicated by some additional equality and inequality conditions.

(ii) *Solution by successive approximation:* In order to locate the minimum of a function of many variables, one often employs the method of steepest descent or successive approximation. Conceptually, one starts the solution by guessing a set of values for the variables. This yields a certain value for the function which is generally not optimum. For this trial or nominal set of values, one computes the gradient of the function. It is well known that the direction specified by the negative gradient is the direction along which the rate of decrease of the function is the greatest. Thus, knowing the gradient one in fact knows how to change the nominal set of values of the variables to obtain an improvement in the criterion function. This process can be repeated for each new set of values of the variables with a small improvement each time in the criterion function (so as not to violate the implicit assumption of linearity when only gradient information is utilized). In the limit, the process hopefully converges to the true minimum.

(iii). *Problem of constraints:* The procedure in (ii) is applicable if no additional constraints are imposed on the system. The effect of the constraints is to make the gradient of the criterion function nonzero at the minimum. In fact, the gradient takes on definite values depending on the constraints.

Consequently, the solution to the general optimal problem reduces to essentially three crucial questions:

(1) How to compute the gradient?

(2) How to make a descent subject to the constraints using the gradient information?

(3) What is the value of the gradient at a minimum (or how can one be sure that a minimum has been attained)?

The answer to these queries can be found by applying some elementary knowledge of differential equations, n-dimensional geometry, and nonlinear programming, respectively. This will be developed in the next section for the special class of linear systems and extended to nonlinear systems in later sections.

[3] Obviously, there are many other means for discretization each entailing a different degree of approximation. However, these are subjects of numerical analysis and are outside the scope of this paper.

[4] The actual functional relationship may be exceedingly complex and impossible to write down analytically. However, in all that follows, we shall never need this functional relationship. All that is required is that the function be computable, i.e., the governing differential equations are given.

4. Linear systems. In this section, the following assumptions are made:

(A) *Plant*—The dynamic system is governed by

$$\dot{x} = F(t)x + g(t)u, \qquad x(t_0) = c. \tag{9}$$

Only single input plants are considered here to simplify the discussion. However, extension to the case of multi-input plants is straightforward and is discussed later.

(B) *Performance criterion.*

$$J = \| x(t_1) \|_R^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j r_{ij} = x'(t_1) R x(t_1) \tag{10}$$

where R is a positive definite or semi-definite matrix.

Although a treatment of more general criterion functions involving integrals can be included as shown in section 3 by defining new state variables, we shall not do so here, since this will add a nonlinear equation to (9) and destroy the simplicity of the linear equations which we wish to preserve in this section.

(C) *Constraints*—One of two types of constraints will be considered:

$$| u(t) | \leqq \gamma(t) \tag{11}$$

$$| x_j(t) | \leqq \beta(t) \tag{12}$$

where x_j is any component of the state vector.

From the above assumptions, it is seen that only a special class of the general problem is considered here. This is done not only because the special problem is interesting in itself but also because it illustrates the salient points in the general method of solution most effectively. Once the technique has been established for this simple case, extension to the general nonlinear case is almost trivial.

(1) *Method of solution.* Following the steps outlined in section 3, the problem is first discretized by letting

$$u(t) = u_k, \; k \leqq t < k + 1, \qquad \text{for } k = 0, \cdots, K - 1 = t_1 - 1 \tag{13}$$

where we have taken the initial time t_0 as the origin and $T = 1$ without loss of generality. Define

$$u = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{K-1} \end{bmatrix} \tag{14}$$

then the determination of $u(t)$ for $0 \leqq t \leqq t_1$ is equivalent to the determination of u, the vector of values of $u(t)$ at different control steps. Hence-

forth, we shall use u and $u(t)$ interchangeably in this section without further explanation.

Now we pose the question *how does the criterion function J change due to a change in control at step k?* This is given via the chain rule for differentiation

$$\frac{\partial J}{\partial u_k} = \sum_{i=1}^{n} \frac{\partial J}{\partial x_i(t_1)} \frac{\partial x_i(t_1)}{\partial u_k}, \qquad \text{for } k = 0, \cdots, K-1. \tag{15}$$

which says that the change in J due to a change in control at step k can be evaluated by first calculating the change in J due to change in the terminal states and then the changes in the terminal states due to the change of control at step k.

Now define $\partial x_i(t_1)/\partial u_k$ as the ikth element of an $(n \times K)$ matrix H, then eqn. (15) can be written as

$$\text{Grad}_u J = H'(\text{Grad}_{x(t_1)} J) = 2H'Rx(t_1). \tag{15}'$$

For any given control u, $(\text{Grad}_{x(t_1)} J)$ can always be easily computed by computing $x(t_1)$ through direct integration of the system equations. Thus, the problem of calculating the gradient is reduced to the problem of calculating the elements of the matrix H, i.e., the change in the terminal state $x_i(t_1)$ due to a change in control u_k. This is provided by differential equation theory [10]. We have

$$x_1(t_1) = \Phi(t_1, t_0)x(t_0) + \int_{t_0}^{t_1} \Phi(t_1, t)g(t)u(t)\, dt, \tag{16}$$

where $\Phi(t, t_0)$, $\Phi(t_1, t)$ satisfy the equations[5]

$$\frac{d}{dt}\Phi(t, t_0) = F(t)\Phi(t, t_0); \Phi(t_0, t_0) = I \tag{17}$$

$$\frac{d}{d\tau}\Phi', (t_1, \tau) = -F'(\tau)\Phi'(t_1, \tau); \Phi(t_1, t_1) = I. \tag{18}$$

Equation (18), the so-called adjoint equation to the system equation, is of direct value to us since from (16) we get,

$$\frac{\partial x_i(t)}{\partial u_k} = \int_k^{k+1} \sum_{j=1}^{n} \phi_{ij}(t_1, t)g_j(t)\, dt = (H)_{ik} \approx \sum_{j=1}^{n} \phi_{ij}(t_1, \xi)g_j(\xi) \tag{19}$$

where $k \leqq \xi \leqq k+1$. Thus, the matrix H can be calculated by integrating (18) backwards from $t = t_1$ with the initial condition $\phi_{ij}(t_1, t_1) = \delta_{ij}$.

Now consider a small variation δu about any given control u, then a first order series expansion yields,

[5] The derivation of (17) and (18) can be found in many texts such as [10].

$$\Delta J = J_{u+\delta u} - J_u = (\mathrm{Grad}_u J)'\delta u = \sum_{k=0}^{K-1} \frac{\partial J}{\partial u_k}\,\delta u_k \,. \tag{20}$$

One sure way to obtain an improvement for the criterion function J is to make δu_k to have the opposite sign of $\partial J/\partial u_k$, i.e.,

$$u_k^{(\mathrm{new})} = u_k^{(\mathrm{old})} - \eta\left(\frac{\partial J}{\partial u_k}\right)$$

where η is a positive scale factor regulating the amount of the change in control. However, such a change in control may not always be possible due to the constraints on u or x_j (11 and 12). Furthermore, even if it were possible, the scale factor η must be determined so as to ensure $J^{\mathrm{new}} - J^{\mathrm{old}} < 0$.

Case I. Constraint on $u(t)$. Suppose we consider the second term in the series expansion of (20), then

$$\Delta J = (\mathrm{Grad}_u J)'\delta u + \delta u' H' R H \delta u. \tag{21}$$

We wish to choose a small δu such that ΔJ in (21) is less than zero. This is done by first defining a dummy vector v proportional to δu,

$$\delta u = \eta v. \tag{22}$$

The vector v is then determined by either of the two following equations, and the proportionality constant η by (25)

$$v_j = \begin{cases} \gamma_j - u_j \quad \text{if} \quad -\dfrac{\partial J}{\partial u_j} > (\gamma_j - u_j) \\ -\dfrac{\partial J}{\partial u_j} \ \text{if} \ -\gamma_j - u_j \leqq -\dfrac{\partial J}{\partial u_j} \leqq (\gamma_j - u_j), \\ -\gamma_j - u_j \ \text{if} \ -\dfrac{\partial J}{\partial u_j} < -\gamma_j - u_j \,, \qquad j = 0, \cdots, K-1 \end{cases} \tag{23}$$

or

$$v_j = \begin{cases} \gamma_j - u_j \quad \text{if} \quad -\dfrac{\partial J}{\partial u_j} \geqq 0 \\ -\gamma_j - u_j \quad \text{if} \quad -\dfrac{\partial J}{\partial u_j} < 0 \end{cases}, \qquad j = 0, \cdots, K-1 \tag{24}$$

and

$$\eta = \mathrm{sat}\left[-\frac{(\mathrm{Grad}_u J)'v}{2v'H'RHv}\right], \tag{25}$$

where

$$\text{sat } (\alpha) = \begin{cases} \alpha & \text{if } |\alpha| < 1 \\ \pm 1 & \text{if } |\alpha| \geqq 1. \end{cases}$$

Equations (22–25) essentially accomplish the following tasks: (a) Equation (20) tells us the direction to move in the "solution space" (the space with $u_0, u_1, \cdots, u_{K-1}$ as coordinates) to obtain the largest change in the criterion function J. (b) However, the constraint (11) may, and often does, prevent us from making a descent in that direction. Consequently, a *feasible* direction of descent has to be chosen as represented by (23) or (24). These are the directions that do not involve a sign change of the gradient components. (c) Equation (25) then tells us how far we should go along this feasible direction to get the best improvement.

Figure 1 provides a simple geometrical interpretation of the process. Here, for the sake of illustration, only a two dimensional solution space is considered, i.e., the control interval consists of only two steps. Every point in the solution space represents a control. The set of admissible controls is inside the rectangle with center at the origin, while the lines of constant J are quadratic surfaces in this solution space. For any point A in this space the gradient separates the space into two halves. Small

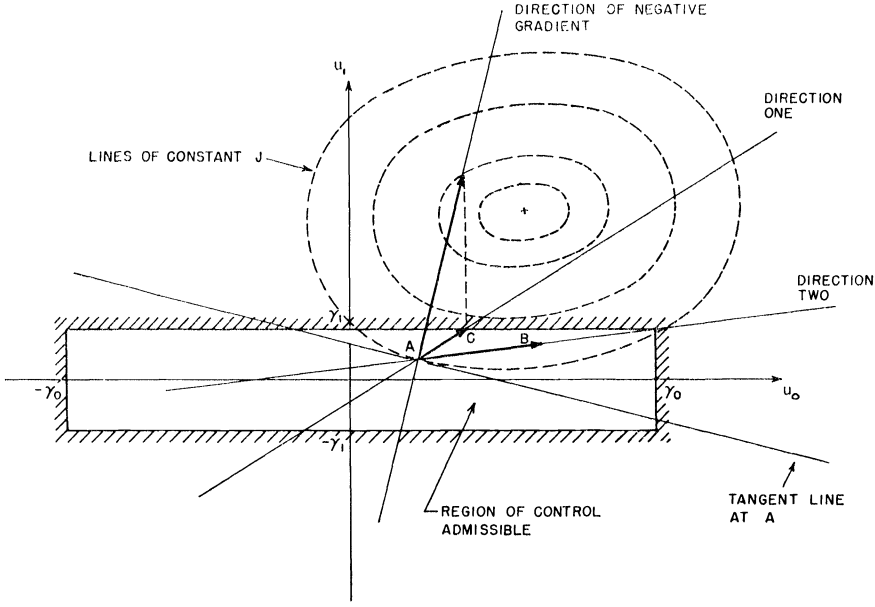

FIG. 1. Solution space interpretation

movement from A into one side causes the function to decrease; into the other side, to increase. Direction AC is simply the projection of the negative gradient onto the boundary and is the direction specified by (23), while direction AB seeks the nearest vertex as specified by (24). The distance to go along either of these directions is determined by the variation of J *along these directions*. In Figure 2b, it is not advantageous to proceed as far as possible, i.e., $\eta < 1$, while the reverse is true in Figure 2a. These geometrical interpretations extend directly without modification to higher dimensions.

Case II. Constraints on x_j. This case can be easily reduced to case I. We note that

$$\delta x_j = B\delta u, \tag{26}$$

where

$$\delta x_j = \begin{bmatrix} \delta x_j(1) \\ \vdots \\ \delta x_j(K) \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \partial x_j(1)/\partial u_0 & & \\ \partial x_j(2)/\partial u_0 & \partial x_j(2)/\partial u_1 & \\ \vdots & \ddots & \ddots \\ \partial x_j(K)/\partial u_0 & \cdots & \partial x_j(K)/\partial u_{K-1} \end{bmatrix}.$$

All the elements in B and $(\mathrm{Grad}_u J)$ are known quantities and can be computed as discussed before. Furthermore, since B is always triangular (a direct consequence of the causality property of the dynamic system), it is feasible to invert B. Then substituting (26) into (21), we get

$$\delta J = (g^+)'\delta x_j + \delta x_j'(B^{-1})'H'RH(B^{-1})\delta x_j \tag{27}$$

where

$$g^+ = B^{-1}\,(\mathrm{Grad}_u J). \tag{28}$$

Now the situation is same as before, a computing scheme analogous to case I is defined by letting

$$\delta x_j = \eta w \tag{29}$$

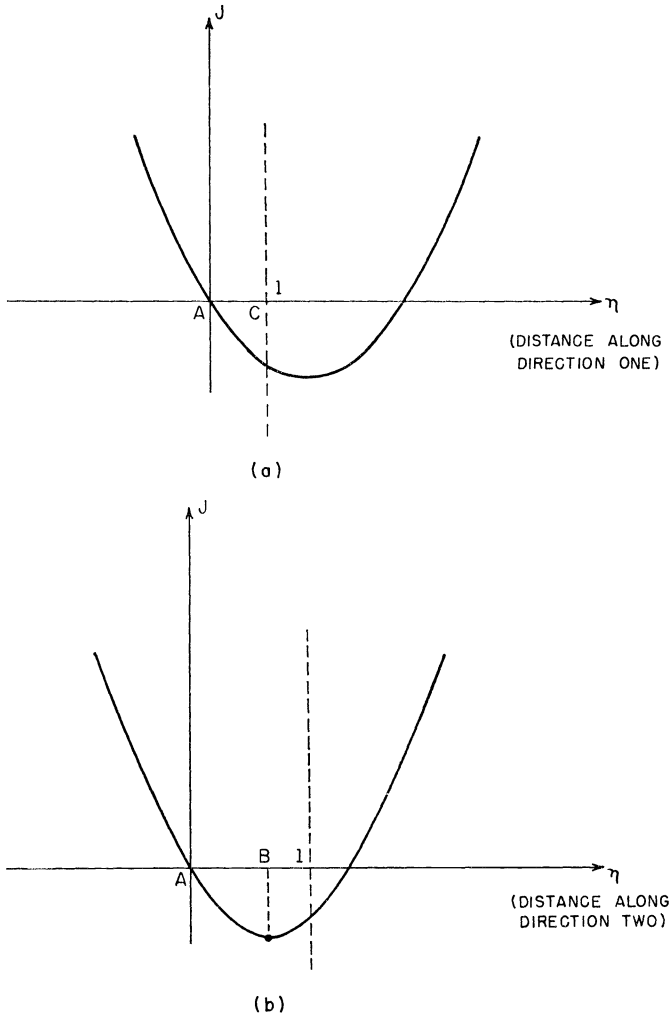

FIG. 2. Distance for descent

$$(30)\quad w_\ell = \begin{cases} \beta_\ell - x_j(\ell) & \text{if} \quad -g_\ell^+ > (\beta_\ell - x_j(\ell)) \\ -g_\ell^+ & \text{if} \quad -\beta_\ell - x_j(\ell) \leqq -g_\ell^+ \leqq (\beta_\ell - x_j(\ell)) \\ -\beta_\ell - x_j(\ell) & \text{if} \quad -g_\ell^+ < -\beta_\ell - x_j(\ell) \end{cases}$$

or

$$(31)\quad w_\ell = \begin{cases} \beta_\ell - x_j(\ell) & \text{if} \quad -g_\ell^+ \geqq 0 \\ -\beta_\ell - x_j(\ell) & \text{if} \quad -g_\ell^+ < 0 \end{cases} \qquad \ell = 1, \cdots, K$$

$$(32)\quad \eta = \text{sat}\left[-\frac{(g^+)'w}{2w'(B^{-1})'H'RHB^{-1}w} \right].$$

Exactly the same geometrical interpretation as in Fig. 1 applies in this case, except that here the space formed with the coordinates $x_j(1)$, $x_j(2)$, $\cdots$, $x_j(K)$, must be considered. Every point in this space defines a trajectory. Changes in trajectory are linearly related to changes in control by the linear transformation B. Thus a hypercube in the x_j-space becomes a polyhedron in the solution space. This is precisely the reason why it is difficult to determine a feasible improvement directly in the solution space for the constrained state variable case. The boundary on u at one time depends on the change in u at other times. To completely specify the boundary of the polyhedron in the solution space the position of 2^K vertices must be determined. For any practical K, say 100, computation time can easily surpass the age of the universe.

(2) *Convergence and optimality.* Section 4, describes a complete cycle of computation for linear optimal control problems subject to control or state variable constraints. Hopefully, by repeated application of this procedure, the true minimum will be monotonically approached in the limit. The purpose of this Section is to indicate that this is indeed so.

Intuitively speaking, there is really no question concerning the convergence of the method. This is clear if we refer back to the geometrical interpretation of the problem in Figure 1. The situation there can be thought of as that of setting free a marble on the surface of a bowl which has constraining walls built in it. Depending on the location of the constraining walls, the marble eventually rolls to a stop at (i) the bottom of the bowl, (ii) against the side of a constraining wall, or (iii) at the intersection of more than one constraining wall. Such a point is obviously the global minimum for the problem. Algebraically, this is expressed by the fact that the computation process comes to a stop only when

$$(33)\quad \begin{aligned} &\gamma_k - u_k = 0 \quad \text{when} \quad (\text{Grad}_u J)_k < 0 \\ &\gamma_k > u_k > -\gamma_k \quad \text{when} \quad (\text{Grad}_u J)_k = 0 \qquad k = 0, \cdots, K-1. \\ &-\gamma_k - u_k = 0 \quad \text{when} \quad (\text{Grad}_u J)_k > 0. \end{aligned}$$

This claim can be seen as follows:

If (33) were not true and the computation stops (i.e. improvement becomes infinitesmal) then (23) or (24) must define a v which is *not* infinitesimal. Thus η given by (25) cannot be arbitrarily small which implies that ΔJ from (21) is not infinitesimally small. This contradicts the assumption that no more improvement can be made.

Equation (33) says that the terminating (hopefully optimal) control has the property that every gradient component is either zero when the control is off the boundary (meaning no improvement can be made even though variation is possible) or pointing outwards when the control is on the boundary (meaning that improvement is possible only by violating the constraints). Thus, the terminating control obviously satisfies the *necessary* conditions for an optimum. To prove that the conditions of (33) are also *sufficient*, some further artifices must be constructed and certain elementary knowledge of nonlinear programming must be invoked. Define,

$$A = \underbrace{\left[\begin{array}{cccc} -1 & 0 & \cdots & 0 \\ 0 & -1 & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & -1 & 0 \\ 0 & \cdots & 0 & -1 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & 1 & 0 \\ 0 & \cdots & 0 & 1 \end{array}\right]}_{K \text{ cols.}} \left.\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array}\right\} 2K \text{ rows} \qquad \gamma_1 = \left[\begin{array}{c} \gamma \\ \gamma \end{array}\right] \left.\begin{array}{c} \\ \\ \end{array}\right\} 2K \text{ rows}$$

$$A = \left[\begin{array}{c} -I \\ \hline I \end{array}\right], \qquad \gamma_1 = \left[\begin{array}{c} \gamma \\ \hline \gamma \end{array}\right].$$

then the optimal control problem with constrained u can be restated as minimize J subject to $Au - \gamma_1 < 0$. Now consider the Lagrangian

$$\phi(u, \lambda) = J + \lambda'(Au - \gamma_1). \tag{34}$$

Differentiation directly verifies that

$$\text{Grad}_{u^0}\phi(u^0, \lambda^0) = 0, \tag{35}$$

$$\text{Grad}_{\lambda}\phi(u^0, \lambda) \leqq 0 \quad \text{for } \lambda \geqq 0, \tag{36}$$

and

$$\lambda^{0\prime}(Au^0 - \gamma_1) = 0. \tag{37}$$

Further, ϕ is convex in u since J is convex in u. If u^0 is given by (33) and λ^0 is given by

$$\lambda_k{}^0 = \begin{cases} \left|\dfrac{\partial J}{\partial u_k}\right| & \text{if} \quad -\gamma_k - u_k = 0 \\ 0 & \text{if} \quad -\gamma_k - u_k \neq 0 \end{cases} \tag{38}$$

$$k = 0, \cdots, K-1$$

$$\lambda_{K+k}^0 = \begin{cases} \left|\dfrac{\partial J}{\partial u}\right| & \text{if} \quad \gamma_k - u_k = 0 \\ 0 & \text{if} \quad \gamma_k - u_k \neq 0, \end{cases} \tag{39}$$

then (35–37) and the convexity say that the particular pair u^0 and λ^0 yield a saddle point for the function ϕ for $\lambda \geqq 0$ i.e.,

$$\phi(u, \lambda^0) \geqq \phi(u^0, \lambda^0) \geqq \phi(u^0, \lambda)$$

for all $\lambda \geqq 0$ and all admissible u. However, by a well-known theorem of nonlinear programming [11] we know that if u^0, λ^0 is a saddle point for ϕ then u^0 must be a global minimum for J. This proves the sufficiency. The same arguments can be employed for the case of constraints on x_j since it is one-to-one related to the u-constrained case. We omit the details.

Actually, the above construction is not as artificial as it may seem at first glance. Equation (34) is highly suggestive of the conventional La-Grange Multiplier method for minimization subject to equality constraints. The method of nonlinear programming merely extends this approach to the case of inequality constraints. The multiplier λ is seen to be intimately related to the gradient of the function to be minimized (39) which is in turn given by the solution of the adjoint system of equations (18). In the classical calculus of variation literature, the adjoint equations are known as the "multiplier rule" and their solutions the "multiplier functions." Viewed in the present context, they acquire more of a physical or intuitive appeal and justify their existence through very simple arguments. The method described in section 4 is simply a constructive approach to the determination of the multiplier λ or the multiplier function $\lambda(t)$ if the vector is viewed as a time function. Reference [8] shows how, in the well-known bang-bang control problem, the gradient ($\text{Grad}_u J$) can be re-interpreted to yield the various known conditions for an optimum.

It is also possible to derive an interesting conclusion when one of the state variables is constrained in magnitude. Equation (26) shows that it is identical to the control variable constrained case except for the additional nonsingular linear transformation B. Since such a transformation preserves all properties in linear spaces, we are led to the following conclusion:

For linear dynamic systems, the minimum time or the minimum terminal error (provided it is finite) problem when one of the state variables is con-

strained will have bang-bang solutions, i.e., the constrained state variable must remain on the boundary at all times except at the instant of switching. Consequently, there must be impulsive actions in the optimal control function.

This conclusion is experimentally verified in section 5 where computer simulations of the techniques discussed here are carried out.

(3) *Remarks and extensions.* Having described the method in previous sections for a special class of systems, we are now in a position to discuss some ready extensions of the method to more involved cases. The framework of a linear dynamic system, however, will be retained.

Multiple control and state variables constraints. For purpose of discussion, we shall assume there are r input and p state variable constraints.

(i) *r constrained inputs, no state variable constraint.* Since the system is linear, the superposition principle holds. Instead of one gradient, there will be several. Equation (20) becomes,

$$\Delta J = (\mathrm{Grad}_{u^1} J)'\delta u^1 + \cdots + (\mathrm{Grad}_{u^r} J)'\delta u^r + O((\delta u)^2). \tag{40}$$

For every constained input, we have

$$v_j^m = \begin{cases} \gamma_j^m - u_j^m & \text{if } -\dfrac{\partial J}{\partial u_j^m} > (\gamma_j^m - u_j^m) \\ -\dfrac{\partial J}{\partial u_j^m} & \text{if } -\gamma_j^m - u_j^m < -\dfrac{\partial J}{\partial u_j^m} < (\gamma_j^m - u_j^m) \\ -\gamma_j^m - u_j^m & \text{if } -\dfrac{\partial J}{\partial u_j^m} < -\gamma_j^m - u_j^m \end{cases} \tag{41}$$

$$m = 1, \cdots, r$$

$$j = 0, \cdots, K-1,$$

i.e., a set of equations similar to (23). The method and the proof of convergence may be made in the same way.

(ii) *r unconstrained inputs and one state variable constraint.* In this case, one of the inputs can be used to keep the system within the state variable constraint, e.g. the first input.

We have $\delta x_j = B_1\delta u^1 + \cdots + B_r\delta u^r$. Then (40) becomes

$$\Delta J = (g^+)'\delta x_j + \sum_{m=2}^{r} [(\mathrm{Grad}_{u^m} J)' - (\mathrm{Grad}_{u^1} J)'B_1^{-1}B_m]\delta u^m. \tag{42}$$

Instead of (41), we use,

$$v_j^m = -j\text{th component } [(\mathrm{Grad}_{u^m}{}' J) - B_m'(B_1^{-1})' (\mathrm{Grad}_{u^1} J)] \tag{43}$$

$$m = 2, \cdots, r$$

$$j = 0, \cdots, K-1.$$

$$(44)\quad w_\iota = \begin{cases} \beta_\iota - x_j(\ell) & \text{if} \quad -g_\iota^+ > (\beta_\iota - x_j(\ell)) \\ -g_\iota^+ & \text{if} \quad -\beta - x_j(\ell) < -g_\iota^+ < (\beta_\iota - x_j(\ell)) \\ -\beta_\iota - x_j(\ell) & \text{if} \quad -g_\iota^+ < -\beta_\iota - x_j(\ell) \end{cases}$$

$$\ell = 1, \cdots, K.$$

which is the same as (30), and

$$\delta u^1 = B_1^{-1}(\delta u_j - B_2\delta u^2 - \cdots - B_r\delta u^r).$$

Equation (43) in this case simply expresses the fact that there are no constraints on inputs 2 to r. Since originally there are no constraints on all inputs, it does not matter which input is chosen to keep the system within x_j limits. In practice, there may be other considerations which will cause us to prefer one input to another.

(iii) *r unconstrained inputs and r constrained state variables.* This is analogous to Case II discussed in section 4. All constraints on the r state variable are reflected as constraints on the r inputs. The relation is provided by a$(rK \times rK)$ matrix which is block triangular. Consequently, its inversion means the inversion of a series of $(r \times r)$ matrices. This may become computationally unfeasible for large r. Otherwise, conceptually the solution is the same as Case II in section 4 with (30) duplicated r times.

(iv) *r unconstrained inputs less than r constrained state variables.* This case is equivalent to (3) plus the addition of $r - p$ unconstrained inputs. Conceptually the solution is the same as (2).

(v) *General input/output constrained problem:* We shall only discuss the case to illustrate the problems involved. Consider the same problem as in Section 4 but with *both* the control constraint (eqn. 11) and the state variable constraint (12) imposed. Then to perform one cycle of the successive approximation computation we must solve the following problem: determine δu such that $(\text{Grad}_u J)'\delta u < 0$ subject to

$$-\beta - x_j \leqq B\delta u = \delta x_j \leqq \beta - x_j$$

$$-\gamma - u \leqq \delta u \leqq \gamma - u.$$

The difficulty with this case is the fact that with both constraints imposed it is no longer convenient to work in either the u-space or the x_j-space. Both constraints cannot be expressed simply in any one space as we have done in Cases I and II. However, the above problem can be recognized as one iteration of the following linear program problem: minimize $(\text{Grad}_u J)'v$ subject to

$$a^- \leqq Bv \leqq a^+$$

$$b^- \leqq v \leqq b^+$$

where we identify v as δu, a^+ as $\beta - x_j$, and b^+ as $\gamma - u$, etc. Algorithms to perform iterations for linear programs, such as the simplex method, are well known. Thus, we find that the general input/output constrained problem can be solved by a series of linear programs with different cost vectors ($\text{Grad}_u J$). This statement applies also to the more general case with several constrained inputs and outputs.

Another terminal criterion. Instead of $J = \| x(t_1) \|_R^2$, one may consider a general terminal criterion $J = \phi(x(t_1))$. The only requirement is that ϕ should be convex which is almost invariably satisfied in practice. Computationally, we simply replace 2 $Rx(t_1)$ by $(\text{Grad}_{x(t_1)}\phi)$ in all the equations in section 4. In (21), (25), (28), and (32) the matrix R is replaced by the matrix formed by $\partial^2\phi/\partial x_i(t_1)\partial x_j(t_1)$. This, of course, is equivalent to fitting a quadratic surface to the general surface at the point in question.

Determination of "η". It was pointed out that the scalar variable "η" in (22) and (29) plays the role of distance of descent along the feasible gradient direction. A reasonable value for "η" was given by (25) or (32) which requires the evaluation of second partial derivatives. Since we are concerned primarily with devising an efficient method for making an improvement, use of (25) or (32) may be computationally uneconomical. A more crude approach would simply assume some "η" and integrate the system equations to verify if an improvement has actually been made. If the new control produces a worse J, then the error in estimating the improvement gives us an estimate of the magnitude of the 2nd order term in (21) or (28). A new and appropriate "η" can then be chosen and tried again. There always exist a $\eta > 0$ such that the new control yields an improvement.

Problem of terminal constraints. In addition to minimizing $\phi(x(t_1))$, the problem may require that the state of the system terminate in some subset of the state space by the additional constraints

$$\psi_i(x(t_1)) = d_i, \qquad \text{for } i = 1, \cdots, m. \tag{45}$$

To handle such a case, we modify the criterion function to

$$J = (\phi(x(t_1)) - \phi_0)^2 + \sum_{i=1}^{m} \mu_i(\psi_i(x(t_1) - d_i)^2 \tag{46}$$

where ϕ_0 is some clearly unattainable minimum of ϕ, and μ_i are large positive constants. It is intuitively reasonable that the solution to the modified problem without the terminal constraints can be made arbitrarily close to the original problem with the terminal constraints, provided the μ_i's are chosen appropriately. Details for doing this can be found in [9].

There is another exact way of incorporating terminal constraints when no in-flight constraints are simultaneously present. This is the gradient projection technique due to Bryson [2]. Essentially, the idea is to project

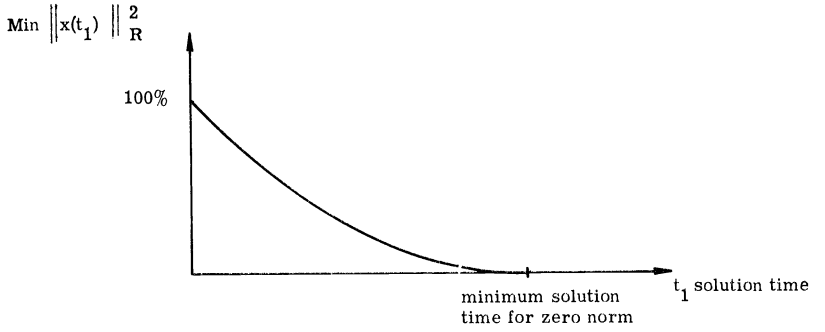

FIGURE 3

the gradient ($\text{Grad}_u J$) onto the intersection of the hyperplanes tangent to the surfaces $\psi_i(x(t_1)) = d_i$, $i = 1, \cdots, m$ at the point in question in the solution space. If the improved solution is obtained by moving along this projected gradient, then the new terminal state due to the improved control does satisfy the terminal constraints excepting for a second order effect. A correction scheme to handle the error accumulated can be easily incorporated. The computation process stops when the projected gradient becomes zero, i.e., the gradient is normal to the subspace defined by $\psi_i(x(t_1)) = d_i$, $i = 1, \cdots, m$.

When in-flight constraints are present, the incorporation of the gradient projection scheme is much more cumbersome. It is not clear at this stage which of the two schemes (gradient projection or penalty function) is more efficient in practice when in-flight constraints are present.

Problem of minimal time. This problem is usually stated as minimizing the time for the system to reach some fixed terminal state. However, an equally satisfactory way is to minimize the error from this fixed terminal state for a fixed time. If the minimum error for a particular fixed time is finite, then it immediately follows that the time optimal problem can not be solved in this amount of time. Thus, by observing the behaviour of the minimum error problem for several fixed times, the minimal time can be found quickly. This situation is illustrated in Figure 3.

General problem of variable terminal time. Instead of a fixed control interval, the problem is to be terminated when a certain stopping condition

$$\Omega(x(t_1)) = 0 \tag{47}$$

is reached where Ω is usually some monotonic function of time such as the altitude. The change in the criterion function due to changes in terminal state now becomes a more complicated expression. We have,

$$\begin{aligned}\delta J &= (\mathrm{Grad}_{x(t_1)}\, J)'\delta x(t_1) + \frac{dJ}{dt_1}\,\delta t_1 \\ &= \sum_{i=1}^{n} \frac{\partial J}{\partial x_i(t_1)}\,\delta x_i(t_1) + \frac{dJ}{dt_1}\,\delta t_1\end{aligned} \tag{48}$$

The second term is contributed by the fact that a change in the terminal state also changes the termination time. From (47),

$$\begin{aligned}\delta\Omega = 0 &= (\mathrm{Grad}_{x(t_1)}\,\Omega)'\delta x(t_1) + \frac{d\Omega}{dt_1}\,\delta t_1 \\ &= \sum_{i=1}^{n} \frac{\partial \Omega}{\partial x_i(t_1)}\,\delta x_i(t_1) + \frac{d\Omega}{dt_1}\,\delta t_1 .\end{aligned} \tag{49}$$

Combining (48) and (49) and eliminating δt_1, we get,

$$\delta J = \left(\mathrm{Grad}_{x(t_1)}\, J - \frac{\dot{J}}{\dot{\Omega}}\,\mathrm{Grad}_{x(t_1)}\,\Omega\right)' \delta x(t_1)$$

or

$$\frac{\delta J}{\delta x_i(t_1)} = \left(\frac{\partial J}{\partial x_i(t_1)} - \frac{\dot{J}}{\dot{\Omega}}\,\frac{\partial \Omega}{\partial x_i(t_1)}\right). \tag{50}$$

The vector $\delta J/\delta x_i(t_1)$, $i = 1, \cdots, n$ plays an equivalent role as the vector $(\mathrm{Grad}_{x(t_1)}J) = 2Rx(t_1)$ in all equations in section 4 or the more general vector $(\mathrm{Grad}_{x(t_1)}\phi)$ as discussed earlier in this section. If this identification is made, all other aspects of the method remain the same. Lastly, (50) can be directly generalized to cases where additional terminal constraints are present. This then becomes the gradient projection technique of Bryson and Denham [2].

5. Nonlinear systems. Actually, almost all the crucial steps towards the developement of a general, computational method for the solution of the optimal control problem as formulated in section 3 have been worked out in the last section. The extension to the general nonlinear case is really quite trivial from a conceptual viewpoint. Essentially, we adopt the following approach: *For any given control to a nonlinear system, there results a specific trajectory. The equation of motion can be linearized about this nominal trajectory. For small perturbation in control and trajectory, the system obeys the linearized set of differential equations. Consequently, the techniques developed in the previous section can be used to devise a small improvement in control. Having obtained a new control, and thus, a new trajectory, the nonlinear equations can be linearized again about this trajectory. The computation cycle repeats.* Mathematically speaking, the idea is expressed by considering a variation of control and trajectory to the system differential equation (1).

$$\dot{x} + \delta\dot{x} = f(x + \delta x; u + \delta u; t), \qquad x(t_0) = c. \tag{51}$$

Expanding the left hand side of (51) by a Taylor series and simpliflying, we obtain the well known variational equation

$$\delta\dot{x} = F(x(t), u(t))\delta x + G(x(t), u(t))\delta u, \qquad \delta x(t_0) = O \tag{52}$$

where the ijth element of the $(n \times n)$ matrix F is $\partial f_i/\partial x_j$ and the ikth element of the $(n \times r)$ matrix G is $\partial f_i/\partial u_k$.

The matrices F and G are functions of the nominal trajectory and control. For specific $x(t)$ and $u(t)$, they are functions of time. The solutions to the time varying but linear differential equations (52) can be written down as

$$\delta x(t_1) = \int_{t_0}^{t_1} \Phi(t_1, t)G(t)\delta u(t)\, dt, \tag{53}$$

where $\phi_{ij}(t_1, t)$ obeys the adjoint set of differential equations to (52)

$$\dot{\Phi}'(t_1, t) = -F'(t)\Phi'(t_1, t), \qquad \Phi(t_1, t_1) = I. \tag{54}$$

With (53) and (54), the matrices H and B in (15), and (26) in section 4 can always be computed. Once H and B are known, the rest of the procedure goes through without modification.

As a practical matter, there is a simplification worth mentioning. In Section 4, we considered the computation of H and B as a separate operation independent of the initial conditions of the dynamic system, i.e., the backward integration of (18) need only be performed once before the actual iterations and it is not necessary to repeat the integration during each computation cycle. This situation is no longer possible for the nonlinear case. Since the matrices F and G depend on the trajectory and the control, (54) must be integrated backwards during each computation cycle to determine the H and B for that particular cycle. This involves the integration of n^2 linear differential equations on the digital computer—a straightforward but nevertheless somewhat cumbersome operation. This labor can, however, be easily reduced to only n equations by the following manipulation. Consider the single input case once again. Note that instead of eqn. (15), the $(\mathrm{Grad}_u J)$ can also be expressed as

$$(\mathrm{Grad}_u J)_k = g'\Phi_k'(\mathrm{Grad}_{x(t_1)} J) = 2g'\Phi_k' Rx(t_1), \quad k = 0, \cdots, K-1 \tag{55}$$

where

$$\Phi_k' = \begin{bmatrix} \phi_{11}(t_1, \xi_k) & \phi_{21}(t_1, \xi_k) & \cdots & \phi_{n1}(t_1, \xi_k) \\ \vdots & & & \\ \phi_{1n}(t_1, \xi_k) & \phi_{2n}(t_1, \xi_k) & \cdots & \phi_{nn}(t_1, \xi_k) \end{bmatrix} \tag{56}$$

where $t_0 + k \leqq \xi_k \leqq t_0 + k + 1$, (cf. (19) and (8)).

Each submatrix Φ_k' is simply the solution of (54) at some instant or small interval $t_0 + k$ to $t_0 + k + 1$. Thus, the vector

$$p(t_1, \xi_k) = 2\Phi_k' R x(t_1)$$

is obtained by taking a linear combination of the columns of $\Phi'(t_1, \xi_k)$ which are solutions of the adjoint equations (54) for specific initial conditions. Equivalently, if we take a linear combination of the specific initial conditions, then we should be able to obtain the vector $p(k)$ by integrating the differential equation

$$\begin{aligned} \dot{p}(t_1, t) &= -F' p(t_1, t) \\ p(t_1, t_1) &= 2Rx(t_1) \end{aligned} \tag{57}$$

backwards from t_1. This is only n equations instead of the n^2 equations specified by (54). The gradient $(\mathrm{Grad}_u J)$ is then obtained by taking a further linear combination of the solution of (57). For general terminal criterion with variable terminal time, one uses the initial condition $\delta J/\delta x_i(t_1)$ as given by (50) in place of $2Rx(t_1)$ to start the backward integration of (57).

To recapitulate: *a general, monotonically converging, computational method for solving general nonlinear optimal control problems with bounded control variables can be described by the following steps*:

(*a*) *Start with any feasible control* $u(t)$ *as a guess.*

(*b*) *Use the* $u(t)$ *and integrate the system equation* (1) *to obtain the nominal trajectory* $x(t)$.

(*c*) *Linearize the system equation about the nominal* $x(t)$ *and* $u(t)$ *to obtain* (52).

(*d*) *Compute the gradient* $(\mathrm{Grad}_u J)$ *by integrating* (57) *backwards from* t_1 *with the proper initial conditions.*

(*e*) *Devise an improvement* $\delta u(t)$ via (21)–(25).

(*f*) *Let the new control be* $u(t) + \delta u(t)$ *and return to step* (*b*).

Executing the above scheme many times one eventually arrives at a computation cycle which has the properties that: (i) For the system equation (1), one has a set of initial conditions $x(t_0)$, a control $u^0(t)$, and a resultant trajectory $x^0(t)$; (ii) For the $x^0(t)$, the adjoint linearized (57) has a set of initial conditions $p^0(t_1, t_1)$ and a solution $p^0(t_1, t)$; and (iii) Conditions (i) and (ii) are related to the fact that the variation of the Hamiltonian

$$H = -\sum_{i=1}^{n} f_i(x; u^0; t) p_i{}^0(t_1, t) \tag{58}$$

cannot be increased anywhere along the trajectory, i.e.,

$$\delta H = -\sum_{i=1}^{n} \frac{\partial f_i}{\partial u} p_i(t_1, t) \delta u(t) = 0, \qquad t_0 \leqq t \leqq t_1$$

or

$$-(\mathrm{Grad}_u J)'\delta u = 0. \qquad \text{(cf. (23))}$$

The Hamiltonian has reached its maximum everywhere along the trajectory.

The connection of conditions (i–iii) with the celebrated Maximum Principle of Pontryagin scarcely needs mentioning.

The computation procedure for state variable constrained case is similar. We omit the details [9]. However, three points of practical importance should be mentioned. First, it should be pointed out that the state variable constraint is handled in the nonlinear case only in an approximate manner since a linearization procedure is involved. A correction routine may or may not be needed depending upon whether or not x_j is a convex function of the u_k's. Secondly, the computation of the matrix B can be most easily obtained by direct integration of the linearized system (52) with appropriate initial conditions rather than via the use of adjoint equation. Lastly, B^{-1} is never needed in practice. Only the solution of (26) is necessary. This is a still simpler task than inverting a triangular matrix.

6. Experimental results.

(1) *Linear systems.* In order to obtain some computational experience with the method, two pilot programs were written in FORTRAN for the IBM 7090. The two programs essentially carry out the method of solution described in section 4 for control variable or state variable constrained linear systems. The specifications of the two programs are as follows:

Program A. Given:

a) a multi-input dynamic system $\dot{x} = Fx + Gu$
   where F = constant square matrix $(n \times n)$
   G = constant matrix $(r \times n)$,

b) an initial condition $x(0)$,

c) a constraint on the controls, of the form $|u(t)| \leqq \gamma(t)$,

d) a constant square matrix R, to evaluate the performance,

e) a terminal time t_1 or equivalently the number of control steps K.

Then Program A finds $u(t)$, $0 < t < t_1$, satisfying the constraint (c) which minimizes the quadratic form

$$V = x'(t_1)Rx(t_1).$$

Program B. This program is similar to Program A, with the exceptions:

a) the constraint is of the form

$$|x_n(t)| \leqq \beta(t)$$

   where $x_n(t)$ is the nth component of the state vector.

b) the dynamic system has to be single input, i.e., $r = 1$.

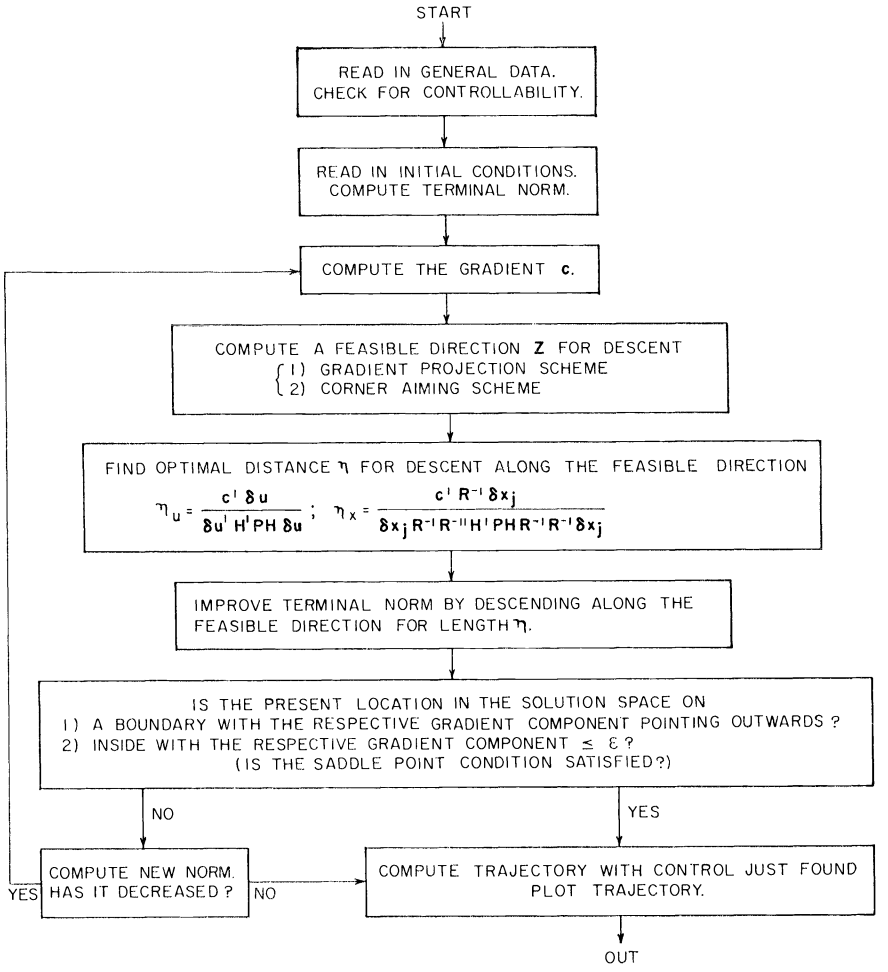

FIG. 4. Flow chart for computer program

In both Program A and B the complete controllability of the system is checked and the figure of merit computed [14]. Furthermore, the transient response of the system under optimal control is also automatically plotted by the computer. The flow chart for the programs is shown in Figure 4.

The programs are useful in solving the following two general types of problems for linear systems with inequality constraints: (i) to get as close as possible to some given terminal state or set of terminal states in a given time period; and (ii) to get to some given terminal state or set of terminal state in shortest time.

Examples. As an example for demonstration, we chose a linear system described by the transfer function

$$G(s) = \frac{(s + 0.5)}{(s + 1)(s + 2)(s^2 + 2s + 2)}.$$

In the state vector notation, the dynamics of the system become:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -4 & -10 & -10 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ -4.5 \end{bmatrix} \cdot u.$$

Assume a disturbance or an equivalent step input

$$\begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \\ x_4(0) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

It is desired to bring x_1, x_2, x_3, x_4 to zero in minimum time, while constraining the input to $|u(t)| \leqq 1$.

Using Program A repeatedly, (with a positive definite matrix R), the results as shown in Figures 5 and 6 are obtained.

Instead of constraining the input it may be desired to constrain one of the components of the state vector, e.g. x_4, $|x_4| \leqq 1$.

Using Program B with the same matrix R one obtains analogous results displayed in Figures 7 and 8. Note that the constrained state variable,

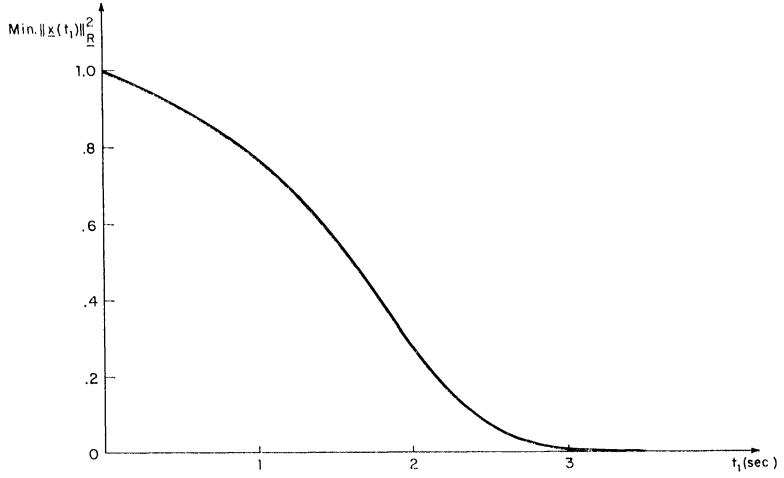

FIG. 5. Minimum terminal norm as a function of terminal time t_1 (control constraint)

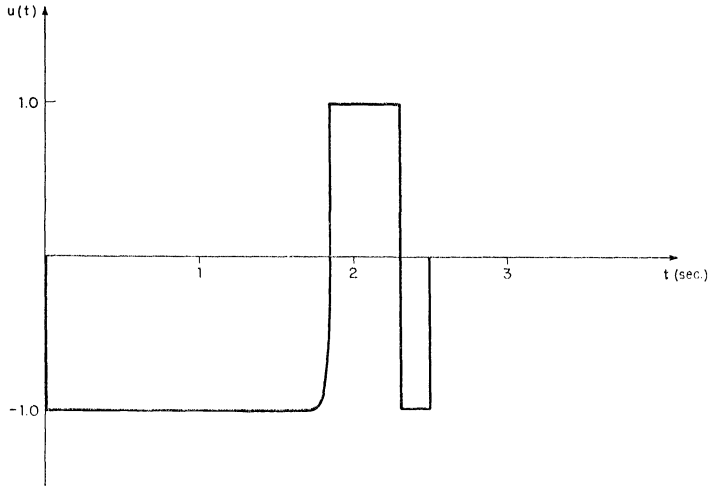

FIG. 6a. Optimal control u for the terminal time $t_1 = 2.5$ sec (control constraint)

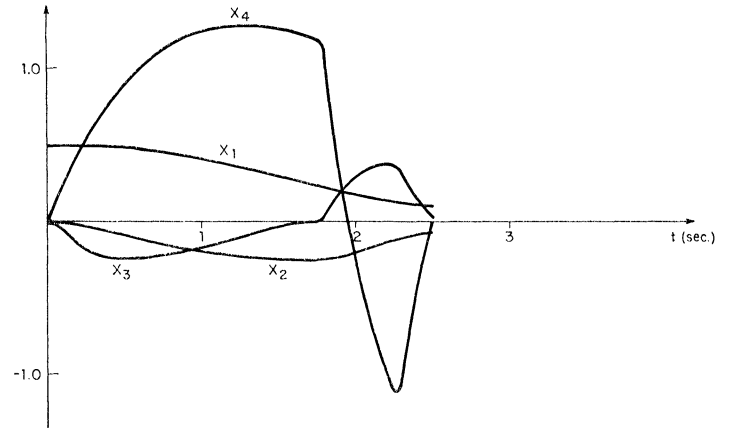

FIG. 6b. State vector components when the optimal control is applied (control constraint).

x_4, indeed exhibits the bang-bang property as predicted in section 4.

(2) *Nonlinear systems.* Optimal control of nonlinear systems with inequality constraints on the control or state variable have not been solved extensively at the time of the writing of this report*. For nonlinear systems without inequality constraints, Bryson and Kelley have reported numerous

* For more recent developments, see Ph.D. thesis by W. F. Denham, Harvard University, Division of Engineering and Applied Physics, June 1963.

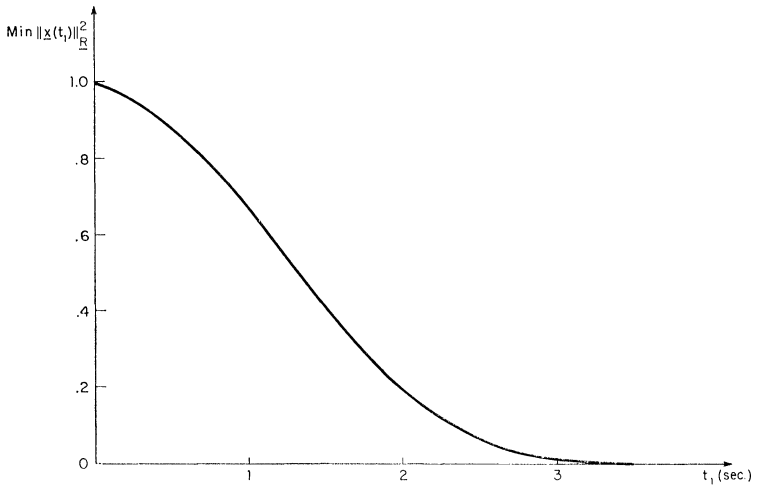

FIG. 7. Minimum terminal norm as a function of terminal time t_1 (constraint on x_4)

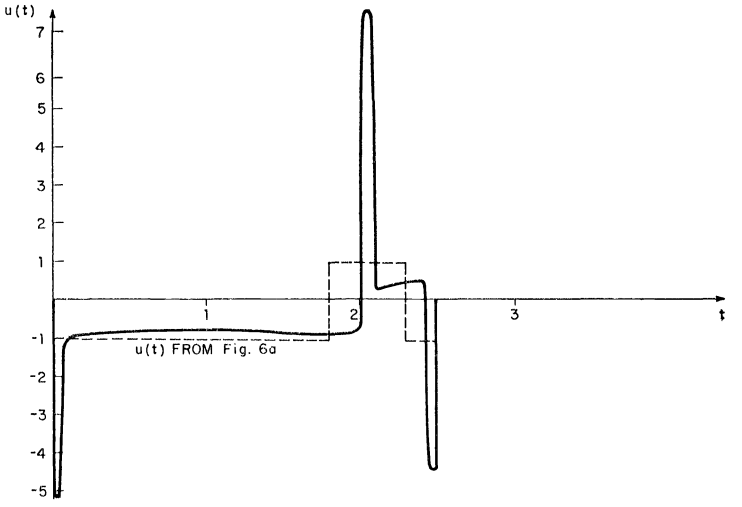

FIG. 8a. Optimal control u for the terminal time $t_1 = 2.5$ sec. (constraint on x_4).

computational examples in [1, 2, 3, and 4]. Kelley has recently demonstrated some examples with control variable constraints using essentially the approach outlined in section 4 [5]. Dreyfus has solved a simple nonlinear state variable constrained problem with no terminal constraint by another approach [7].

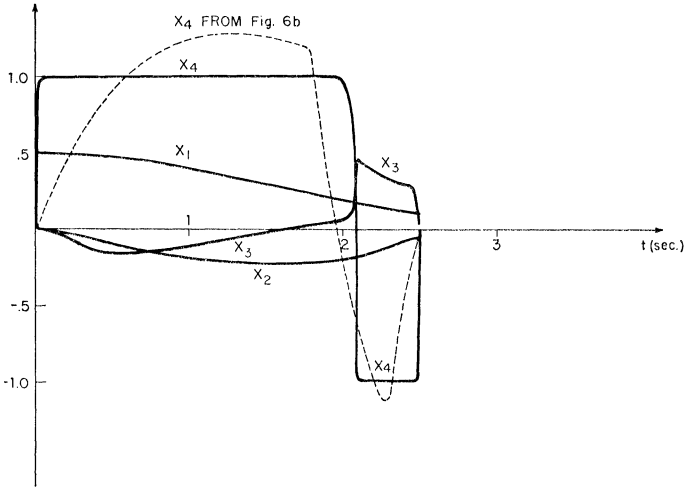

FIG. 8b. State vector components when the optimal control is applied (constraint on x_4).

The difficulty with writing general nonlinear programs is the problem of taking the partial derivatives required by (52). These have to be programmed separately for each given problem.

7. Conclusion and open problems. In the above sections, we have attempted to develop in a systematic manner a general method for the computational solution of complex optimal control problems. The derivation used is only one of the several possible ways in arriving at the solution. Others are the original variational approach used by Bryson and Kelley when they first developed the method, and the dynamic programming viewpoint taken by Dreyfus [6]. Our approach can also be interpreted in the light of the recent work of Rosen and Zoutendijk [12, 18] on nonlinear programming. However, here we take advantage of the properties of the dynamic system and the nature of the constraint to make the generalized gradient projection (cf. (23), (24), (30) and (31)) computationally much simpler. The present approach emphasizes the role of the inequality constraints and the discrete nature of the computational process, and regards the problem as one of nonlinear programming. In fact, to the best of the authors' knowledge, the computational solutions of the optimal control problem provided the first unifying interpretation to the three different disciplines, namely, variational calculus, dynamic programming, and nonlinear programming. For simple pedagogical reasons, the readers are urged to consult the above-listed references.

At the present stage of development of this method, there is still con-

siderable art associated with the application of this technique to complex problems. While there is little doubt concerning the validity of the general approach, ingenuity on the part of the user can still help the effectiveness of the method greatly. For one thing, one good initial guess to the optimal solution can speed the convergence very much.

A further question in this connection is how close does this technique allow one to approach the actual optimal solution? This is difficult to answer, because in many cases we do *not* know even the form of the actual optimal solution, particularly with nonlinear problems. Also, in the latter case, the method does not distinguish between a local and a global minimum, or worse yet, between a local minimum and an inflection point if no further assumptions are made on the nonlinear functions involved.

The method of successive approximation described in this paper is an infinite process. In practice, of course, we must terminate our computation within a finite number of iterations. How close does the computed solution approach the true solution thus further depends on the criterion for termination of the computation process, control of round-off error, and other numerical approximations. These are all subjects of numerical analysis and require separate treatment. At present, relatively little is known along these lines concerning this technique as applied to problems of optimal control. What one can say numerically is that "Whatever you can do, I can do better or at least just as good". This is often sufficient for practical purposes.

In the case of linear dynamical systems, our experience suggests the following qualitative answers: (1) Convergence to the *exact* optimal solution appears to be slow with the descent schemes of (22–25); (2) On the other hand, most of the descent was realized in the first ten to fifty iterations. Then a large number of iterations was used in obtaining fractions of percent of the total improvement; (3) The number of control steps employed does not seem to affect the rate of convergence in any significant way; and (4) The rate of convergence is dependent on the controllability of the dynamic system.

These remarks (1–4) are more or less what is to be expected in any gradient method of computation. It is believed that they apply also to nonlinear problems in the same qualitative fashion. However, it is also expected that we can utilize many of the known techniques (such as anti-zig-zag procedures) for the gradient computation and improve the convergence rate in calculating optimal controls.

For the same reasons it is also difficult at this stage to evaluate the various methods for obtaining an optimal control for nonlinear systems. Numerically, of course one can always compare two methods by solving a series of identical problems using both methods. This is probably what has

to be done if the effectiveness of the various versions of the method are to be compared.

Lastly, it should be pointed out that at the present stage of computer technology the gradient method described herein can solve effectively practical problems involving single in-flight constraints of the type (6) or (7). The problem of multiple constraints discussed in section 4, especially the case where there are more inequality constraints than control variables, cannot as yet be solved rapidly on the computer. In other words, it is not yet feasible to solve a general nonlinear program involving a very large number of variables.

8. Acknowledgment. The authors gratefully acknowledge the helpful suggestions of Drs. J. Wing, B. Whalen and S. Kodama of the University of California at Berkeley and Mr. J. R. Eckel of the University of Wisconsin in the writing of the paper.

The work reported in the paper, particularly the development of computer program and the experimental results, were performed under NASA Contract No. NASr-27. It was also supported in part by NONR Contract No. 1866(16) administered through Harvard University.

This article is published posthumously after the unfortunate early death of the second author in late 1962. The first named author would like to register his personal loss of a good friend and talented collaborator.

REFERENCES

[1] A. E. Bryson, et al, *Lift program that minimizes re-entry heating*, Journal of Aerospace Science 29 (1962), pp. 420–431.

[2] A. E. Bryson and W. Denham, *A steepest ascent method for solving optimal programming problems* J. Appl. Mech., 29, 2 (1962), pp 247–257.

[3] H. J. Kelley, *Gradient theory of optimal flight paths* J. Amer. Rocket Soc., 30 (1960) pp. 947–953.

[4] H. J. Kelley, *Method of gradient* Ch. 6 of *Optimization Techniques*, edited by G. Leitmann. Academic Press, 1962.

[5] H. J. Kelley, R. Kopp, and H. Moyer, *Successive approximation techniques for trajectory optimization* Proc. Inst. Aerospace Sci. Symposium on Vehicle Systems Optimization, Garden City, N. Y., Nov. 1961.

[6] S. Dreyfus, *Variational problems with inequality constraints* J. Math. Anal. Appl. 4 (1962) pp. 297–305.

[7] S. Dreyfus, *Numerical solution of variational problems*, J. Math. Anal. Appl., 5 (1962) pp. 30–45.

[8] Y. C. Ho, *A successive approximation technique for optimal control systems subject to input saturation*, J. Basic Engrg. Trans. A.S.M.E. Ser. D: 84 (1962) pp. 33–40.

[9] Y. C. Ho, *Computational procedure for optimal control problem with state variable constraint* J. Math. Anal. Appl., 5 (1962) pp. 216–224.

[10] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* McGraw Hill, New York, 1955, Ch. 3.

[11] H. KUHN AND A. W. TUCKER, *Nonlinear Programming*, Second Berkeley Symposium of Math. Statistics and Probability, University of California Press, Berkeley, 1951.

[12] J. B. ROSEN, *The gradient projection method for nonlinear programming, Part I*, J. Soc. Indust. Appl. Math. 8 (1960) pp. 181–217.

[13] J. V. BREAKWELL, J. SPEYER, AND A. E. BRYSON, JR. *Optimization and Control of Nonlinear Systems using the Second Variation* this Journal, Vol. 1, No. 2, pp. 193–223.

[14] R. E. KALMAN, Y. C. HO, AND K. S. NARENDRA, *Controllability of linear dynamic systems*, Contributions to Differential Equations Vol. 1, John Wiley, New York (1963).

[15] E. B. LEE, *Mathematical aspects of the minimum response time problem* Trans. I.R.E., Prof. Group Automatic Control, (1960) pp. 283–290.

[16] L. NEUSTADT, *Synthesizing time optimal controls* J. Math. Anal. Appl. 1 (1960) pp. 484–493.

[17] R. KALABA, *Dynamic programming* Ch. 9 of *Optimization Techniques*, edited by G. LEITMANN, Academic Press, 1962.

[18] G. ZOUTENDIJK, *Method of Feasible Directions*, Elsevier Pub. Co. 1961.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

A SOLUTION OF THE GODDARD PROBLEM*

BORIS GARFINKEL†

Abstract. The problem of optimizing the thrust of a vertically ascending rocket is solved here under the assumption of isothermal atmosphere in two important cases: 1) the jet Mach number and the fuel supply are sufficiently large; 2) the drag is a convex function of the velocity.

The first case embraces all physical drags and is valid for the Earth; the second extends to all atmospheres, but is restricted to drags that are fairly common.

With impulsive boosts in velocity admitted, the solution is shown to contain a finite number of such boosts in the sonic region of the rocket velocity, and to contain no coasting arcs except in the terminal stage.

An absolute minimum is proved with the aid of a *sufficient condition* applicable to problems of optimum control.

1. Introduction. The problem of maximizing the summit altitude of a vertically ascending rocket, of which the Goddard Problem (1919) is a variant, has received considerable attention in the literature. Despite the notable advance achieved by Tsien and Evans (1951), numerous gaps in the theory still remain to be filled. As has been pointed out by Leitman, Ross, et al., the problem continues to be beset by the difficulty arising from the requirement that the mass be monotone. Solutions that meet this requirement have been obtained only in a few very special cases, typified by the work of Miele (1958), who treated flight in vacuum and the power law of drag.

In the present paper, which is an outgrowth of the author's unpublished work of 1949, the class of soluble cases is considerably broadened. With the assumption of isothermal atmosphere and the admissibility of infinite thrust, a solution is obtained in the following two cases: (1) The jet Mach number and the fuel supply are sufficiently large; and (2) The drag is a convex function of the velocity. The first case is valid for the Earth; the second is restricted to a class of drags that are fairly common. The remaining case, where neither 1) nor 2) holds, is being left as a subject for future investigation.

A recapitulation of the relevant existing theory, designed to provide the necessary background for the current development, is incorporated in sections 2 and 5.

2. Formulation of the problem. The equation of motion of the rocket, subject to forces of gravity, drag, and thrust, is

$$\dot{m}c + m\dot{V} + \tfrac{1}{2}C_D(V, X)\rho(X)V^2S + mg(X) = 0, \tag{1}$$

* Received by the editors December 18, 1962.

† Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland.

where m is the mass, V the velocity of the rocket, C_D the drag-coefficient, X the altitude, ρ the density of the air, S the cross-section, g the acceleration of gravity, c the jet velocity, and the superscript dot indicates the differentiation with respect to the time.

We shall introduce the simplifying assumptions:

$$\begin{aligned} &\text{(i) } C_D \text{ is a function of } V \text{ only,} \\ &\text{(ii) } g = \text{const.,} \\ &\text{(iii) } \rho = \rho_0 \exp(-X/\ell), \quad \ell = \text{const.,} \end{aligned} \tag{2}$$

define the dimensionless parameters α, β by

$$\begin{aligned} &\alpha \equiv g\ell/c^2, \qquad \beta \equiv 2m_0 g/c^2\rho_0 S, \\ &0 < \alpha < \infty, \qquad 0 < \beta < \infty, \end{aligned} \tag{3}$$

where m_0 is the initial mass, and dimensionless variables x, v, ω, y, and f by

$$\begin{aligned} &x \equiv gX/c^2, \qquad v \equiv V/c, \qquad \omega \equiv \log m_0/m, \\ &y \equiv \omega - v - x/\alpha, \qquad f \equiv C_D v^2 e^v/\beta. \end{aligned} \tag{4}$$

Then (1) becomes

$$\begin{aligned} &\phi \equiv -y' + fe^y/v + 1/v - 1/\alpha = 0, \\ &x_0 \leqq x \leqq x_1, \end{aligned} \tag{5}$$

the prime indicating the differentiation with respect to x. The initial conditions are

$$x_0 = 0, \qquad v(0) = 0, \qquad y(0) = 0; \tag{6}$$

the terminal conditions are not specified.

The quantities m and $\dot{m}$ in (1) are bounded by the inequalities

$$m \geqq m_{\min}, \qquad 0 \leqq -\dot{m} < \infty, \tag{7}$$

if infinite thrust is admitted as a mathematical convenience. Such a thrust, operating for an infinitesimal time, produces a finite positive jump Δv, while y and $(\omega - v)$ remain continuous in virtue of (5) and (4). In terms of the new variables, (7) can be written

$$\begin{aligned} &\psi_1 \equiv \omega_{\max} - y - v - x/\alpha \geqq 0, \\ &\psi_2 \equiv y' + v' + 1/\alpha \geqq 0. \end{aligned} \tag{8}$$

The two unknown functions $y(x)$, $v(x)$ are connected by a differential constraint $\phi = 0$; the system therefore has one degree of freedom, which can be realized physically by a choice of an arbitrary $v(x)$, ideally regulated

by a servo-mechanism controlling the flow rate $\dot{m}$. Functions $y(x)$, $v(x)$ will be admissible if y, v, y' satisfy the constraints (5) and (8) with the initial conditions (6), and if they are continuous except at corners, where y' and v may be discontinuous with $\Delta v \geqq 0$. In the class of admissible functions we seek $v(x)$ that minimizes $-x_1$.

The problem is thus identified with the Problem of Mayer in the Calculus of Variations, complicated by the presence of algebraic and differential inequality constraints.

3. The auxiliary problem. The differential constraint $\psi_2 \geqq 0$, assuring the monotonicity of the mass, admits subarcs on which $\psi_2 = 0$ while $\psi_1 > 0$; i.e., the "burning" regime may be interrupted by the insertion of "coasting" subarcs. In order to avoid such complications let us consider an auxiliary problem characterized by the absence of the constraint $\psi_2 \geqq 0$. While such a formulation, used by Tsien et al., automatically eliminates the aforesaid complication, it creates another one by admitting $\psi_2 < 0$ and $\Delta v < 0$. Of course, negative fuel consumption is a physical absurdity! The resulting solution would not be of physical interest, were it not for the curious fact that such an occurrence is precluded in certain practical cases. Indeed, if the constraints are satisfied anyway in the form $\psi_2 > 0$, $\Delta v \geqq 0$, it is clear that the auxiliary and the actual problems have the same solutions. In particular, that such is the situation in both cases treated here will be shown in Theorems 1 and 2 of section 12. In terms of the quantities α and v, the two cases can be respectively characterized by:

(i) α is sufficiently small; $\omega_{\max}$ is sufficiently large;

(ii) $C_D v^2$ is convex.

Accordingly, we shall attack the *auxiliary problem,* which is in the standard form of the *problem of optimum control*: *We seek a function* $u(x)$ *satisfying*

$$\phi = -y' + g(x, y, u) = 0, \qquad x_0 \leqq x \leqq x_1, \tag{9}$$

subject to the boundary conditions and the inequalities:

$$x_0 = a, \qquad y(x_0) = b, \qquad \Phi(x_1, y(x_1)) = 0, \qquad \psi(x, y, u) \geqq 0, \tag{10}$$

and minimizing some prescribed function

$$G(x_1, y(x_1)). \tag{11}$$

Here y, u, Φ, ψ are vectors of n, m, p, r components respectively, with $p < n + 1$. In our problem $n = m = r = 1$, $p = 0$; $G = -x_1$, $u = v$, and

$$
\begin{aligned}
g &\equiv (fe^y + 1)/v - 1/\alpha, \\
\psi &\equiv \omega_{\max} - y - v - x/\alpha = \psi_1 .
\end{aligned} \tag{12}
$$

Since v' has disappeared from the problem, v has assumed the role of a "control" variable, which enters $g(x, y, v)$ non-linearly. That the problem is non-singular is shown in section 8; the solution is obtained in sections 5–14 by the application of the *necessary conditions* I–IV and the *fundamental sufficiency condition* of Weierstrass. The first one is the *multiplier rule*, comprising the Euler, the *convexity*, the *transversality*, and the *corner conditions*, treated respectively in sections 5, 13, 6, and 7.

The existence and the character of the solution intimately depend on the nature of the drag coefficient $C_D(v)$, which is the subject of the next section.

4. Some properties of the drag. We shall assume the usual positiveness and the continuity of $C_D(v)$, the monotonicity of the drag,

$$
\frac{d}{dv}(C_D v^2) > 0, \tag{13}
$$

and the asymptotic expansions

$$
\begin{aligned}
C_D &= A_0 + A_1 v + A_2 v^2 + \cdots \qquad \text{as } v \to 0, \\
C_D &= B_0 + B_1/v + B_2/v^2 + \cdots \qquad \text{as } v \to \infty, \\
A_i &\geqq 0, \qquad B_i \geqq 0, \qquad i = 0, 1, \cdots \infty .
\end{aligned} \tag{14}
$$

Then the logarithmic derivatives k and k' defined by

$$
k \equiv d \log C_D / d \log v, \qquad k' \equiv dk/d \log v \tag{15}
$$

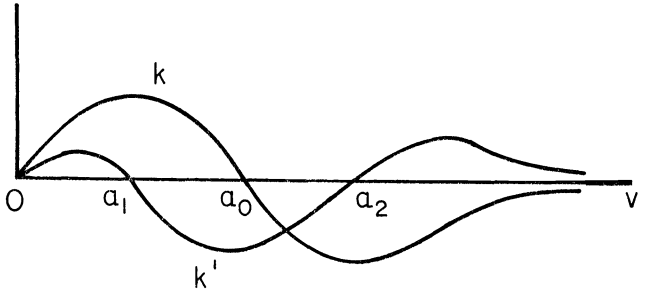

FIG. 1. Logarithmic derivatives $k(v)$ and $k'(v)$

have the properties:

$$
\begin{aligned}
&k(0) = k(\infty) = k'(0) = k'(\infty) = 0,\\
&k(0+) > 0, \qquad k(\infty -) < 0,\\
&k'(0+) > 0, \qquad k'(\infty -) > 0,\\
&k + 2 > 0.
\end{aligned} \tag{16}
$$

Furthermore, let $C_D(v)$ have a single maximum at, say a_0. Then k has a maximum at a_1, a zero at a_0, and a minimum at a_2, while k' has zeros at a_1 and a_2. It follows, in view of (16), that

$$
\begin{aligned}
&(a_0 - v)k > 0, \qquad (v - a_1)(v - a_2)k' > 0,\\
&0 < a_1 < a_0 < a_2 < \infty.
\end{aligned} \tag{17}
$$

In the analysis, the function $f(v)$, defined in (4), and the derived functions $H(v)$, $h(v)$, defined herewith, will be extremely useful:

$$
\begin{aligned}
f &\equiv C_D(v)v^2e^v/\beta > 0,\\
H &\equiv vf_v - f,\\
h &\equiv H - \alpha f_v = (v - \alpha)f_v - f,
\end{aligned} \tag{18}
$$

with literal subscripts denoting the argument of differentiation.

In terms of k and k', these functions and their derivatives can be exhibited as follows:

$$
\begin{aligned}
f_v &= (f/v)(2 + v + k) > 0,\\
f_{vv} &= (f/v^2)[(2 + v + k)(1 + v + k) + v + k'],\\
H &= f(1 + v + k),\\
H_v &= vf_{vv},\\
h &= f[(1 - \alpha/v)(2 + v + k) - 1],\\
h_v &= (v - \alpha)f_{vv}.
\end{aligned} \tag{19}
$$

Special properties of these functions, obtained with the aid of (16), are tabulated below:

$$
\begin{aligned}
&f(0) = f_v(0) = H(0) = H_v(0) = h(0) = 0,\\
&f(\infty) = f_v(\infty) = f_{vv}(\infty) = H(\infty) = h(\infty) = \infty,\\
&\tfrac{1}{2}f_{vv}(0) = C_D(0)/\beta > 0,\\
&h_v(0) = -\alpha f_{vv}(0) < 0,\\
&2f = f_{vv}(0)v^2 + \cdots \quad \text{as } v \to 0,\\
&2H = f_{vv}(0)v^2 + \cdots \quad \text{as } v \to 0.
\end{aligned} \tag{20}
$$

5. The Euler equations. Since the Lagrangian function of the *auxiliary problem* is $F = \lambda\phi + \mu\psi$, the extremals must satisfy the equations

$$
\begin{aligned}
&y' = g(x, y, u),\\
&\lambda' + \lambda g_y + \mu\psi_y = 0,\\
&\lambda g_u + \mu\psi_u = 0,\\
&\mu\psi = 0, \qquad \psi \geqq 0,
\end{aligned}
\tag{21}
$$

where $\lambda(x)$, $\mu(x)$ are Lagrange multipliers. The substitution from (12) into the Euler equations (21.2) and (21.3) now yields, in view of (18.2),

$$
\begin{aligned}
&\lambda' + \lambda f e^y/v - \mu = 0,\\
&(\lambda/v^2)(He^y - 1) - \mu = 0,
\end{aligned}
\tag{22}
$$

leading to

$$
\lambda = \lambda(0) \exp \int_0^x [(H - vf)e^y - 1]\, dx/v^2. \tag{23}
$$

The use of the "switching function" $\mu(x)$ permits simultaneous consideration of subarcs lying in the region $\psi_1 > 0$, where $\mu = 0$, and of subarcs lying in the boundary $\psi_1 = 0$, where $\mu \neq 0$. Three regimes are distinguished, designated by I, B, and C:

I, Impulsive thrust, $\Delta v \neq 0$,

B, Burning, $\quad \psi_1 > 0, \quad \mu = 0$,

C, Coasting, $\quad \psi_1 = 0, \quad \mu \neq 0$.

An extremal is compounded of a B-subarc, with impulsive thrusts I occurring at a finite number of points, and a C-subarc appearing in the terminal stage only.

During the burning stage $\mu = 0$, and the "optimality" condition (22.2) yields

$$
e^y H(v) = 1. \tag{24}
$$

That a solution $v(y)$ of (24) exists follows from (20), which gives the range of H as $(0, \infty)$; that this solution is unique will be shown in section 9, with the aid of Condition II. Several conclusions can now be drawn. First, (8.1) implies $y < \infty$; then from (24) and (18) there follows

$$
H > 0, \qquad v \neq 0, \tag{25}
$$

and therefore $v > 0$. Since the initial value $v(0) = 0$ violates the requirement (25), the burning stage must be preceded by an impulsive launching with

a velocity v_0 that satisfies (24) with the initial condition $y(0) = 0$; i.e.,

$$H(v_0) = 1. \tag{26}$$

The initial discontinuity is thus specified by $v_-(0) = 0$, $v_+(0) = v_0$, and $\Delta\omega = \Delta v$. Of historical interest is the value of the gravity-drag ratio mg/D, which equals $1/fe^y$ in our symbols, and is optimized by (24) into H/f, or

$$mg/D = 1 + v + k. \tag{27}$$

The solution of the Euler equations is obtained by the differentiation of (24) with respect to x, yielding

$$-Hy' = H_v v', \tag{28}$$

followed by the substitution from (24) into (5), which in view of (18) now becomes

$$-\alpha H y' = h. \tag{29}$$

Then, from (28) and (29),

$$v' = h/\alpha H_v, \tag{30}$$

and, provided $h \not\equiv 0$, $v(x)$ is obtained by the inversion of the quadrature

$$\begin{aligned} x/\alpha &= \int_{v_0}^{v} dH/h \\ &= \chi(v) - \chi(v_0), \end{aligned} \tag{31}$$

where

$$\chi(v) \equiv \int_1^v dH/h \tag{32}$$

defines a "rocket function" dependent only on the form of $C_D(v)$ and on the value of the parameter α. The equation of the extremal subarc now appears in the parametric form $x = x(v)$, $y = y(v)$, in consequence of (24) and (31). The special case $h = 0$ is solved in section 11.

During the coasting stage $\psi_1 = 0$, $\mu \not\equiv 0$, and (5) becomes

$$vv' + f(v) \exp\left(-v - \frac{x}{\alpha} + \omega_{\max}\right) + 1 = 0 \tag{33}$$

with the initial conditions corresponding to the "burnout", i.e., the solution of the equation $\psi_1 = 0$ with $y(v)$ and $x(v)$ furnished by (24) and (31).

That the *convexity conditioo* $\mu \leqq 0$ is satisfied on the C-subarc will be proved in Theorem 3 of section 13.

6. The transversality condition. In the control problem of section 3 the relation

$$[(G_x + \lambda \cdot g)\, dx + (G_y - \lambda) \cdot dy]_{x_1} = 0 \tag{34}$$

must hold for all dx and dy satisfying the differentiated equation $\Phi = 0$, the dot placed between vectors indicating their inner product. In the *auxiliary problem* $n = 1$, $G = -x_1$, and $p = 0$, so that dx and dy are arbitrary, and (34) reduces to

$$\begin{aligned} -1 + \lambda g &= 0, \\ \lambda &= 0 \end{aligned} \tag{35}$$

at $x = x_1$. Three conclusions can be drawn. First, since both λ and μ cannot vanish simultaneously, $\mu(x_1) \neq 0$, so that $\psi_1(x_1) = 0$. Second, noting that $|g(x_1)| = \infty$ from (35), and recalling that $y < \infty$, $\alpha > 0$, $f < \infty$, we deduce from (12) that $v(x_1) = 0$. Of course, both conclusions are physically obvious: x_1 must be reached with zero velocity after coasting with fuel consumed. The third conclusion,

$$\lambda(x_1 - 0) > 0 \tag{36}$$

follows from the observation that: 1) As $v \to 0$ the asymptotic value of g is $g \sim 1/v > 0$, in view of (12), (20), and (25), and that 2) $\lim (\lambda g) = 1$ as $x \to x_1$, in view of (35). Now, since $\lambda(x)$ cannot change its sign in virtue of (23), the inequality (36) implies

$$\lambda(x) > 0 \quad \text{for} \quad x_0 \leqq x < x_1, \tag{37}$$

which requirement can be satisfied by choosing

$$\lambda(0) = 1. \tag{38}$$

For future use, we note the following asymptotic values as $v \to 0$:

$$g \sim 1/v, \qquad \lambda \sim v, \qquad \lambda' \sim -1/v, \qquad \mu \sim -1/v, \tag{39}$$

which can now be obtained from (35), (22), and (20).

The existence and the continuity of the multipliers $\lambda(x)$, $\mu(x)$, required by the *multiplier rule*, are now assured between corners of a minimizing arc.

7. The corner condition. At a "free" corner, the relations

$$\Delta(\lambda \cdot g) = 0, \qquad \Delta\lambda = 0 \tag{40}$$

must hold, with Δ denoting a jump. Noting that in our problem, with $n = 1$, (40) implies $\Delta g = \Delta y' = 0$, and recalling that $\Delta y = 0$ in virtue of (5), we deduce from (24), (29), and (18) that

$$\Delta H = \Delta h = \Delta f_v = 0 \tag{41}$$

on a B-subarc. The definition of H now implies

$$\frac{\Delta f}{\Delta v} = f_v\,, \qquad \Delta f_v = 0, \tag{42}$$

from which the transition values v_- and v_+ can be determined.

Since $f_{vv}(0) > 0$ and $f_{vv}(\infty) > 0$, according to (20), it follows that f_{vv} has an even number of zeros, I_i, $i = 1, 2, \cdots, 2N$. Two types of drag function are of practical importance:

$$\begin{aligned} &\textit{Type } 1 && N = 0, \\ &&& f_{vv} > 0, \end{aligned} \tag{43}$$

for which (42) has only the trivial solution $\Delta v = 0$, so that no corner occurs.

$$\begin{aligned} &\textit{Type } 2 && N = 1, \\ &&& (v - I_1)(v - I_2) f_{vv} > 0. \end{aligned} \tag{44}$$

Then there exists a line tangent to the curve $f(v)$ at two points v_1 and v_2 (Fig. 2 and 3), which satisfy the Corner Condition (42), and the inequality

$$v_1 < I_1 < I_2 < v_2\,. \tag{45}$$

The transition values at the corner are then determined as follows:

$$\begin{aligned} &v_- = v_1\,, \quad v_+ = v_2 \quad \text{if} \quad v_1' > 0 \\ &v_- = v_2\,, \quad v_+ = v_1 \quad \text{if} \quad v_2' < 0 \end{aligned} \tag{46}$$

It will be shown in section 9 that such a jump in velocity is required by the Weierstrass Condition II whenever the *corner condition* is satisfied.

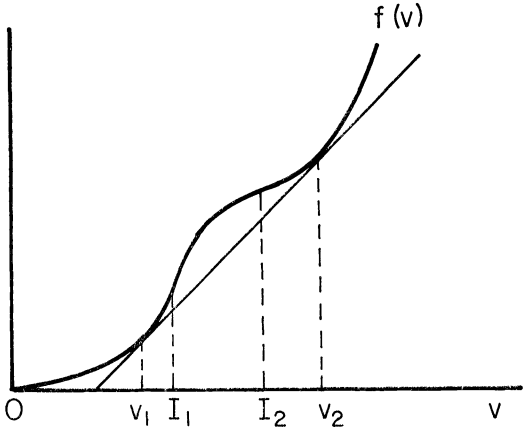

FIG. 2. Double tangent and points of inflection

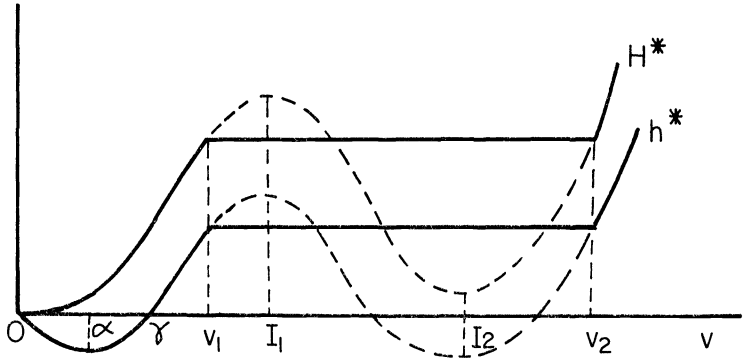

FIG. 3. $H(v)$, $h(v)$ - - - - -
$H^*(v)$, $h^*(v)$ ——

The results of the last paragraph can be easily generalized for any N, with (44) replaced by

$$f_{vv}\prod_{1}^{2N}(v - I_i) > 0, \tag{47}$$

there being a velocity jump for each one of the N double tangents.

At the junction of the B and C-subarcs the *corner condition* is satisfied with

$$\Delta v = \Delta y = \Delta y' = \Delta\lambda = \Delta\mu = 0, \qquad \Delta v' < 0. \tag{48}$$

8. The Hilbert condition. With $n = m = r = 1$, the four unknown functions $y(x)$, $\lambda(x)$, $u(x)$, $\mu(x)$ are related by the four equations (21). The highest derivatives being (y', λ', u, μ), the non-vanishing of the Jacobian determinant is the Hilbert Condition

$$\begin{vmatrix} F_{uu} & \psi_u \\ \mu\psi_u & \psi \end{vmatrix} \neq 0, \tag{49}$$

or

$$|F_{uu}| \neq 0 \quad \text{if} \quad \psi > 0. \tag{50}$$

Here F is the Lagrangian function, and F_{uu} is generally an $m \times m$ matrix. The condition assures the existence of the highest derivatives listed above, as well as their piecewise continuity of class C^{k-2} if g and ψ are of class C^k, and is a direct consequence of the Legendre Condition III′.

In our problem (50) becomes

$$\frac{\lambda f_{vv}}{vH} \neq 0, \tag{51}$$

and, since λ, v, and H are positive by (25) and (37),

$$f_{vv} \neq 0. \tag{52}$$

Provided this requirement is met on the B-subarc, the Hilbert Condition is satisfied, and since g is analytic in our problem, the functions (y, λ, v, μ) are analytic between corners. It is noteworthy that the use of the velocity v as a control variable, in place of the thrust $v\omega'$, removes the apparent singularity of the original problem.

9. Conditions of Legendre and Weierstrass. The necessary conditions III and II, modified by the inclusion of the control variable u among the set of slope functions, can be written for the one dimensional case, $n = m = 1$, as

$$\begin{aligned} [\lambda g_{uu}(x, y, u) + \mu\psi_{uu}]\delta u^2 &\geqq 0, \\ E \equiv \lambda[g(x, y, \bar{u}) - g(x, y, u)] &\geqq 0, \end{aligned} \tag{53}$$

for all (x, y, u, λ, μ) belonging to the minimizing extremal, and for all $\bar{u} \neq u$ and satisfying $\phi = 0$. In our problem (53), with the aid of (12) and (24), reduces to

$$\begin{aligned} \frac{\lambda f_{vv}}{vH} &\geqq 0, \\ \left(\frac{\lambda}{\bar{v}H}\right)[\bar{f} - f - (\bar{v} - v)f_v] &\geqq 0, \end{aligned} \tag{54}$$

where $\bar{f} \equiv f(\bar{v})$, and finally, since v, H, and λ are positive, to the requirements that

$$\begin{aligned} f_{vv} &\geqq 0, \\ \bar{f} - f &\geqq (\bar{v} - v)f_v \end{aligned} \tag{55}$$

hold on every B-subarc.

In the language of geometry, (55) implies that v must be restricted to the domain where the tangent to the curve $f(v)$ lies entirely below the curve. For drag of type 1, (55) is automatically satisfied; for drag of type 2, where (44) and (45) hold, (55) is equivalent to the requirement

$$(v - v_1)(v - v_2) \geqq 0, \tag{56}$$

where v_1 and v_2 are the points of contact of the double tangent. The exclusion of the interval (v_1, v_2) from the B-subarc then demands that a corner occur when v reaches the values v_1 or v_2, as described in (46). Conversely, the occurrence of such a corner satisfies the requirement $E \geqq 0$, the equality holding only at corners for $v = v_1$, $\bar{v} = v_2$, and conversely. Consequently, $\int E\,dx > 0$ on the B-subarc, to which we shall refer as Condition II*.

That this condition is also satisfied on the C-subarc will be proved in Theorem 3 of section 13.

A fortiori, the strengthened Legendre-Clebsch Condition III′; i.e.,

$$f_{vv} > 0 \tag{57}$$

also holds, from which two consequences follow. First, (57) establishes the Hilbert Condition (52); second, with the aid of (19.4) it implies that $H_v > 0$. We conclude that $H(v)$ is monotonic in the domain defined by (56), and has an inverse H^{-1}, thus assuring the uniqueness of the solution $v(y)$ of the equation (24). In view of this fact, it is convenient to replace H and h in all the equations referring to the B-subarc by H^* and h^* (see Fig. 3) defined by

$$\begin{aligned} &H^* \equiv H, \qquad h^* \equiv h && \text{if} \quad (v - v_1)(v - v_2) \geqq 0, \\ &H^* \equiv H(v_1), \qquad h^* \equiv h(v_1) && \text{if} \quad (v - v_1)(v - v_2) \leqq 0. \end{aligned} \tag{58}$$

In the future, if no confusion results, the asterisks of H and h will be dropped.

10. The Jacobi condition. The second variation in the control problem defined by (9)–(11) with $n = m = r = 1$, $p = 0$ can be written

$$\begin{aligned} d^2J = {} & [(F_x - y'F_y)\,dx^2 + 2F_y\,dy\,dx + d^2G]_{x_1} \\ & + \int_{x_0}^{x_1} (F_{yy}\,\delta y^2 + 2F_{yv}\,\delta y \delta v + F_{vv}\,\delta v^2)\,dx, \end{aligned} \tag{59}$$

where F is defined by $F = \lambda\phi + \mu\psi$, $\phi = -y' + g(x, y, v)$, $\mu\psi = 0$, $\psi \geqq 0$. The necessary Jacobi Condition is that

$$d^2J \geqq 0 \tag{60}$$

must hold on a minimizing extremal for all dx, dy, and for all δy, δv satisfying the differentiated equations $\phi = 0$, $\mu\psi = 0$.

Observe that $F_y = -\lambda'$ from (21.2), and that our problem possesses the following special features: 1) $G = -x_1$, so that $d^2G = 0$; 2) $F_x = -\mu/\alpha$ in view of (12); 3) $u = v$, and $F_{vv} = \lambda g_{vv}$, since ψ_1 is linear in v; 4) At $x = \xi - 0$ the relation $d\psi = \delta\psi + \psi'\,dx = 0$ holds. Since $\psi' = -\psi_2 < 0$ from section 12, the burnout point $\bar{\xi} = \xi + d\xi$ of a "weak" comparison curve lies in an arbitrarily small neighborhood of ξ. Clearly, $\delta\psi = 0$ for $x > \bar{\xi}$, and $\mu\delta\psi = 0$ except on the interval $(\xi, \xi + d\xi)$, whose contribution to (59) can be neglected. With this exception, δy and δv must satisfy

$$\begin{aligned} -\delta y' + g_y\delta y + g_v\delta v &= 0, \\ \mu(\delta y + \delta v) &= 0. \end{aligned} \tag{61}$$

Furthermore, on the B-subarc, $x_0 \leqq x \leqq \xi$, where ξ is the "burnout" point, observe that $\mu = 0$, $g_v = 0$ by (21.3), and the initial condition $y(0) = 0$ implies $\delta y(0) = 0$, and hence the solution of (61) is $\delta y \equiv 0$, δv arbitrary. On the other hand, on the C-subarc, $\xi \leqq x \leqq x_1$, $\mu \neq 0$, and the continuity of δy in (61) implies $\delta y(\xi + 0) = \delta y(\xi - 0) = 0$. Hence the solution of (61) is $\delta y = \delta v \equiv 0$. Now, since $\delta y = 0$ everywhere, $dy = y'\,dx$, and (59) becomes

$$d^2J = \left[\left(\frac{-\mu}{\alpha} - \lambda' g\right) dx^2\right]_{x_1} + \int_{x_0}^{\xi} \lambda g_{vv}\, \delta v^2. \tag{62}$$

Finally, at $x = x_1$ (39) yields $\mu \sim -1/v$ and $\lambda' g \sim -1/v^2$ as $v \to 0$; on the B-subarc $\lambda g_{vv} > 0$ by (53.1) and (57). Since α and v are positive, we conclude that the Jacobi Condition is satisfied in its strengthened form IV′, $d^2J > 0$.

An immediate consequence is that in our problem, with $n = 1$, IV′ assures the existence of a field. Let a family of extremals $y(x, \theta)$ be defined by (9)–(11) with the initial condition (10.1) replaced by

$$x_0 = a, \qquad y(x_0) = b + \theta, \tag{63}$$

where θ is a family parameter. That the region bounded by $x = a$, $\Phi(x, y) = 0$ is, indeed, a field follows from the following considerations: 1) The extension of IV′ to $\theta \neq 0$ is trivial; 2) IV′ assures the simple covering of the region; i.e., the existence of the function $\theta(x, y)$, and hence of the slope functions $u(x, y)$ and multipliers $\lambda(x, y)$, $\mu(x, y)$; 3) in a one-dimensional problem, the Euler equations suffice to assure that the Hilbert integral is independent of the path.

The conclusion of the last paragraph is not affected by the special circumstance $p = 0$ of our problem. In the absence of a prescribed terminal boundary $\Phi = 0$, a natural boundary of the field is furnished by the relation $y + x/\alpha = \omega_{max}$, which follows from $v(x_1) = 0$ and $\psi(x_1) = 0$.

11. The steady states of motion. Aside from their intrinsic interest, the lemmas of this section are required in the proof of the basic theorems of section 12.

LEMMA 1. *The function $h^*(v)$ has one and only one positive zero, γ.*

Proof. The proof proceeds from (18), (19), (20), (57), (58). Two cases are distinguished:

Case 1. α is outside (v_1, v_2) (See Fig. 3.) Then the relations $h = (v - \alpha)f_v - f$, $h_v = (v - \alpha)f_{vv}$, $f_{vv} > 0$ imply that h^* has one and only one stationary point,

$$\min h^* = h(\alpha) = -f(\alpha) < 0. \tag{64}$$

Since the minimum is negative, the relations $h^*(0) = 0$, $h^*(\infty) = \infty$, and the continuity of h^* imply the conclusion of the lemma, with

$$\begin{gathered} \alpha < \gamma < v_1 \quad \text{or} \quad v_2 < \alpha < \gamma, \\ (v - \gamma)h^* > 0. \end{gathered} \tag{65}$$

Case 2. α is inside (v_1, v_2). Then h^*, stationary on the interval (v_1, v_2), attains there a minimum, $\min h^* = h(v_1)$. That the minimum is again negative is implied by $h(0) = 0$, $h(\alpha) < 0$, and $0 < v_1 < \alpha$; finally, $h(v_1) = h(v_2) < 0$ and $h^*(\infty) = \infty$ imply the conclusion of the lemma, with

$$\begin{gathered} v_1 < \alpha < v_2 < \gamma, \\ (v - \gamma)h^* > 0. \end{gathered} \tag{66}$$

LEMMA 2. *On the burning subarc of an extremal, the acceleration of the rocket cannot change its sign.*

Proof. The proof proceeds from (30) and (19), leading to

$$\begin{gathered} v' = \frac{h}{\alpha H_v}, \qquad h_v = \left(1 - \frac{\alpha}{v}\right) H_v, \\ h = h_0 \exp \int_0^x \left(\frac{1}{\alpha} - \frac{1}{v}\right) dx. \end{gathered} \tag{67}$$

Noting that $\alpha > 0$, and that $H_v > 0$ by (19), (25), and (57), we conclude that

$$\operatorname{sgn} v' = \operatorname{sgn} h^* = \operatorname{sgn} h_0^*. \tag{68}$$

THE COROLLARY, $h_0 = 0$ *implies* $h(x) \equiv 0$ *and* $v \equiv \gamma$, follows immediately. Three types of trajectory are thus distinguished:

a) $v_0 < \gamma$, $h(x) < 0$, $v' < 0$ deceleration

b) $v_0 > \gamma$, $h(x) > 0$, $v' > 0$ acceleration

c) $v_0 = \gamma$, $h(x) \equiv 0$, $v \equiv \gamma$ steady state.

In case c) the solution (24) and (31), of the Euler equations, must be replaced by $y \equiv 0$, $v \equiv \gamma$.

12. The basic theorems. Having constructed the solution of the *Auxiliary Problem*, we shall show that under the assumptions of Theorems 1 or 2 it satisfies the constraints $\psi_2 > 0$ and $\Delta v \geqq 0$ on the B-subarc.

THEOREM 1. $2\alpha < \min(a_1, v_1)$ *implies*: 1) $\alpha < \gamma < 2\alpha$, 2) $\psi_2 > 0$, 3) $\Delta v \geqq 0$, *with the last two relations holding on the B-subarc of an extremal of the Auxiliary Problem.*

Proof. To prove 1), observe that from (18.3)

$$\begin{aligned} h(\alpha) &= -f(\alpha), \\ h(2\alpha) &= f(2\alpha)[\alpha + \tfrac{1}{2}k(2\alpha)], \end{aligned} \tag{69}$$

and that the hypothesis and (17) imply

$$\begin{aligned} 0 < 2\alpha < a_1 < a_0\,, \\ k(2\alpha) > 0. \end{aligned} \tag{70}$$

Then (69) and $f > 0$ imply $h(\alpha) < 0$ and $h(2\alpha) < 0$. Furthermore, from the hypothesis, $\alpha < 2\alpha < v_1 < v_2$, so that $h = h^*$ on $(\alpha, 2\alpha)$, in view of (58). The conclusion is now implied by the continuity of h^* and by Lemma 1.

To prove 2), observe that ψ_2, defined by (8.1) as

$$\psi_2 \equiv \omega' = y' + v' + \frac{1}{\alpha}, \tag{71}$$

can be exhibited as a function of v, with the aid of (29), (30), (18), (19), in two alternate forms:

$$\psi_2(v) = \frac{h}{\alpha H_v} + \frac{f_v}{H} = \frac{H}{\alpha H_v} + \left(\frac{f_v}{vH_v}\right)\left[2 + k + \frac{(v + k')}{(1 + v + k)}\right]. \tag{72}$$

The positiveness of $v, f, H, f_v, H_v, 2 + k, 1 + v + k$ is assured by (16), (18), (19), (20), and (57). There are two possibilities: Either $h \geqq 0$ or $h < 0$. If $h \geqq 0$, then $\psi_2 > 0$ in the first line of (72). On the other hand, if $h < 0$, then $v < \gamma$ in (65); the previous conclusion, $\gamma < 2\alpha$, and the hypothesis, $2\alpha < a_1$, imply $v < a_1$; then (17) implies $k' > 0$, and finally $\psi_2 > 0$ in the last line of (72).

To prove 3), recall that $v_1 > 2\alpha > \gamma$, and that by Lemma 2, $v > \gamma$ implies $h > 0$, $v' > 0$, and conversely. Then note, with the aid of (41), that $h(v_1) = h(v_2) > 0$. Therefore $v' > 0$ if $v = v_1$ or $v = v_2$, thus excluding the possibility $\Delta v < 0$ in the second line of (46).

For rough practical purposes, the hypothesis of the theorem may be replaced by

$$\kappa M > 4, \tag{73}$$

where M is the jet Mach number, and κ is the ratio of the atmospheric specific heats. To derive this result observe that: 1) $\alpha \equiv g\ell/c^2 = 1/\kappa M^2$ by the law of perfect gas and the formula for the sonic velocity, 2) the constants a_1 and v_1, which are the values of v at the maximum of $k(v)$ and at the first point of contact of the double tangent to $f(v)$, respectively, lie in the sonic region, 3) the sonic velocity corresponds to $a_0 \sim 1/M$,

and both a_1 and v_1 are generally sufficiently near a_0 to justify the inequalities $a_1/a_0 > 1/2$, $v_1/a_0 > 1/2$. Thus, (73) implies $2\alpha < a_1$ and $2\alpha < v_1$.

For the Earth, with $c \sim 2000\ m/s$, $g \sim 9.8\ m/s^2$, $\ell \sim 8000\ m$, $\kappa = 1.4$, we calculate

$$\begin{gathered} \alpha \sim 0.02, \qquad M \sim 6, \\ \kappa M \sim 8.4, \end{gathered} \tag{74}$$

concluding that the hypothesis of the theorem is satisfied for terrestrial rockets.

The vacuum case, $\rho \equiv 0$, solved by Miele, corresponds to $\alpha = 0$ and is, therefore, a subcase of the theorem. On the other hand, the constant-density atmosphere, $\rho = \text{const.}$, corresponds to $\alpha = \infty$ and hence lies outside the scope of Theorem 1. Indeed, Leitman succeeded in solving this case only by invoking the quadratic law of drag, $C_D = \text{const.}$, bringing the problem within the scope of Theorem 2.

THEOREM 2. *If $C_D v^2$ is convex, then $\psi_2 > 0$ and $\Delta v \geqq 0$ on the B-subarc of an extremal of the Auxiliary Problem.*

In the proof, note that the hypothesis, in view of (15), implies

$$(k + 1)(k + 2) + k' > 0, \tag{75}$$

and, consequently, $\psi_2 > 0$ in the last line of (72). Furthermore, observe that (75), (16.4), and (19.2) imply the convexity of $f(v)$. Hence the drag belongs to type 1 of section 7, with no corners on the B-subarc, and with $\Delta v = 0$.

The cases of quadratic law of drag and, the more general, power law of drag, also treated by Miele, appear as subcases of the theorem.

13. The coasting arc. The Weierstrass Condition $E \geqq 0$ and the Convexity Condition $\mu \leqq 0$ will be established on the coasting arc under the assumption that $\omega_{\max}$ is suitably restricted. The proof in Theorem 3 involves two definitions and a lemma.

For drag of type 2, define a constant V_2 (see Fig. 3) by

$$H(V_2) = H(I_1), \qquad V_2 \neq I_1, \tag{76}$$

and a function $u(y)$ as the least root of (24),

$$H(u)e^y = 1. \tag{77}$$

Note that, in view of Theorem 1,

$$V_2 > v_2 > I_1 > v_1 > 2\alpha > \gamma, \tag{78}$$

and that u is confined to the domain

$$u < I_1, \qquad u \geqq V_2, \tag{79}$$

where

$$H > 0, \qquad H_v > 0. \tag{80}$$

Furthermore, observe that the function $g(y, v)$ is minimized with respect to v at $v = u$, since

$$g_v(y, u) = 0, \qquad g_{vv}(y, u) > 0, \tag{81}$$

and that it has no other stationary points for $v < u$. An immediate consequence is that

$$g_v(y, v) < 0 \quad \text{if} \quad v < u. \tag{82}$$

Hereafter we shall use the notation $u(y(x)) = u(x)$.

The function u is not to be confused with the optimum velocity v on the B-subarc, defined by

$$H^*(v)e^y = 1, \tag{83}$$

and confined to the domain of (56). Note that $v(x) \geqq u(x)$ on the B-subarc, the inequality holding only for $v_2 < v < V_2$.

LEMMA 3: "*If* $[v(\xi) - v_1][v(\xi) - V_2] \geqq 0$, *then* $v(\xi) \leqq u(\xi)$ *and* $v(x) < u(x)$ *on* (ξ, x_1)."

Proof: If ξ belongs to a B-subarc, then the hypothesis and the preceding paragraph imply $v(\xi) = u(\xi)$. The special case $v(\xi) < u(\xi)$ arises when the B-subarc is absent because of insufficient fuel supply, with

$$\omega_{\max} < H_*^{-1}(1), \qquad \xi = 0, \tag{84}$$

so that $v_+(0) = \omega_{\max} < v_0$.

If $v = u$ at some point x_* of $[\xi, x_1]$, then the element $(x, y, u = v)$ at x_* belongs to a fictitious burning rocket. Therefore, the values y', v', u' at x_* satisfy the relations

$$\begin{aligned} y' + v' + \frac{1}{\alpha} &= 0, \\ y' + u' + \frac{1}{\alpha} &= \psi_2(u). \end{aligned} \tag{85}$$

The result follows from (71) with $\psi_2 \equiv 0$ on the C-subarc, and with v, v' replaced by u, u' for the fictitious burning rocket. Although u is not the optimum velocity, (71) and (72) remain formally valid, and the conclusion $\psi_2 > 0$, of Theorems 1 and 2, holds in view of (80). Therefore, $u' - v' > 0$ whenever $u - v = 0$. Since $u - v \geqq 0$ at $x = \xi$, it follows that $u > v$ on (ξ, x_1).

THEOREM 3: "*If* $[v(\xi) - v_1][v(\xi) - V_2] \geqq 0$, *then* $E \geqq 0$ *and* $\mu \leqq 0$ *on* (ξ, x_1)."

Proof: Let a bar over a letter refer to a comparison curve. Observe that for a given (x, y),

$$\begin{aligned} \bar{\omega} &> \omega_{\max}, \quad \bar{v} < v(x) \quad \text{on} \quad (\xi, \bar{\xi}), \\ \bar{\omega} &= \omega_{\max}, \quad \bar{v} = v(x) \quad \text{on} \quad (\bar{\xi}, x_1), \end{aligned} \tag{86}$$

with the aid of (4). From (87) and Lemma 3, $\bar{v} \leqq v(x) < u(x)$, so that, by (83), $g_v < 0$ for all values of v between v and $v(x)$. The Mean Value Theorem now leads to $g(\bar{v}) \geqq g(v)$; (53) and $\lambda > 0$ complete the proof that $E \geqq 0$. To show that $\mu \leqq 0$, note that $\lambda g_v - \mu = 0$.

The physical interpretation of the Theorem is the requirement that the fuel supply be sufficient to preclude a burnout with $v(\xi)$ in the interval (v_1, V_2). In order to relate this requirement to $\omega_{\max}$, note that

$$\begin{aligned} \omega(v) &= v - \log H + \chi(v) - \chi(v_0), \\ \omega(v(\xi)) &= \omega_{\max}, \end{aligned} \tag{87}$$

in view of (4), (24), (31). The hypothesis of Theorem 3 is then equivalent to

$$[\omega_{\max} - \omega(v_1)][\omega_{\max} - \omega(V_2)] \geqq 0. \tag{88}$$

Clearly, it is sufficient but not necessary that the fuel supply be so large that

$$\omega_{\max} \geqq \omega(V_2). \tag{89}$$

The results can be extended to a drag with any $N > 0$, as was done in (47). Of course, if $N = 0$, as in Theorem 2, v_1 and V_2 do not exist. Then the conclusions of Lemma 3 and Theorem 3 are automatic.

14. The sufficiency condition. We resort to the following variant of the *fundamental sufficiency condition* of Weierstrass, proved in the Appendix:

Let a family $y(x, \theta)$ of extremals of a control problem be generated by the initial conditions (63), *involving θ as a parameter. If this family constitutes a field, and if each extremal of the field satisfies I and II* with the appropriate initial conditions, then the extremal for $\theta = 0$ yields an absolute minimum of the control problem.*

Note that Conditions I and II* have been established in sections 5–13 for $\theta = 0$, and that their extension to the family defined by (63) is trivial. Furthermore, the existence of a field has been proved in section 10. We conclude that the hypothesis of the theorem is satisfied, and that our extremal therefore yields an absolute minimum of the *auxiliary problem.*

15. Summary. Under the assumptions of Theorems 1 and 3 or of Theorem 2, a typical solution is characterized by the structural formula

$$(IB)_{N+1}C; \tag{90}$$

i.e., the burning stage B, preceded by an impulsive launching I, contains

N additional impulsive thrusts, N being the number of double tangents of the curve $f(v)$, and is followed by the coasting stage without fuel. The solution therefore includes as a special case the results of Tsien and Evans, where $N = 0$. An absolute minimum has been established with the aid of the second variation and a variant of the *sufficiency principle* that is particularly useful in problems of optimum control.

Theorem 1 applies to terrestrial launching and any drag function with some very general properties listed in section 4; Theorem 2 covers extra-terrestrial launching but is restricted to a fairly common class of drags that includes all the cases previously treated in the literature.

APPENDIX

To prove the Sufficiency Condition stated in section 14, define $w(x, y)$ by

$$
\begin{aligned}
& w \equiv G(x, y) + I^*, \\
& I^* \equiv \int \{[\lambda \cdot g(x, y, u) + \mu\psi]\, dx - \lambda \cdot dy\},
\end{aligned}
\tag{91}
$$

where I^* is the Hilbert integral of the control problem and u, λ, μ belong to the field. With Δ denoting an increment, observe that: 1) $\Delta w = 0$ on a closed path; 2) $\Delta w = 0$ on a boundary subarc in $\Phi = 0$, in virtue of the *transversality condition* (34) and $\mu\psi = 0$; 3) $\Delta w = \Delta G$ on an extremal of the field, in view of $y' = g(x, y, u)$, where $u(x, y)$ is the "slope function".

Next let 0 and 1 denote the end-points of the extremal for $\theta = 0$, and let C_{02} be any admissible arc connecting 0 and a terminal point 2 lying in $\Phi = 0$. Then there follows from the properties of w listed above that

$$
\begin{aligned}
& w_0^1 = G(1) - G(0), \\
& w_1^2 = 0, \\
& w_0^2 = G(2) - G(0) + I^*(C_{02}), \\
& w_0^2 = w_0^1 + w_1^2,
\end{aligned}
\tag{92}
$$

leading to

$$
G(1) - G(2) = I^*(C_{02}). \tag{93}
$$

Finally, note that, in view of (53) and $\mu\psi = 0$, the expression for I^* in (91) can be also exhibited as

$$
I^* = \int \lambda \cdot (g - \bar{g})\, dx = -\int E\, dx, \tag{94}
$$

$$
\bar{g} \equiv g(x, y, \bar{u}); \qquad \bar{u} \neq u,
$$

and that II* implies $I^* < 0$ in (94) and (93). The conclusion

$$G(1) < G(2) \tag{95}$$

follows immediately.

Acknowledgment. The author wishes to express his appreciation to Dr. J. Breakwell and to Dr. W. R. Hazeltine, who read the manuscript critically and made several constructive suggestions.

REFERENCES

[1] G. A. Bliss, *Lectures on the Calculus of Variations*, Univ. of Chicago Press, 1946; pp. 202–204, 223, 224, 227, 238, 241, 257.

[2] J. V. Breakwell, *The optimization of trajectories*, J. Soc. Indust. Appl. Math. 2 (1959) pp. 242–246.

[3] B. Garfinkel, *Minimal problems in airplane performance*, Quart. Appl. Math. 9, No. 2 (1951) pp. 154, 159.

[4] R. H. Goddard, *A method of reaching extreme altitudes*, Smithsonian Inst. Publs. Misc. Collection 71, No. 2 (1919).

[5] G. Hamel, *Über eine mit den problem der rakete zusammenhängende aufgabe der variationsrechnung*, Z. angew. Math. Mech. 7, No. 6 (1927).

[6] G. Leitman, *An elementary derivation of the optimum control conditions*, Proc. 12th Int. Astronaut. Congr., Washington, D.C., 1961.

[7] G. Leitman, *Progress in Astronautical Sciences*, vol. 1, p. 154, North Holland Publishing Co., Amsterdam, 1962.

[8] A. Miele, *Optimization Techniques*, Ch. 4, p. 159, Academic Press 1962.

[9] A. Miele, *Generalized variational approach to the optimum thrust programming for the vertical flight of a rocket*, Z. Flugwiss 6, No. 3 (1958).

[10] H. Oberth, *Wege zur Raumschiffahrt*, (R. Oldenburg, Munich and Berlin, 1929).

[11] S. Ross, *Minimality for problems in vertical and horizontal rocket flight*, ARS Journal 28, No. 1 (1958) pp. 55–56.

[12] H. S. Tsien and R. C. Evans, *Optimum thrust programming for a sounding rocket*, ARS Journal 21, No. 5 (1951), pp. 99–107.

J.SIAM CONTROL
Ser. A, Vol. 1, No. 3
Printed in U.S.A., 1963

A REMARK ON "A NEW PARTIAL DIFFERENTIAL EQUATION FOR THE STABILITY ANALYSIS OF TIME INVARIANT CONTROL SYSTEMS"*

G. P. SZEGÖ‡ AND G. R. GEISS†

1. Introduction. In [1] Szegö presented a generalization of the Zubov method for solving the stability problem associated with autonomous control systems. This note discusses some extensions of [1], examines its last example in detail and presents a correction, due to the second author, to equation (57). Some interesting structural aspects of the stability investigation are further emphasized and illustrated by examples.

2. Some extensions. In [1] Szegö was mainly concerned with the stability properties of the equilibrium point $x = 0$ of the dynamical system [2]

$$\dot{x} = f(x), \qquad f(0) = 0. \tag{1}$$

Two methods were developed for the stability investigation of such a system. The first method is based upon the integration of the partial differential equation (18 of [1])

$$\dot{v}_1 = \langle \operatorname{grad} v_1(x), f(x) \rangle = \frac{\psi(x)}{\beta_1(v_1)} \tag{2}$$

where the function $\psi(x)$ is supposed to be definite along the trajectories of the system (1) and $\beta_1(v_1)$ to be such that the Liapunov function $\alpha_1^*(x)$ converges, i.e.,

$$\alpha_1^*(x) = \int_0^{v_1(x)} \beta_1(s)\, ds < \infty, \qquad \alpha_1^*(0) = 0, \qquad \dot{\alpha}_1^* = \psi(x). \tag{3}$$

The second method is based upon the integration of the partial differential equation (23 in [1]):

$$\dot{v}_2 = \langle \operatorname{grad} v_2(x), f(x) \rangle = \theta(v_2), \tag{4}$$

where $\theta(v_2)$ is such that:

$$\alpha_2^*(x) = \int_0^{v_2(x)} \beta_2(s)\, ds < \infty, \qquad \alpha_2^*(0) = 0, \qquad \dot{\alpha}_2^* = \Omega(v_2(x)), \tag{5}$$

* Received by the editors August 1, 1963.

† Research Department, Grumman Aircraft Engineering Corporation, Bethpage, New York.

‡ Instituto di Meccanica Applicata del Politecnico, Milano. Visitor at the Research Institute of Advanced Studies (RIAS), Baltimore, Maryland.

where

$$(6) \quad \beta_2(v_2) = \frac{\Omega(v_2)}{\theta(v_2)} \quad \text{and} \quad \Omega = \Omega(v_2) \text{ is any semidefinite scalar function.}$$

In the case of equation (4) there may arise what is called a "degenerate case", i.e., the case in which the scalar function $\alpha_2^*(x)$ itself is semidefinite. Since the scalar function $\alpha_2^*(x)$ has continuous first partial derivatives it follows that on the manifold M on which $\alpha_2^*(x)$ vanishes $\dot{\alpha}_2^*(x)$ also vanishes. Thus M is an integral manifold of (1). In the case of the scalar functions $\alpha_1^*(x)$, derived from the integration of the partial differential equation (2), the same situation may arise only if the condition that $\psi(x)$ has to be definite along the trajectories of (1) is relaxed and $\psi(x)$ is allowed to be semidefinite. Under these relaxed conditions equation (2) contains as a particular case equation (4) and both the partial differential equations suggested by Zubov ([2] Theorem 19 and [3] Theorem 52). In this case, if the resulting Liapunov function vanishes on M, the *stability properties* of M will be defined by the following:

THEOREM. *Consider the dynamical system* (1). *Let*

i) $v(x)$ *be a continuous scalar function with continuous first partial derivatives in the whole space* E^n.

ii) $\theta(v)$ *be a continuous scalar function.*

iii) M *be the manifold on which* $v(x) = 0$.

Assume that:

iv) $\theta(v(x)) \equiv 0$ *in all points of* M, $\theta(v(x)) \not\equiv 0$ *for* $x \notin M$.

v) *The partial differential equation*

$$(7) \qquad \langle \operatorname{grad} v(x), f(x) \rangle = -\theta(v)$$

is satisfied in the whole space E^n.

vi)

$$(8) \qquad v(x)\theta(v(x)) \geqq 0$$

in the whole space E^n.

vii) *The trivial solution* $v = 0$ *of the equation*

$$(9) \qquad \dot{v} = -\theta(v)$$

is globally asymptotically stable.

viii)

$$(10) \qquad \begin{aligned} &\text{I)} \qquad a(\rho(x, M)) \leqq |v(x)|, \\ &\text{II)} \qquad a(\rho(x, M)) \leqq |v(x)| \leqq b(\rho(x, M)) \end{aligned}$$

where $\rho(x, M)$ *is the euclidean distance of the point* x *from the set* M, $a(r)$

and $b(r)$ *are positive definite scalar functions, and* $a(r)$ *is such that* $\lim_{r\to\infty} a(r) = \infty$; *then if* viii I) *is satisfied* $\lim_{t\to\infty} \rho(x(t), M) = 0$ *for all initial conditions, and if* viii II) *is satisfied* M *is globally asymptotically stable.*

Proof: The assumptions i)–vi) imply that the manifold M is stable in the v-norm, vii) implies that M is globally asymptotically stable in the v-norm, viii) translates the v-norm properties into euclidean norm properties.

This theorem can be easily extended to include Theorem 2 of [5] and all instability cases.

In the particular case in which M is a minimal set containing the equilibrium point $x = 0$, asymptotic stability of M implies asymptotic stability of $x = 0$.

It is interesting to examine what form M can have and what consequences its form will have on the outcome of the stability analysis. We shall discuss the case $n = 2$ first. If in this case M is a closed, bounded curve and does not contain the point $x = 0$ it corresponds to a periodic motion. If it contains equilibrium points it may be a "path polygon" [6] or all its points may be equilibrium points. If M is unbounded, then either all its points are equilibrium points or M corresponds to a singular solution of (1).

The same conclusions as in the case $n = 2$ may be reached, for the case $n = 3$, $M = \mathfrak{M}^1$. If $M = \mathfrak{M}^2$ and if M is compact, then by a theorem due to Schwartz [7] we know that the only minimal sets on it will be equilibrium points, closed orbits or the whole M which in this case must be a torus $T = \mathfrak{M}^2$. If M is not compact no immediate conclusions on its structure can be reached.

The stability problem of (1) will in any case be reduced upon the identification of M to a problem of dimension at most $n - 1$, the dimension of M.

In the following example (example 3 of [1]) we shall illustrate the case $n = 2$, M noncompact.

3. Example 1. The system with which we are concerned is

$$\begin{aligned} &\text{a)}\ \dot{x} = y^3 - x \\ &\text{b)}\ \dot{y} = x - \tfrac{1}{2}y. \end{aligned} \tag{11}$$

It is shown in [1] that the function

$$v = 2x^2 - y^4 \tag{12}$$

satisfies the relation

$$\dot{v} = -2v \tag{13}$$

and thus the solution

$$2x^2 - y^4 = 0 \tag{14}$$

is globally asymptotically stable. This result is interesting for a number of reasons.

The solution (14) is a singular solution, i.e., write (11) as

$$\frac{dx}{dy} = \frac{2y^3 - 2x}{2x - y} \tag{15}$$

and then let

$$u = x/y^2 \tag{16}$$

to obtain

$$\frac{du}{dy} = \frac{2 - 4u^2}{2uy - 1}. \tag{17}$$

Now, observe that $\frac{du}{dy} = 0$ when $u^2 = \frac{1}{2}$, i.e., when

$$2x^2 - y^4 = 0. \tag{14}$$

The equilibrium points of system (11) are contained in the solution curve (14). In particular, the system (11) has equilibrium points at

$$\begin{aligned} x &= 0, \pm\frac{\sqrt{2}}{4} \\ y &= 0, \pm\frac{\sqrt{2}}{2} \end{aligned} \tag{18}$$

and they lie on the branch of (14) given by

$$\sqrt{2}x = y\,|\,y\,|. \tag{19}$$

The linear approximations to (11), in the neighborhood of the equilibrium points, are:

$$\left.\begin{aligned} \dot{x} &= -x \\ \dot{y} &= -\tfrac{1}{2}y + x \end{aligned}\right\} \text{near} \quad \begin{aligned} x &= 0 \\ y &= 0 \end{aligned} \tag{20}$$

and

$$\left.\begin{aligned} \dot{x} &= -\left(x \mp \frac{\sqrt{2}}{4}\right) + \frac{3}{2}\left(y \mp \frac{\sqrt{2}}{2}\right) \\ \dot{y} &= \left(x \mp \frac{\sqrt{2}}{4}\right) - \frac{1}{2}\left(y \mp \frac{\sqrt{2}}{2}\right) \end{aligned}\right\} \text{near} \quad \begin{aligned} x &= \pm\frac{\sqrt{2}}{4} \\ y &= \pm\frac{\sqrt{2}}{2} \end{aligned}. \tag{21}$$

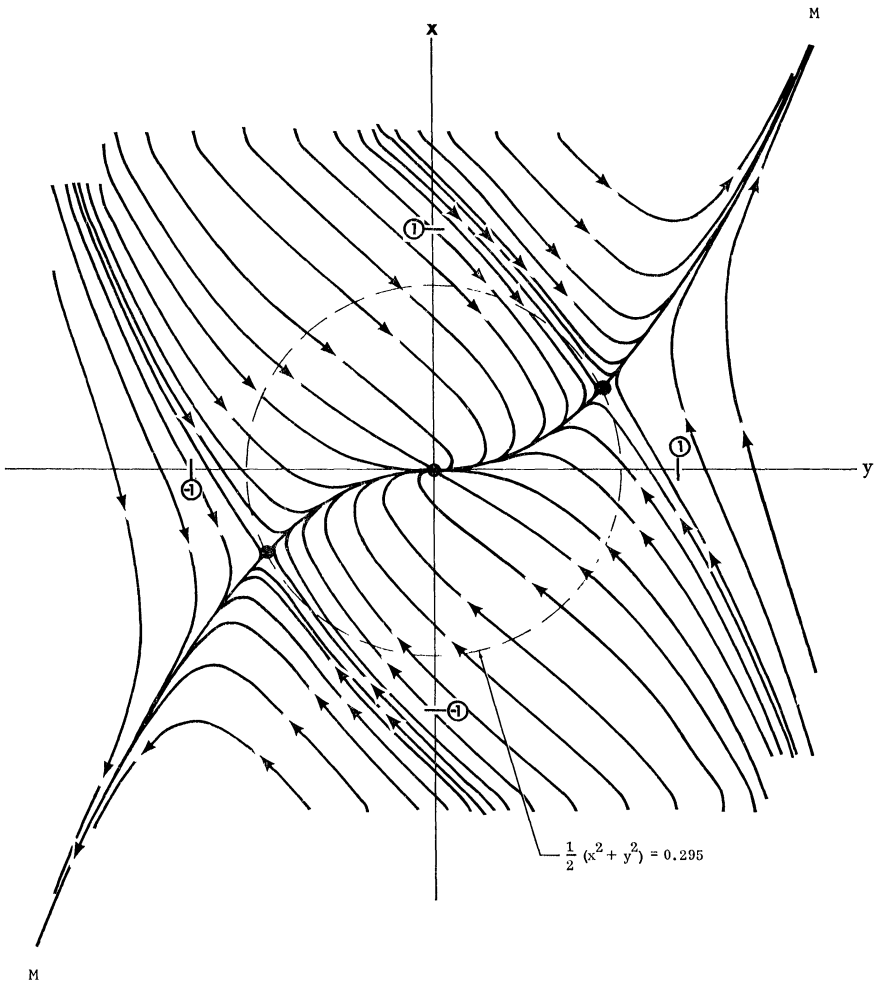

FIG. 1. Phase Portrait of System (11)

Thus, the origin is a stable node (the eigenvalues are $-\frac{1}{2}$, -1) and the other equilibrium points are saddle points (the eigenvalues are $+\frac{1}{2}$, -2).

Now, consider the Liapunov function,

$$V = \tfrac{1}{2}(x^2 + y^2) \tag{22}$$

whose derivative along trajectories of (11) is

$$\dot{V} = -\tfrac{1}{2}x^2 - \tfrac{1}{2}(x - y)^2 + xy^3. \tag{23}$$

It can easily be shown that $V > 0$ and $\dot{V} < 0$ in the region Ω_l,

$$\Omega_l : V = \tfrac{1}{2}(x^2 + y^2) < 0.295; \tag{24}$$

therefore, all solutions beginning in Ω_l tend toward the origin as $t \to \infty$. (See LaSalle and Lefschetz [8].)

Finally, one obtains the phase portrait shown in Fig. 1. One may express the solution curves in terms of elliptic functions of the first kind, i.e., by inverting (17) obtain

$$\frac{dy}{du} + \frac{2u}{4u^2 - 2}\, y = \frac{1}{4u^2 - 2} \tag{25}$$

and then

$$y = (4u^2 - 2)^{-1/4} \int (4u^2 - 2)^{-3/4}\, du, \tag{26}$$

from which the elliptic function expression is obtained [9]. The time, T, taken to move from one point to another on a given nonsingular trajectory is obtained by integrating (13), viz.,

$$T = \frac{1}{2} \ln \left[\frac{2x^2(0) - y^4(0)}{2x^2(T) - y^4(T)} \right]. \tag{27}$$

The parametric form of the singular solution (14) is obtained as follows: Solve (14) for x, i.e.,

$$x = \pm \frac{1}{\sqrt{2}}\, y^2 \tag{28}$$

and substitute this in (11b) to obtain

$$\dot{y} = \pm \frac{1}{\sqrt{2}}\, y^2 - \frac{1}{2}\, y \tag{29}$$

which may be integrated to yield

$$y = (\pm\sqrt{2} + ce^{1/2\, t})^{-1}.$$

Finally, substitution of the proper value for c and use of (28) result in

$$\begin{aligned} &\text{a)} \quad y(t) = \left([\operatorname{sgn} x(0)]\sqrt{2}\,(1 - \epsilon^{1/2\, t}) + \frac{1}{y(0)}\, \epsilon^{1/2t} \right)^{-1} \\ &\text{b)} \quad x(t) = \frac{\operatorname{sgn} x(0)}{\sqrt{2}} \left([\operatorname{sgn} x(0)]\sqrt{2}(1 - \epsilon^{1/2\, t}) + (2x^2(0))^{-1/4} \epsilon^{1/2\, t} \right)^{-2}. \end{aligned} \tag{30}$$

The expression (30b) does not agree with (57) of [1]; however, (57) is obviously in error since it does not allow the equilibrium solutions (18).

It is also interesting to note that if (11b) is rewritten as

$$x = \dot{y} + \tfrac{1}{2}y$$

and then substituted into (11a) one obtains the unforced Duffing equation, viz.,

$$\ddot{y} + \tfrac{3}{2}\dot{y} + \tfrac{1}{2}y - y^3 = 0. \tag{31}$$

Thus, in correspondence with the previous development, (31) has the singular phase solution (29). In fact, one can show that systems of the form

$$\begin{aligned} \dot{x} &= y^3 - \alpha x \\ \dot{y} &= x - \beta y, \end{aligned} \tag{32}$$

or equivalently,

$$\ddot{y} + (\alpha + \beta)\dot{y} + \alpha\beta y - y^3 = 0 \tag{33}$$

have the respective singular solutions:

$$x = \pm \frac{1}{\sqrt{2}} y^2 \tag{34}$$

and

$$\dot{y} = -\beta y \pm \frac{1}{\sqrt{2}} y^2 \tag{35}$$

if $\alpha/\beta = 2$.

In the following example we shall illustrate the case $n = 3$, M non-compact and containing a periodic orbit.

4. Example 2. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= -x_1 - x_2 - x_3 + \epsilon(1 - x_1^2 - 2x_1x_2)x_2 + \epsilon(1 - x_1^2)x_3\,, \quad \epsilon > 0 \end{aligned} \tag{36}$$

and the scalar function

$$v = -x_1 + \epsilon x_2 - \epsilon x_1^2 x_2 - x_3, \tag{37}$$

whose total time derivative with respect to the system (36) is

$$\dot{v} = x_1 - \epsilon x_2 + \epsilon x_1^2 x_2 + x_3 = -v. \tag{38}$$

We conclude that the manifold M on which the scalar function (37) vanishes

$$M: \qquad x_3 = -x_1 + \epsilon x_2 - \epsilon x_1^2 x_2 \tag{39}$$

is asymptotically stable.

By substituting (39) into (36) the third equation becomes an identity and we obtain the familiar Van der Pol equation

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + \epsilon x_2 - \epsilon x_1^2 x_2 \,. \end{aligned} \tag{40}$$

We conclude that the equilibrium point $x = 0$ of equation (36) (which is its only equilibrium point) is unstable and that the equation (36) has one asymptotically stable orbit which lies on the surface (39) and is defined by the equation (40).

5. Acknowledgments. The work reported in this note was partly supported by the Aeronautical Systems Division, Air Force Systems Command and the Advanced Research Projects Agency under Contract AF 33(657)-9185, Project No. 8219, Task No. 04; and by the U. S. Air Force under Contract No. AF 33(657)-8559.

REFERENCES

[1] G. P. Szegö, *On a new partial differential equation for the stability analysis of time invariant control systems*, J. Soc. Indust. Appl. Math. Ser. A. 1, (1) (1962), pp. 63–75.

[2] V. I. Zubov, *Methods of A. M. Liapunov and their applications*, Izdatel'stvo Leningradskogo Universiteta, Leningrad (1957), (English Trans. AEC-tr-4439).

[3] V. I. Zubov, *Mathematical methods of investigating automatic regulation systems*, Gosudarstvennoe Soyuznoe Izdatel'stvo Sudoistroitel'noi Promyshlennosti, Leningrad (1959), (English Trans. AEC-tr-4494).

[4] J. P. LaSalle, *Some extensions of Liapunov's second method*, IRE Trans. PGCT, Vol. CT-7, 1960, pp. 520–527.

[5] G. P. Szegö, *A contribution to Liapunov's second method: nonlinear autonomous systems*, Proc. of the International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York (1963), pp. 421–430.

[6] V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton Univ. Press, Princeton (1960).

[7] A. Schwartz, *The Poincaré-Bendixson theorems on two-manifolds*, (to be published in the Proc. Am. Math. Soc.).

[8] J. P. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press (1962).

[9] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals*, Springer (1954), p. 149, 271.51.