SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

ASYMPTOTIC STABILITY IN THE LARGE OF A CLASS OF SINGLE-LOOP FEEDBACK SYSTEMS*

R. A. BAKER† AND C. A. DESOER‡

The purpose of this paper is to obtain some sufficient conditions for asymptotic stability in the large of a large class of systems. The basic idea is due to O'Shea [1]. We consider a system whose block diagram representation is shown in Fig. 1. Our results extend those of O'Shea in several directions:

(a) The linear time-invariant subsystem, denoted by G in Fig. 1, is allowed to belong to a much broader class. By describing G by a convolution operator we allow in the class not only systems described by differential equations but also systems discussed by difference differential equations [2, p. 189], [3]. Also allowed are systems whose internal dynamics require partial differential equations, say, because of diffusion process or wave propagation.

(b) The conditions on the nonlinearity ϕ are less restrictive.

(c) The results are stated more sharply in terms of the disturbance η.

The input-output relation of the linear time-invariant subsystem is

$$\sigma_e(t) = \int_0^t g(t - \tau)e(\tau)\, d\tau, \qquad t \geqq 0. \tag{1}$$

In the following, all the time functions are defined for $t \geqq 0$, and we shall use the symbol $*$ to denote the convolution of such functions. Thus (1) is written as

$$\sigma_e = g * e.$$

The input-output relation of the nonlinearity is

$$c(t) = \phi[\sigma(t)]. \tag{2}$$

The specific assumptions which apply throughout are the following:

(N1) $\phi: R \to R, \phi(0) = 0$, where R denotes the set of all real numbers.

* Received by the editors January 23, 1967, and in revised form September 20, 1967. This research was supported in part by the National Aeronautics and Space Administration under Grant NsG-354, Supplement 3, Ford Foundation support through the Forgivable Predoctoral Loan to Future Engineering Teachers and the Adolph C. and Mary Sprague Miller Institute for Basic Research in Science.

† Department of Electrical Engineering and Electronics Research Laboratory, University of California, Berkeley, California 94720. On leave from Washington State University, Pullman, Washington.

‡ Department of Electrical Engineering and Electronics Research Laboratory, University of California, Berkeley, California 94720.

(N2) For some finite k, $|\phi(\sigma)| < |k\sigma|$ for all $\sigma \neq 0$.

(N3) $0 \leqq \dfrac{\phi(\sigma_1) - \phi(\sigma_2)}{\sigma_1 - \sigma_2} \leqq k$ for all σ_1, σ_2 except $\sigma_1 \neq \sigma_2$.

(N4) If $\phi(\sigma) = 0$ for some $\sigma \neq 0$, then there is a $\sigma_1 > 0$ such that $\phi(\sigma) = 0$ for all $\sigma \in [-\sigma_1, \sigma_1]$.

(G1) $$g \in L_2(0, \infty).$$

The distributional derivative $\dot{g}$ of g is of the form

(G2) $$\dot{g}(t) = \dot{g}_1(t) + \sum_{i=1}^{\infty} a_i \delta(t - t_i),$$

where

(G3) $$\dot{g}_1 \in L_1(0, \infty), \quad \sum |a_i| < \infty.$$

(E1) $$\eta \in L_1(0, \infty),$$

(E2) η is differentiable and $\dot{\eta} \in L_1(0, \infty)$.

Observe that (G2) and (G3) imply that g is bounded on $[0, \infty)$ and that $g(t) \to 0$ as $t \to \infty$. The same holds for η. Call

$$\eta_M = \sup_{t \geqq 0} |\eta(t)|, \quad g_M = \sup_{t \geqq 0} |g(t)|.$$

Throughout the paper we add the subscript M to the name of a function to denote the sup of the absolute value of that function; the subscript M suggests the idea of "maximum." In some manipulations to follow, it is useful to consider the functions g, e, σ, η and c to be defined for all t, all of them being identical to zero for $t < 0$. We use $\hat{}$ to denote Fourier transforms; e.g.,

$$\hat{g}(i\omega) = \int_0^{\infty} g(t) e^{-i\omega t}\, dt.$$

We use $\| \cdot \|$ to denote L_1 norms; e.g.,

$$\|\eta\| = \int_0^{\infty} |\eta(t)|\, dt.$$

The system shown in Fig. 1 obeys the equation

(3) $$\sigma(t) = \eta(t) - \int_0^t g(t - t')\phi[\sigma(t')]\, dt', \qquad t \geqq 0,$$

or, in abbreviated notation,

$$\sigma(t) = \eta(t) - (g * c)(t),$$

where we use (2).

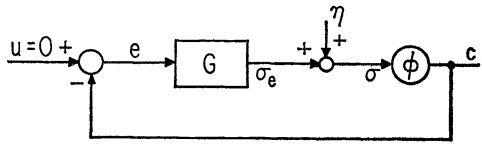


FIG. 1. *Feedback system under consideration*

We come now to the main result of the paper.

THEOREM. *Consider the system shown in Fig.* 1. *Suppose that assumptions* (N1) *to* (N4), (G1) *to* (G3), (E1) *and* (E2) *hold. Let* y *be any real-valued function which has a Fourier transform* $\hat{y}$ *such that* $y(t) = 0$ *for* $t < 0$, $y(t) \leqq 0$ *for* $t \geqq 0$ *and* $\| y \| < 1$. *Under these conditions, if for some* $\alpha > 0$,

$$(4)\quad \operatorname{Re}\{[1 + i\omega\alpha + \hat{y}(i\omega)][\hat{g}(i\omega) + 1/k]\} \geqq 0 \quad \textit{for all} \quad \omega \in (-\infty, \infty),$$

then

(i) $\sup_{t \geqq 0} | \sigma(t) | < \infty$,

(ii) $\sigma(t) \to 0$ *as* $t \to \infty$,

(iii) *as* $\| \eta \| + \| \dot{\eta} \| \to 0$, *the corresponding* σ *has the property that* $\sup_{t \geqq 0} | \sigma(t) | \to 0$.

Note 1. By (N3), the output c has the same properties.

Note 2. If ϕ is identically zero, the conclusions are immediate consequences of (E1) and (E2). From now on, ϕ is assumed not identically zero.

The proof of this theorem is somewhat involved. In order to simplify it we quote a lemma (see [5]–[7]).

LEMMA. *Let* x *and* y *be in* $L_2(-\infty, \infty)$. *Let, for each* $t \in R$, $(x(t), y(t)) \in \varphi$, *where* φ *is a monotonically increasing relation (i.e.,* $\xi_1, \xi_2 \in R$ *implies* $[\varphi(\xi_1) - \varphi(\xi_2)](\xi_1 - \xi_2) \geqq 0$); *then for all* $\tau \in R$,

$$\int_{-\infty}^{\infty} x(t)y(t)\, dt \geqq \int_{-\infty}^{\infty} x(t)y(t - \tau)\, dt.$$

If, in addition, φ *is odd (i.e.,* $(\xi, \eta) \in \varphi$ *implies* $(-\xi, -\eta) \in \varphi$), *then the inequality above holds with absolute value signs on both integrands.*

Proof of the theorem. I. The system shown in Fig. 1 is characterized by (3). The given function η is continuous and bounded, g is bounded and, by (N3), ϕ satisfies a Lipschitz condition; then solving (3) by iteration we can apply the standard arguments to show that the resulting sequence converges uniformly on every bounded interval, and that (3) has a unique solution which is continuous. For brevity, let $L_{\infty e}$ be the class of all measurable functions which are bounded on every bounded interval. Thus $\sigma \in L_{\infty e}$; clearly, $c \in L_{\infty e}$ and $e \in L_{\infty e}$.

II. Let T be an arbitrary positive number. Let $\sigma_m = \sigma + \sigma * y$,

$c_m = c + c * y$, and, in general, given any function x, we define $x_m = x + x * y$. Then

$$\int_0^T \left(\sigma_m(t) - \frac{c_m(t)}{k}\right)c(t)\,dt = \int_0^T \left(\sigma(t) - \frac{c(t)}{k}\right)c(t)\,dt \int_0^T + \left(y * \left(\sigma - \frac{c}{k}\right)\right)(t)c(t)\,dt, \tag{5}$$

where all integrals are finite since $y \in L_1$, $\sigma \in L_{\infty e}$, and $c \in L_{\infty e}$. Now let the subscript T denote the truncation of a function to the interval $[0, T]$: thus, $f_T(t) = f(t)$ on $[0, T]$, and $f_T(t) = 0$ elsewhere. Considering the second integral in (5), we define

$$R(\tau) = \int_0^\infty \left[\sigma_T(t - \tau) - \frac{c_T(t - \tau)}{k}\right]c_T(t)\,dt, \tag{5a}$$

and observe that by Fubini's theorem,

$$\int_0^T \left(y * \left(\sigma - \frac{c}{k}\right)\right)(t)c(t)\,dt = \int_0^\infty y(\tau)R(\tau)\,d\tau. \tag{6}$$

Observe that, for each t, the real numbers $c_T(t)$ and $\sigma_T(t) - c_T(t)/k$ are monotonically related: indeed, denoting $c_T(t_i)$ by c_i and $\sigma_T(t_i)$ by σ_i, we have

$$\left[(\sigma_1 - \sigma_2) - \frac{(c_1 - c_2)}{k}\right](c_1 - c_2) = (\sigma_1 - \sigma_2)(c_1 - c_2)\left[k - \frac{(c_1 - c_2)}{(\sigma_1 - \sigma_2)}\right]k^{-1} \geqq 0,$$

where the inequality follows from (N3). Consequently, by Lemma 1, $R(\tau) \leqq R(0)$; and since $y \leqq 0$, (6) gives

$$\int_0^\infty y(\tau)R(\tau)\,d\tau \geqq -R(0)\|y\|. \tag{6a}$$

Thus, the left-hand side integral of (5) is larger than or equal to $(1 - \|y\|)\cdot R(0) \geqq 0$. In other words, for each $T > 0$, there is a finite $b(T) \geqq (1 - \|y\|) > 0$ such that

$$\int_0^T \left[\sigma_m(t) - \frac{c_m(t)}{k}\right]c(t)\,dt = b(T)R(0) \geqq 0. \tag{7}$$

III. From the block diagram, $\sigma = \sigma_e + \eta$; hence,

$$\int_0^T \left[\sigma_m(t) + \alpha\dot{\sigma}(t) - \frac{c_m(t)}{k}\right]c(t)\,dt = \int_0^T \left[\sigma_{em}(t) + \alpha\dot{\sigma}_e(t) - \frac{c_m(t)}{k}\right]c(t)\,dt + \int_0^T [\eta_m(t) + \alpha\dot{\eta}(t)]c(t)\,dt. \tag{8}$$

In order to show by Fourier methods that the first integral of the right-hand side of (8) is nonpositive, observe that if $\sigma_e' = -g * c_T$ and $\dot{\sigma}_e' = -\dot{g} * c_T$, then, on $[0, T]$, $\sigma_e = \sigma_e'$, $\sigma_{em} = \sigma_{em}'$, and $\dot{\sigma}_e = \dot{\sigma}_e'$. Similarly, c may be replaced by c_T. Now $c \in L_{\infty e}$, hence $c_T \in L_1 \cap L_2$. With $g \in L_2$, this implies $\sigma_e' \in L_2$; hence, since $\dot{g}$ and $y \in L_1$, $\dot{\sigma}_e \in L_2$ and $\sigma_{em}' \in L_2$ (see [4]). Therefore the first integral in (8) is the product of two L_2-functions. Using Parseval's theorem, and noting that odd functions of ω contribute nothing to the integral, we obtain

$$\int_0^T \left[\sigma_{em}'(t) + \alpha\dot{\sigma}_e'(t) - \frac{c_m'(t)}{k}\right] c(t)\, dt$$

$$= \frac{-1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{Re}\left\{[1 + \alpha i\omega + \hat{y}(i\omega)]\left[\hat{g}(i\omega) + \frac{1}{k}\right]\right\} \hat{c}_T(i\omega)\hat{c}_T^*(i\omega)\, d\omega \leqq 0,$$

where the inequality follows by (4). Thus (8) implies that, for all $T > 0$,

$$(9) \quad \int_0^T \left[\sigma_m(t) + \alpha\dot{\sigma}(t) - \frac{c_m(t)}{k}\right] c(t)\, dt \leqq \int_0^T [\eta_m(t) + \alpha\dot{\eta}(t)]c(t)\, dt.$$

This is the fundamental inequality.

IV. Using (7) in (9), we conclude that, for all $T > 0$,

$$(10) \qquad \alpha \int_0^T \dot{\sigma}(t)c(t)\, dt \leqq \int_0^T [\eta_m(t) + \alpha\dot{\eta}(t)]c(t)\, dt.$$

Let $c_{TM} = \sup_t |\, c_T(t)\,|$. Since $\| \eta_m \| \leqq \| \eta \| + \| y \| \| \eta \|$, we conclude that

$$(11) \quad \alpha \int_0^T \dot{\sigma}(t)c(t)\, dt \leqq c_{TM}[(1 + \| y \|)\| \eta \| + \alpha \| \dot{\eta} \|] = c_{TM} M,$$

where M denotes the bracket of the right-hand side of the inequality. Call $\Phi(x) = \int_0^x \phi(\sigma)\, d\sigma$; then, with $\alpha > 0$, (11) implies that for all $T > 0$,

$$(12) \qquad \Phi[\sigma(T)] \leqq \Phi[\sigma(0)] + c_{TM} M\alpha^{-1}.$$

The slope condition (N3) on ϕ implies that $\Phi(x) \geqq [\phi(x)]^2/2k$; hence for all $T > 0$,

$$\frac{1}{2k}\,|\, c(T)\,|^2 \leqq \Phi[\sigma(0)] + c_{TM} M\alpha^{-1}.$$

This inequality implies that $c \in L_\infty$; recalling that c is continuous this is easily shown by contradiction. In fact,

$$(15) \qquad \sup_{t \geqq 0} |\, c(t)\,| \leqq \{(kM\alpha^{-1})^2 + 2k\Phi[\sigma(0)]\} + kM\alpha^{-1}.$$

Since $\sigma_e(0) = 0$ by (1), as $\| \eta \| + \| \dot{\eta} \| \to 0$ both M and $\Phi[\sigma(0)]$ tend to zero and so does $\sup_{t \geqq 0} |\, c(t)\,|$.

V. Let us show that

$$\int_0^T \left(\sigma(t) - \frac{c(t)}{k}\right) c(t)\, dt$$

is bounded. By (5a), this integral is $R(0)$. If we let $c_M = \sup_{t \geqq 0} | c(t) |$, then (7), (9) and (11) give

$$b(T) \int_0^T \left(\sigma(t) - \frac{c(t)}{k}\right) c(t)\, dt \leqq c_M M + \alpha\Phi[\sigma(0)].$$

Observing that ϕ is monotonic and that $\sigma(0) = \eta(0)$, we obtain

$$\Phi[\sigma(0)] \leqq \eta(0)c(0) \leqq \| \dot{\eta} \| c_M \leqq M c_M \alpha^{-1}.$$

Hence,

$$\int_0^T \left(\sigma(t) - \frac{c(t)}{k}\right) c(t)\, dt \leqq 2c_M M(1 - \| y \|)^{-1}.$$

Since the integrand is nonnegative by (N3) and since the right-hand side is independent of T, we have

$$\int_0^\infty \left(\sigma(t) - \frac{c(t)}{k}\right) c(t)\, dt \leqq 2c_M M(1 - \| y \|)^{-1}. \tag{16}$$

Note that as $\| \eta \| + \| \dot{\eta} \| \to 0$, the bound on the right-hand side of (16) tends to zero, because both $c_M \to 0$ and $M \to 0$.

VI. To complete the proof we must consider the two possible behaviors of the nonlinear characteristic in the neighborhood of the origin.

Case 1. $\phi(\sigma) = 0$ implies $\sigma = 0$. We know that $\sigma = \sigma_e + \eta$ and that $\eta \to 0$ as $t \to \infty$. Since $\dot{\sigma}_e = -\dot{g} * c$, where $\dot{g} \in L_1$ and $c \in L_\infty$, it follows that $\dot{\sigma}_e \in L_\infty$; hence σ_e is *uniformly* continuous on $[0, \infty)$. If σ_e did not tend to 0 as $t \to \infty$, then σ does not go to zero; using the uniform continuity of σ_e we can easily show that the area under the function

$$\left(\sigma(t) - \frac{\phi[\sigma(t)]}{k}\right) \phi[\sigma(t)] \tag{17}$$

would then be infinite [8].[1] This contradicts (16). Hence $\sigma \to 0$ as $t \to \infty$. This, together with $\sigma \in L_{\infty e}$, implies that $\sigma \in L_\infty$. Thus (i) and (ii) are established, and (iii) follows by contradiction: if $\| \eta \| + \| \dot{\eta} \| \to 0$ and $\sup_{t \geqq 0} | \sigma(t) |$ does not go to zero, then, because of the uniform continuity of σ_e, the bound on the integral in (16) could not go to zero.

Case 2. $\phi(\sigma) = 0$ for all $\sigma \in [-\sigma_1, \sigma_2]$ with $\sigma_1 > 0$, $\sigma_2 > 0$. Using the

[1] Note that in the proof of [8, Lemma 1, p. 58] only the *uniform* continuity of f is needed.

monotonicity of ϕ and inequality (16), we obtain

$$2c_M M(1 - \| y \|)^{-1} \geqq \int_0^\infty \left(\sigma(t) - \frac{c(t)}{k}\right) c(t)\, dt \geqq \sigma_1 \int_0^\infty c^-(t)\, dt + \sigma_2 \int_0^\infty c^+(t)\, dt. \tag{18}$$

Here the superscripts $^+$ and $^-$ are used in their usual sense:

$$c^+(t) = \sup\{c(t), 0\}, \qquad c^-(t) = \sup\{-c(t), 0\}.$$

Hence $c \in L_1$. But already we know $c \in L_\infty$, hence $c \in L_2$. Now $\sigma_e = -g * c$, hence $\hat{\sigma}_e = \hat{g}\hat{c}$, where $\hat{g}, \hat{c} \in L_2$. Consequently, $\hat{\sigma}_e \in L_1$. Now by the Riemann-Lebesgue lemma, $\sigma_e(t) \to 0$ as $t \to \infty$, hence $\sigma(t) \to 0$ as $t \to \infty$. Now, as $\| \eta \| + \| \dot{\eta} \| \to 0$, (18) implies that $\| c \| \to 0$. But

$$| \sigma_e(t) | \leqq g_M \| c \|,$$

and consequently, $\sup_{t \geqq 0} | \sigma_e(t) | \to 0$ and so does $\sup_{t \geqq 0} | \sigma(t) |$. Therefore (i), (ii) and (iii) have been established.

COROLLARY. *If ϕ, the characteristic of the nonlinearity, is an odd function, then the theorem still holds without the requirement that $y(t) \leqq 0$ for $t \geqq 0$.*

Proof. By the lemma, since ϕ is odd, $R(0) \geqq | R(\tau) |$ for all τ. The previous assumption that $y(t) \leqq 0$ was used only in the derivation of (7) from (6). Under the present conditions,

$$\left| \int_0^\infty y(\tau) R(\tau)\, d\tau \right| \leqq \int_0^\infty | y(\tau) |\, | R(\tau) |\, d\tau \leqq R(0) \| y \|;$$

hence, the previous equation (6a) still holds. The remainder of the proof requires no modifications.

REFERENCES

[1] R. P. O'SHEA, *A combined frequency time-domain stability criterion for autonomous continuous systems*, Trans. IEEE Automatic Control, AC-11 (1966), pp. 477–484.

[2] R. BELLMAN AND K. L. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.

[3] A. HALANAY, *Differential Equations: Stability, Oscillations, Time Lags*, Academic Press, New York, 1966.

[4] S. BOCHNER AND K. CHANDRASEKHARAN, *Fourier Transforms*, Princeton University Press, Princeton, 1949.

[5] R. A. BAKER AND C. A. DESOER, *On scalar products of signals passing through memoryless nonlinearities with delay*, Memo. ERL-M198, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, 1966, to be published.

[6] P. L. FALB AND G. ZAMES, *On cross-correlation bounds and the positivity of certain*

nonlinear operators, IEEE Trans. Automatic Control, AC-12 (1967), pp. 219–221.

[7] J. C. WILLEMS AND M. GRUBER, *Comments on 'A combined frequency-time domain stability criterion for autonomous continuous systems,'* Ibid., AC-12 (1967), pp. 217–219.

[8] M. A. AIZERMAN AND F. R. GANTMACHER, *Absolute Stability of Regulator Systems*, Holden-Day, San Francisco, 1964.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

NECESSARY CONDITIONS FOR CONTROL PROBLEMS WITH VARIABLE TIME LAGS*

H. T. BANKS†

Introduction. In [5] Kharatishvili extended Pontryagin's proof of the maximum principle to obtain a pointwise maximum principle for a class of nonlinear control problems involving system equations with constant lags in the state variables. The class of admissible controls consisted of piecewise continuous control functions. Friedman [3] also considered a hereditary control problem to which he applied Pontryagin's methods to obtain necessary conditions. In a recent book [4] Oguztoreli showed that the results of several recent papers on control problems with systems of ordinary differential equations could be extended to obtain corresponding results for systems involving delays.

Gamkrelidze [1] (see also Neustadt [2]) gave a very elegant proof of a maximum principle in integral form which includes Pontryagin's pointwise maximum principle as a special case. In this paper we consider a general nonlinear system with variable time dependent lags. A maximum principle in integral form is obtained for a class of Lebesgue measurable control functions. Transversality conditions are also given for variable initial functions. The proofs are generalizations of Gamkrelidze's methods involving quasi-convex families of functions.

Throughout this paper, by a solution to a differential equation we shall mean an absolutely continuous (A. C.) function which satisfies the equation almost everywhere with respect to Lebesgue measure. Unless it is specifically stated otherwise, by a measurable function we shall mean a Lebesgue measurable function. Vector matrix notation will be used. Following common usage we shall not distinguish between a vector and its transpose when it is clear which we mean. The notation $|A|$ will denote the Euclidean norm of A in whatever space A lies. For example, if A is a matrix (a^{ij}), then $|A|^2 = \sum_i \sum_j (a^{ij})^2$. If g is a vector function depending on the vector x, g_x will denote the Jacobian matrix with elements $\partial g^i/\partial x^j$. By a nontrivial vector function we shall mean one which is not identically zero. Finally, the scalar t will denote time and $\dot{x}$ will denote the derivative of x with respect to t.

1. Preliminary results. Let $F = F(x(\cdot), t)$ be an n-vector functional defined on $C[\alpha_0, t_1] \times [t_0, t_1]$, where $\alpha_0 < t_0$. By the notation $F(x(\cdot), t)$ we

* Received by the editors June 30, 1967, and in revised form October 27, 1967.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported by a National Science Foundation Graduate Fellowship at the Division of Mathematical Sciences, Purdue University, Indiana.

mean that at each t, F depends on t and on the values $x(s)$, $\alpha_0 \leqq s \leqq t$, where $x \in C[\alpha_0, t_1]$, x an n-vector.

Suppose that F is measurable in t and continuous in x on $C[\alpha_0, t_1]$ for almost all fixed t in $[t_0, t_1]$. Assume there exists an $m \in L_1(t_0, t_1)$ such that $|F(x(\cdot), t)| \leqq m(t)\|x\|_t$ for $(x, t) \in C[\alpha_0, t_1] \times [t_0, t_1]$, where $\|x\|_t = \sup\{|x(s)| : s \in [\alpha_0, t]\}$.

THEOREM 1. *Let F satisfy the above assumptions and let $\phi \in C[\alpha_0, t_0]$. Then*

$$\begin{aligned} \dot{x}(t) &= F(x(\cdot), t) \quad \text{on} \quad [t_0, t_1], \\ x(t) &= \phi(t) \qquad\quad \text{on} \quad [\alpha_0, t_0] \end{aligned} \tag{1.1}$$

has a solution on $[t_0, t_1]$.

THEOREM 2. *Let $D(t)$ be an n-vector function, $D \in L_1(t_0, t_1)$. In addition to the assumptions of Theorem 1, suppose there exists $p \in L_1(t_0, t_1)$ such that F satisfies*

$$|F(x(\cdot), t) - F(y(\cdot), t)| \leqq p(t)\|x - y\|_t$$

for each (x, t) and (y, t) in $C[\alpha_0, t_1] \times [t_0, t_1]$. Then the solution to

$$\begin{aligned} \dot{x}(t) &= F(x(\cdot), t) + D(t) \quad \text{on} \quad [t_0, t_1], \\ x(t) &= \phi(t) \qquad\qquad\quad \text{on} \quad [\alpha_0, t_0] \end{aligned} \tag{1.2}$$

exists and is unique.

The proofs of Theorems 1 and 2 are straightforward extensions of standard proofs due to Carathéodory and will not be given here.

As an example to which Theorems 1 and 2 are applicable, consider the system

$$\begin{aligned} \dot{x}(t) &= \int_{\alpha_0}^{t} g(t, s)x(s)\, ds \quad \text{on} \quad [t_0, t_1], \\ x(t) &= \phi(t) \qquad\qquad\quad \text{on} \quad [\alpha_0, t_0], \end{aligned}$$

where $\phi \in C[\alpha_0, t_0]$, $g \in L_1([t_0, t_1] \times [\alpha_0, t_1])$. It is easy to show that this system satisfies the hypotheses of the theorems and hence the existence of a unique solution on $[t_0, t_1]$ is guaranteed. This system is a simple example of an integrodifferential equation. Such equations are important in the study of physical systems involving hereditary processes, such as magnetic hysteresis and elastic torsion, and have been discussed in detail by Volterra [9]. Another example where Theorems 1 and 2 are applicable and which is of interest to us in this paper will be discussed next.

Throughout this paper, we shall assume that $\theta(t)$ is a nonnegative real-valued C^1 function defined on R^1 satisfying $d_1 \leqq \dot{\theta}(t) \leqq d_2 < 1$ on R^1.

We define $\omega(t) = t - \theta(t)$. Then ω is an A. C. strictly increasing function with $\dot{\omega}(t) = 1 - \dot{\theta}(t) \geqq 1 - d_2 > 0$ and $\dot{\omega}(t) = 1 - \dot{\theta}(t) \leqq 1 - d_1$. Hence ω is C^1 with a bounded derivative. Since ω is strictly increasing, we have that $r = \omega^{-1}$ exists. In fact, it follows from the inverse mapping theorem that r is C^1. Furthermore, since $r(\omega(t)) = t$, we have $(dr/ds)(d\omega/dt) = 1$, which implies that $\dot{r}(s) > 0$. Moreover, $\dot{r}(s) = 1/\dot{\omega}(t) \geqq 1/(1 - d_1) > 0$ and $\dot{r}(s) \leqq 1/(1 - d_2) < \infty$. Hence $r = \omega^{-1}$ is a strictly increasing A. C. function. It is, in fact, C^1 with a bounded derivative. Also, ω and $r = \omega^{-1}$ are C^1 functions with derivatives bounded away from zero.

Under the above assumptions, the next two corollaries follow from applications and modifications of Theorem 1 and Theorem 2.

COROLLARY 2.1. *Let A and B be $n \times n$ measurable matrices and C be a measurable n-vector satisfying $|A(t)| \leqq m(t)$, $|B(t)| \leqq m(t)$, $|C(t)| \leqq m(t)$, where $m \in L_1(t_0, t_1)$. Let $\omega(t) = t - \theta(t)$, $\omega_0 = \omega(t_0)$. Assume ϕ is continuous on $[\omega_0, t_0]$. Then*

$$\dot{x}(t) = A(t)x(t) + B(t)x(\omega(t)) + C(t) \quad \text{on} \quad [t_0, t_1],$$

$$x(t) = \phi(t) \quad \text{on} \quad [\omega_0, t_0]$$

has a unique solution on $[t_0, t_1]$.

COROLLARY 2.2. *Let A and B be as in Corollary 2.1. Suppose ϕ to be continuous on $[\omega(\bar{t}), \bar{t}]$. Let $r = \omega^{-1}$. Then the advanced system*

$$\dot{y}(s) = A(s)y(s) + B(s)y(r(s)) \quad \text{on} \quad [t_0, \omega(\bar{t})],$$

$$y(s) = \phi(s) \quad \text{on} \quad [\omega(\bar{t}), \bar{t}]$$

has a solution on $[t_0, \bar{t}]$.

We next prove a lemma which will be needed in studying lagged equations.

LEMMA 1. *Suppose ω is C^1 on $[t_0, t_1]$ with $\dot{\omega}(t) \geqq \delta > 0$ for $t \in [t_0, t_1]$. Let f be measurable on $[\omega(t_0), \omega(t_1)]$. Then $f \circ \omega$ is measurable on $[t_0, t_1]$.*

Proof. From the hypotheses on ω, we have ω^{-1} exists as a 1–1 C^1 function on $[\omega(t_0), \omega(t_1)]$. Also, f measurable on $[\omega(t_0), \omega(t_1)]$ implies there exists a sequence $\{g_n\}$ of continuous functions on $[\omega(t_0), \omega(t_1)]$ such that g_n converges to f a.e. on $[\omega(t_0), \omega(t_1)]$. Let E be the subset of $[\omega(t_0), \omega(t_1)]$ such that $g_n(s) \to f(s)$ for s in E, where $\mu(E) = \omega(t_1) - \omega(t_0)$. Let $F = \omega^{-1}(E)$. Then $F \subset [t_0, t_1]$ and $\mu(F) = t_1 - t_0$, since the A. C. function ω^{-1} takes sets of measure zero into sets of measure zero. Now let $t \in F$. Then there exists $s \in E$ such that $s = \omega(t)$. But $s \in E$ implies $g_n(s) \to f(s)$. Hence, $g_n(\omega(t)) \to f(\omega(t))$ for $t \in F$ or $g_n \circ \omega \to f \circ \omega$ a.e. on $[t_0, t_1]$. But $g_n \circ \omega$ is measurable. Therefore $f \circ \omega$ is measurable on $[t_0, t_1]$.

Now let $\theta_1, \cdots, \theta_\nu$ be real-valued C^1 functions on R^1 satisfying the

same hypotheses as θ above. Define $\omega_j(t) = t - \theta_j(t)$ and $r_j(t) = \omega_j^{-1}(t)$, $j = 1, 2, \cdots, \nu$. Assume that

$$\theta_1(t) \leqq \theta_2(t) \leqq \cdots \leqq \theta_\nu(t)$$

so that $\omega_j(t) \leqq \omega_{j-1}(t)$ and $r_j(t) \geqq r_{j-1}(t)$. We also assume that there exist no increasing function ω and positive integers $k_1, k_2, \cdots, k_\nu$ such that

$$\omega_j(t) = \omega^{\circ k_j}(t), \qquad j = 1, 2, \cdots, \nu,$$

where $\omega^{\circ k} = \omega \circ \omega \circ \cdots \circ \omega$ (k times).[1]

THEOREM 3. *Let $A, B_j, j = 1, 2, \cdots, \nu$, be $n \times n$ measurable matrices and C be a measurable n-vector satisfying $|A(s)| \leqq m(s)$, $|B_j(s)| \leqq m(s)$, $|C(s)| \leqq m(s)$, where $m \in L_1(t_0, t_1)$. Let $t_0 < t \leqq t_1$. Let ω_j and $r_j = \omega_j^{-1}$ be as defined above. Define $\chi_j(s)$ to be the characteristic function of $[t_0, \omega_j(t)]$ (or $[\omega_j(t), t_0]$). Let $\Lambda(s, t)$ be a matrix solution to*

$$\Lambda(s, t) = 0, \qquad s > t,$$

$$\Lambda(t, t) = I,$$

$$\frac{\partial \Lambda}{\partial s}(s, t) + \Lambda(s, t)A(s) + \sum_{j=1}^{\nu} \chi_j(s)\Lambda(r_j(s), t)B_j(r_j(s))\dot{r}_j(s) = 0 \tag{1.3}$$

for $s \in [t_0, t]$.

Then the vector solution x to

$$\dot{x}(s) = A(s)x(s) + \sum_{j=1}^{\nu} B_j(s)x(\omega_j(s)) + C(s)$$

$$\text{for} \quad s \in [t_0, t_1], \tag{1.4}$$

$$x(s) = \phi(s) \qquad \text{for} \quad s \in [\omega_\nu(t_0), t_0],$$

where $\phi \in C[\omega_\nu(t_0), t_0]$, is given by

$$x(t) = \Lambda(t_0, t)\phi(t_0) + \int_{t_0}^{t} \Lambda(s, t)C(s)\, ds + \sum_{j=1}^{\nu} \int_{\omega_j(t_0)}^{t_0} \Lambda(r_j(s), t)B_j(r_j(s))\phi(s)\dot{r}_j(s)\, ds$$

for $t > t_0$.

Proof. Let $\Lambda(s, t)$ be a solution to (1.3). That such a solution exists follows from an easy extension of Corollary 2.2. Multiplying the first equa-

[1] The author is indebted to the referee for pointing out that without this last assumption, the control problem with lags to be discussed below can be reduced to a control problem involving ordinary differential equations.

tion in (1.4) by $\Lambda(s, t)$ and integrating over s from t_0 to t gives

$$\int_{t_0}^{t} \Lambda(s, t)\dot{x}(s)\, ds = \int_{t_0}^{t} \Lambda(s, t)A(s)x(s)\, ds + \int_{t_0}^{t} \Lambda(s, t)C(s)\, ds + \sum_{j=1}^{\nu} \int_{t_0}^{t} \Lambda(s, t)B_j(s)x(\omega_j(s))\, ds.$$

Since $\Lambda(s, t)$ is A. C. in s, we can integrate by parts in the term on the left side of the equation obtaining

$$\int_{t_0}^{t} \Lambda(s, t)\dot{x}(s)\, ds = x(t) - \Lambda(t_0, t)x(t_0) - \int_{t_0}^{t} \frac{\partial \Lambda}{\partial s}(s, t)x(s)\, ds.$$

Consider $\int_{t_0}^{t} \Lambda(s, t)B_j(s)x(\omega_j(s))\, ds$. Letting $\tau = \omega_j(s)$ or $s = r_j(\tau)$ in this integral gives

$$\int_{t_0}^{t} \Lambda(s, t)B_j(s)x(\omega_j(s))\, ds = \int_{\omega_j(t_0)}^{\omega_j(t)} \Lambda(r_j(\tau), t)B_j(r_j(\tau))x(\tau)\dot{r}_j(\tau)\, d\tau$$

$$= \int_{\omega_j(t_0)}^{t_0} \Lambda(r_j(s), t)B_j(r_j(s))x(s)\dot{r}_j(s)\, ds + \int_{t_0}^{t} \chi_j(s)\Lambda(r_j(s), t)B_j(r_j(s))x(s)\dot{r}_j(s)\, ds.$$

Combining these results and the fact that $x(s) = \phi(s)$ on $[\omega_\nu(t_0), t_0]$, we obtain

$$x(t) = \Lambda(t_0, t)\phi(t_0) + \int_{t_0}^{t} \Lambda(s, t)C(s)\, ds + \int_{t_0}^{t} \left\{ \frac{\partial \Lambda}{\partial s}(s, t) + \Lambda(s, t)A(s) + \sum_{j=1}^{\nu} \chi_j(s)\Lambda(r_j(s), t)B_j(r_j(s))\dot{r}_j(s) \right\} x(s)\, ds + \sum_{j=1}^{\nu} \int_{\omega_j(t_0)}^{t_0} \Lambda(r_j(s), t)B_j(r_j(s))\phi(s)\dot{r}_j(s)\, ds.$$

The term in brackets vanishes almost everywhere on $[t_0, t]$ since $\Lambda(s, t)$ satisfies (1.3). Hence we have the desired results.

To conclude this section, we prove a lemma of du Bois-Raymond type.

LEMMA 2. *Let* $[a, b] \subset R^1$. *Let* $M(t) > 0$ *be an arbitrary but fixed function in* $L_1(a, b)$. *Define* $D(M)_b = \{g : g$ is A. C. on $[a, b]$, $|\dot{g}(t)| \leqq M(t)$, and

$g(b) = 0\}$. *Suppose* $f \in L_1(a, b)$. *Then* $\int_a^b fg\,dt \leqq 0$ *for each* $g \in D(M)_b$ *implies* $f(t) = 0$ *a.e. on* $[a, b]$.

Proof. Suppose $f \in L_1(a, b)$ and $\int_a^b fg\,dt \leqq 0$ for $g \in D(M)_b$. Define $F(t) = \int_a^t f(s)\,ds$. Then F is A. C. on $[a, b]$ and $\int_a^b \dot{F}(t)g(t)\,dt \leqq 0$ for $g \in D(M)_b$. Since both F and g are A. C., we may integrate by parts obtaining $\int_a^b F(t)\dot{g}(t)\,dt \geqq 0$ since $g(b) = F(a) = 0$. This last inequality implies $F(t) = 0$ on (a, b). Suppose not. Then there exist t_0 in (a, b) and $\delta > 0$ such that $F(t) > 0$ on $(t_0 - \delta, t_0 + \delta)$. (The proof for $F(t_0) < 0$ is similar.) We choose a particular g in $D(M)_b$ as follows:

$$g(t) = 0 \quad \text{for} \quad t \geqq t_0 + \delta,$$

$$\dot{g}(t) < 0 \quad \text{for} \quad t \in (t_0 - \delta, t_0 + \delta) \quad (\text{with} \quad |\dot{g}| \leqq M),$$

$$\dot{g}(t) = 0 \quad \text{for} \quad t \leqq t_0 - \delta.$$

Then for this g we have $\int_a^b F\dot{g}\,dt = \int_{t_0-\delta}^{t_0+\delta} F\dot{g}\,dt < 0$, which is a contradiction. Hence, $F(t) = 0$ on (a, b). This implies $f(s) = 0$ a.e. on $[a, b]$.

2. Formulation of problem and statement of necessary conditions. Suppose t_0 is fixed in R^1. Let $\bar{x} = (x^1, \cdots, x^{n-1})$, $\bar{y}_j = (y_j^1, \cdots, y_j^{n-1})$, $j = 1, 2, \cdots, \nu$, $\bar{g} = (g^1, \cdots, g^{n-1})$, and $u = (u^1, \cdots, u^v)$. Let $\bar{Y} = (\bar{y}_1, \cdots, \bar{y}_\nu)$; that is, $\bar{Y}$ is a ν-vector, each component of which is an $(n-1)$-vector. Given a set S in R^{n-1}, the notation $S \times S^\nu$ will mean $S \times S \times \cdots \times S$ ($\nu + 1$ times). Let I be a bounded open interval in R^1 containing $[\omega_\nu(t_0), t_0]$. Put $I' = I \cap \{t : t > t_0\}$. Let $\bar{G}$ be an open convex region in R^{n-1}. (Although the notation $\bar{G}$ usually means the closure of G, in this paper this is not so unless specifically stated.) Suppose that T is a given C^1 differentiable manifold in R^{2n-1}. We assume that T has dimension $< 2n - 1$ and that $T \subset \bar{G} \times \bar{G} \times I'$. Points in T will be written $(\bar{x}_0, \bar{x}_1, t_1)$. Let R_U be a given subset of R^v. Suppose that $g^i = g^i(\bar{x}, \bar{y}_1, \cdots, \bar{y}_\nu, u, t) = g^i(\bar{x}, \bar{Y}, u, t)$, $i = 0, 1, \cdots, n-1$, are defined on $\bar{G} \times \bar{G}^\nu \times R_U \times I'$ with range in R^1. Further, let each g^i be C^1 in $\bar{x}$, $\bar{Y}$ and Borel measurable in u, t. Let U be a mapping defined on I:

$$U : t \in I \to U(t) \subset R_U \subset R^v.$$

Put $\Omega = \{u : u$ is measurable on I, $u(t) \in U(t)$ for each $t \in I\}$. Let $\tilde{g} = (g^0, \bar{g})$. We make the further assumption on the g^i's that:

For every compact $\bar{X} \subset \bar{G}$ and $u \in \Omega$, there exists an $\bar{m} \in L_1(I')$ ($\bar{m}$ de-

pending on $\bar{X}$, u) such that

$$|\tilde{g}(\bar{x}, \bar{Y}, u(t), t)| \leqq \bar{m}(t),$$
$$|\tilde{g}_{\bar{x}}(\bar{x}, \bar{Y}, u(t), t)| \leqq \bar{m}(t),$$
$$|\tilde{g}_{\bar{y}_j}(\bar{x}, \bar{Y}, u(t), t)| \leqq \bar{m}(t), \qquad j = 1, \cdots, \nu,$$

for each $(\bar{x}, \bar{Y}) \in \bar{X} \times \bar{X}^\nu$ and each $t \in I'$.

Let $\bar{\Phi}$ be the class of A. C. $(n-1)$-vector functions on $[\omega_\nu(t_0), t_0]$ into $\bar{G}$. That is, $\bar{\Phi} = \{\bar{\phi} : \bar{\phi} \in AC([\omega_\nu(t_0), t_0], \bar{G})\}$.

PROBLEM. Minimize

$$J[\bar{\phi}, u, \bar{x}, t_1] = \int_{t_0}^{t_1} g^0(\bar{x}(t), \bar{x}(\omega_1(t)), \cdots, \bar{x}(\omega_\nu(t)), u(t), t)\, dt$$

over $\bar{\Phi} \times \Omega \times R^{n-1} \times I'$ subject to

(a) $\dot{\bar{x}}(t) = \tilde{g}(\bar{x}(t), \bar{x}(\omega_1(t)), \cdots, \bar{x}(\omega_\nu(t)), u(t), t)$ for $t \in [t_0, t_1]$,
$\bar{x}(t) = \bar{\phi}(t)$ for $t \in [\omega_\nu(t_0), t_0]$,

(b) $(\bar{x}(t_0), \bar{x}(t_1), t_1) \in T$.

A solution $(\bar{\phi}^*, u^*, \bar{x}^*, {t_1}^*)$ to the above problem will be called an optimal solution and the corresponding solution $\bar{x}^*$ of (a) will be called an optimal trajectory. We now state necessary conditions for an optimal solution.

THEOREM 4. *Let $(\bar{\phi}^*, u^*, \bar{x}^*, {t_1}^*)$ be an optimal solution. In addition to the previous assumptions, suppose that ${t_1}^*$ is a regular (Lebesgue) point for $\tilde{g}^*(t) \equiv \tilde{g}(\bar{x}^*(t), \bar{x}^*(\omega_1(t)), \cdots, \bar{x}^*(\omega_\nu(t)), u^*(t), t)$. Then there exists a nontrivial A. C. n-vector function $\tilde{\psi}(t) = (\psi^0(t), \bar{\psi}(t)) = (\psi^0(t), \psi^1(t), \cdots, \psi^{n-1}(t))$ defined on $[t_0, {t_1}^*]$ satisfying:*

(i) $\psi^0(t) = \text{const.} \leqq 0$,
$\dot{\tilde{\psi}}(t) + \tilde{\psi}(t)\tilde{g}_{\bar{x}}^*(t) + \sum_{j=1}^{\nu} \chi_j(t)\tilde{\psi}(r_j(t))\tilde{g}_{\bar{y}_j}^*(r_j(t))\dot{r}_j(t) = 0$ *on* $[t_0, {t_1}^*]$, *where χ_j is the characteristic function for* $[t_0, \omega_j({t_1}^*)]$.

(ii) $\int_{t_0}^{{t_1}^*} \tilde{\psi}(t) \cdot \tilde{g}^*(t)\, dt$
$\geqq \int_{t_0}^{{t_1}^*} \tilde{\psi}(t) \cdot \tilde{g}(\bar{x}^*(t), \bar{x}^*(\omega_1(t)), \cdots, \bar{x}^*(\omega_\nu(t)), u(t), t)\, dt$ *for all* $u \in \Omega$,

(iii) $\sum_{i=0}^{j-1} \tilde{\psi}(r_{\nu-i}(t))\tilde{g}_{\bar{y}_{\nu-i}}^*(r_{\nu-i}(t))\dot{r}_{\nu-i}(t) = 0$ *a. e. on* $[\omega_{\nu+1-j}(t_0), \omega_{\nu-j}(t_0)]$ *for* $j = 1, 2, \cdots, \nu$, *where* $\omega_0(t) \equiv t$.

(iv) *The $(2n-1)$-dimensional vector* $(-\bar{\psi}(t_0), \bar{\psi}({t_1}^*), -\tilde{\psi}({t_1}^*) \cdot \tilde{g}^*({t_1}^*))$ *is orthogonal to the tangent plane to T at* $(\bar{x}^*(t_0), \bar{x}^*({t_1}^*), {t_1}^*)$.

(v) $\tilde{\psi}(t) = (\psi^0, \bar{\psi}(t))$ *is nonzero on* $[\omega_1({t_1}^*), {t_1}^*]$. $\tilde{\psi}$ *can have zeros in* $[t_0, \omega_1({t_1}^*)]$. *If, however, we have*
(a) $\theta_\nu(t) > \theta_{\nu-1}(t) > \cdots > \theta_1(t) > 0$ *on* $[t_0, {t_1}^*]$,
(b) $\tilde{g}_{\bar{y}_j}^*(t)$ *is nonsingular for* $t \in [t_0, {t_1}^*]$, $j = 1, 2, \cdots, \nu$,
then $\psi^0 \not\equiv 0$ *and hence* $\tilde{\psi}$ *is nonzero on* $[t_0, {t_1}^*]$.

The proof of Theorem 4, which will follow from Theorem 5, will be given

in §5. First, however, let us make a few remarks about the problem and theorem discussed above.

Remark 1. The conditions that $u(t) \in U(t)$ (see definition of Ω) are simply a way of putting constraints on the control functions. For example, if $U(t) = \{u \in R^v : |u| \leqq \rho(t)\}$, then $U(t)$ is the ball of radius $\rho(t)$ in R^v (where ρ is some given function) and the constraints on $u(t)$ are $\rho(t) - |u(t)| \geqq 0$.

Remark 2. T as defined above is a manifold consisting of possible initial state values $(\bar{x}_0)$ and possible terminal state and time values $(\bar{x}_1, t_1)$. Statement (iv) of Theorem 4 is just the well-known transversality condition. As it will be seen in the following proofs, the assumption that t_1^* be a regular point of $\tilde{g}$ is needed only to prove (iv).

Remark 3. Statement (ii) of Theorem 4 is the maximum principle in integral form. Under the additional assumptions that (a) $U(t) = U$, a fixed subset of R_U, for each $t \in I$, and (b) $\tilde{g} = (g^0, \bar{g})$ is continuous in all of its arguments, one can show that statement (ii) implies a Pontryagin type maximum principle, i.e., a pointwise maximum principle (see [5]).

In §3 and §4 we shall formulate a general extremal problem similar to that of Gamkrelidze and give necessary conditions (Theorem 5) for solutions to be extremals.

3. Formulation of an extremal problem. In §3 and §4, $x, y, y_1, \cdots, y_\nu, z$ will denote n-vectors. The ν-vector $(y_1, \cdots, y_\nu)$ will be denoted by Y. Let t_0 be fixed in R^1 and let I and I' be the subsets of R^1 as defined in §2. Let G be an open convex region in R^n. We shall denote by F a family of n-vector functions $f(x, Y, t) = f(x, y_1, \cdots, y_\nu, t)$ defined on $G \times G^\nu \times I'$. For any integer k we define $P^k = \{\alpha : \alpha = (\alpha_1, \cdots, \alpha_k), \alpha_i \geqq 0, \sum_1^k \alpha_i = 1\}$.

DEFINITION 3.1. The family F is *quasi-convex* if:

1. $f(x, Y, t)$ is C^1 in x, Y for fixed $t \in I'$. $f(x, Y, t)$ is measurable on I' for fixed x, Y.

2. For any $f \in F$ and any compact convex X contained in G, there is an integrable function m (depending on X, f) such that $|f(x, Y, t)| \leqq m(t)$, $|f_x(x, Y, t)| \leqq m(t)$, and $|f_{y_j}(x, Y, t)| \leqq m(t)$, $j = 1, \cdots, \nu$, for all $(x, Y) \in X \times X^\nu$ and all $t \in I'$.

3. For every compact convex X contained in G and finite collection $f_1, f_2, \cdots, f_k$ in F and $\epsilon > 0$, there exists for each $\alpha \in P^k$ an $f_\alpha \in F$ (depending on X, the f_i, ϵ) such that $g(x, Y, t; \alpha) = \sum_1^k \alpha_i f_i(x, Y, t) - f_\alpha(x, Y, t)$ satisfies:

   (i) There exists an $\bar{m} \in L_1(I')$, depending on X and the f_i, such that $|g(x, Y, t; \alpha)| < \bar{m}(t)$, $|g_x(x, Y, t; \alpha)| < \bar{m}(t)$, and $|g_{y_j}(x, Y, t; \alpha)| < \bar{m}(t)$, $j = 1, 2, \cdots, \nu$, for all $(x, Y) \in X \times X^\nu$, $\alpha \in P^k$, and $t \in I'$.

(ii) $\left|\int_{\tau_1}^{\tau_2} g(x, Y, t; \alpha)\, dt\right| < \epsilon$ for all $(x, Y) \in X \times X^\nu$, $\alpha \in P^k$, and τ_1, τ_2 in I'.

(iii) If $\{\alpha^i\}$ is a sequence in P^k such that $\alpha^i \to \bar{\alpha} \in P^k$, then $\{g(x, Y, t; \alpha^i)\}$ converges in measure on I' as a function of t to $g(x, Y, t; \bar{\alpha})$ for all $(x, Y) \in X \times X^\nu$.

Define $\Phi = \{\phi : \phi \in AC([\omega_\nu(t_0), t_0], G)\}$. For $f \in F$ and $\phi \in \Phi$ consider the solution to

$$\begin{aligned} \dot{x}(t) &= f(x(t), x(\omega_1(t)), \cdots, x(\omega_\nu(t)), t) \quad \text{for} \quad t > t_0, \\ x(t) &= \phi(t) \qquad\qquad\qquad\qquad\qquad\quad \text{for} \quad t \in [\omega_\nu(t_0), t_0]. \end{aligned} \tag{3.1}$$

Let $z(t)$, $t_0 \leqq t \leqq \tau_1$, be a solution to (3.1) corresponding to $(f, \phi) \in F \times \Phi$. Define $q_z = (z(t_0), z(\tau_1), \tau_1)$. Then $q_z \in R^{2n+1}$. Define $Q =$ the set of all such q_z for solutions of (3.1) for $(f, \phi) \in F \times \Phi$. Let N be a given C^1 manifold in R^{2n+1} with boundary $\tilde{M} = \partial N$. For $q \in \tilde{M}$ let $N_T(q)$ be the tangent half-plane to N at q and let $M_T(q)$ be the tangent plane to $\tilde{M}$ at q.

DEFINITION. A solution $z(t)$, $t_0 \leqq t \leqq \tau$, to (3.1) corresponding to $(f, \phi) \in F \times \Phi$ is called an F, N, Φ *extremal* if

(i) $q_z \in \tilde{M}$,

(ii) there exists a neighborhood V of q_z such that

$$V \cap Q \cap N \subset \tilde{M}.$$

THEOREM 5. *Suppose F is quasi-convex. Let $x(t)$, $t_0 \leqq t \leqq t_1$, be an F, N, Φ extremal corresponding to $(\hat{f}, \hat{\phi}) \in F \times \Phi$. Suppose t_1 is a regular point for $\hat{f}(t) = \hat{f}(x(t), x(\omega_1(t)), \cdots, x(\omega_\nu(t)), t)$. Then there exists a nontrivial A. C. n-vector function $\psi(t)$ defined on $[t_0, t_1]$ satisfying the following:*

(i) $\dot{\psi}(t) + \psi(t)\hat{f}_x(t) + \sum_{j=1}^{\nu} \chi_j(t)\psi(r_j(t))\hat{f}_{y_j}(r_j(t))\dot{r}_j(t) = 0$ *on $[t_0, t_1]$, where χ_j is the characteristic function for $[t_0, \omega_j(t_1)]$.*

(ii) $\int_{t_0}^{t_1} \psi(t) \cdot \hat{f}(x(t), x(\omega_1(t)), \cdots, x(\omega_\nu(t)), t)\, dt$

$\geqq \int_{t_0}^{t_1} \psi(t) \cdot f(x(t), x(\omega_1(t)), \cdots, x(\omega_\nu(t)), t)\, dt$ *for all $f \in F$.*

(iii) $\sum_{i=0}^{j-1} \psi(r_{\nu-i}(t))\hat{f}_{y_{\nu-i}}(r_{\nu-i}(t))\dot{r}_{\nu-i}(t) = 0$ *a. e. on $[\omega_{\nu+1-j}(t_0), \omega_{\nu-j}(t_0)]$ for $j = 1, 2, \cdots, \nu$.*

(iv) *The $(2n + 1)$-dimensional vector $(-\psi(t_0), \psi(t_1), -\psi(t_1) \cdot \hat{f}(t_1))$ is orthogonal to $\tilde{M}$ at $q_x = (x(t_0), x(t_1), t_1)$.*

4. Proof of Theorem 5. In this section we shall give the proof of Theorem 5. However, to simplify the notation, we shall consider the case with a single lag (i.e., $\nu = 1$) $\omega(t) = t - \theta(t)$ in most of the proof below. We shall drop the subscripts on θ_1, ω_1, y_1 and call them θ, ω, y, respectively. In Definition

3.1, for example, we then replace $f(x, Y, t)$ by $f(x, y, t)$. It will be clear that the same arguments given below will work with only slight modifications for the case where $\nu > 1$, ν finite. We shall return to the case of multiple lags only when it becomes necessary (i.e., when the arguments for the case $\nu > 1$ are substantially different from those for $\nu = 1$). Before beginning the proof of Theorem 5, we first prove a lemma that will be needed.

Let m^* be a nonnegative real-valued function defined on I with $m^* \in L_1(I)$.

DEFINITION 4.1. A family F is *m^*-quasiconvex* if it satisfies the conditions in Definition 3.1 with 3(ii) replaced by

(ii′) $\left| \int_{\tau_1}^{\tau_2} g(x(t), x(\omega(t)), t; \alpha)\, dt \right| < \epsilon$ for every $\alpha \in P^k$, $\tau_1, \tau_2 \in I'$, and any A. C. $x(t)$, $x: I \to X$, with $|\dot{x}(t)| \leqq m^*(t)$ a. e. on I.

LEMMA 4.1. *F quasi-convex implies F m^*-quasi-convex for every nonnegative $m^* \in L_1(I)$.*

Note that the converse of Lemma 4.1 is not true.

Proof. Suppose F is quasi-convex. Let X be a given compact convex subset of G; let $\{f_i\}_{i=1}^k$ in F and $\epsilon > 0$ be given. Let m^* be any nonnegative function, $m^* \in L_1(I)$. Then there exists $\delta_1 > 0$ such that $\int_E m^*(t)\, dt < \epsilon \Big/ \left(8 \int_{I'} \bar{m}\right)$ for $E \subset I$, $\mu(E) < \delta_1$, where $\bar{m}$ is the $L_1(I')$ function depending on X, $\{f_i\}_1^k$ in statement 3(i) of Definition 3.1.

Since θ is A. C. on I', there exists $\delta_2 > 0$ such that $|\theta(t) - \theta(s)| < \delta_1$, whenever $t, s \in I'$, $|t - s| < \delta_2$. Let $s_0, s_1, \cdots, s_p$ be a partition P^* of I' with $s_i < s_{i+1}$ such that $|s_{i+1} - s_i| < \delta \equiv \min\{\delta_1, \delta_2\}$. Let $x(t)$ be any A. C. mapping on I into X satisfying $|\dot{x}(t)| \leqq m^*(t)$ a.e. on I. Then for $t \in [s_i, s_{i+1}]$ we have

$$|x(t) - x(s_i)| = \left| \int_{s_i}^{t} \dot{x}(s)\, ds \right| \leqq \int_{s_i}^{t} m^*(s)\, ds < \epsilon \Big/ \left(8 \int_{I'} \bar{m}\right).$$

Since $x(\omega(t))$ is A. C. in t (x is A. C., ω is A. C. strictly increasing), we have for $t \in [s_i, s_{i+1}]$,

$$|x(\omega(t)) - x(\omega(s_i))| = \left| \int_{s_i}^{t} \frac{d}{ds} x(\omega(s))\, ds \right| = \left| \int_{s_i}^{t} \dot{x}(\omega(s))\dot{\omega}(s)\, ds \right|,$$

and letting $\sigma = \omega(s)$, we obtain

$$\left| \int_{\omega(s_i)}^{\omega(t)} \dot{x}(\sigma)\, d\sigma \right| \leqq \int_{\omega(s_i)}^{\omega(t)} m^*(\sigma)\, d\sigma = \int_{s_i - \theta(s_i)}^{t-\theta(t)} m^*(\sigma)\, d\sigma$$

$$= \int_{s_i - \theta(s_i)}^{t-\theta(s_i)} m^*(\sigma)\, d\sigma + \int_{t-\theta(s_i)}^{t-\theta(t)} m^*(\sigma)\, d\sigma$$

$$< \epsilon \Big/ \left(8 \int_{I'} \bar{m}\right) + \epsilon \Big/ \left(8 \int_{I'} \bar{m}\right) = \epsilon \Big/ \left(4 \int_{I'} \bar{m}\right)$$

since $|t - s_i| < \delta = \min\{\delta_1, \delta_2\}$ and $|\theta(t) - \theta(s_i)| < \delta_1$.

From the quasi-convexity of F we get the existence of an $f_\alpha \in F$ for each $\alpha \in P^k$ such that $g(x, y, t; \alpha) = \sum_{i=1}^{k} \alpha_i f_i(x, y, t) - f_\alpha(x, y, t)$ satisfies 3(ii) of Definition 3.1 with ϵ replaced by $\epsilon/2(p + 2)$. Let τ_1, τ_2 be any two points of I'. Adding these two points to P^* and reindexing where necessary, we get a partition $s_0 \leqq s_1 \leqq \cdots \leqq s_{p+2}$ of I' containing τ_1, τ_2 and such that $s_{i+1} - s_i < \delta$ for $i = 0, 1, \cdots, p + 1$.

Then we have

$$\left|\int_{\tau_1}^{\tau_2} g(x(t), x(\omega(t)), t; \alpha)\, dt\right| \leqq \sum_{i=0}^{p+1} \left|\int_{s_i}^{s_{i+1}} g(x(t), x(\omega(t)), t; \alpha)\, dt\right|. \tag{4.1}$$

But

$$\begin{aligned} &\sum_{i=0}^{p+1} \left|\int_{s_i}^{s_{i+1}} g(x(t), x(\omega(t)), t; \alpha)\, dt\right| \\ &\leqq \sum_{i=0}^{p+1} \left[\left|\int_{s_i}^{s_{i+1}} g(x(s_i), x(\omega(s_i)), t; \alpha)\, dt\right| \right. \\ &\quad \left. + \left|\int_{s_i}^{s_{i+1}} \{g(x(t), x(\omega(t)), t; \alpha) - g(x(s_i), x(\omega(s_i)), t; \alpha)\}\, dt\right|\right]. \end{aligned} \tag{4.2}$$

By 3(ii) of Definition 3.1 we obtain

$$\left|\int_{s_i}^{s_{i+1}} g(x(s_i), x(\omega(s_i)), t; \alpha)\, dt\right| < \epsilon/2(p + 2) \quad \text{for each} \quad i = 0, 1, \cdots, p + 1.$$

Also,

$$\begin{aligned} &\sum_{i=0}^{p+1} \left|\int_{s_i}^{s_{i+1}} \{g(x(t), x(\omega(t)), t; \alpha) - g(x(s_i), x(\omega(s_i)), t; \alpha)\}\, dt\right| \\ &\leqq \sum_{i=0}^{p+1} \int_{s_i}^{s_{i+1}} \left[\max_{(x,y)\in X\times X} |g_x(x, y, t; \alpha)|\, |x(t) - x(s_i)|\right. \\ &\quad \left. + \max_{(x,y)\in X\times X} |g_y(x, y, t; \alpha)|\, |x(\omega(t)) - x(\omega(s_i))|\right] dt \\ &< \left\{\epsilon \Big/ \left(4 \int_{I'} \bar{m}\right)\right\} \sum_{i=0}^{p+1} \int_{s_i}^{s_{i+1}} \{\bar{m}(t) + \bar{m}(t)\}\, dt = \epsilon/2. \end{aligned}$$

Combining these two estimates with (4.1) and (4.2) one obtains

$$\left|\int_{\tau_1}^{\tau_2} g(x(t), x(\omega(t)), t; \alpha)\, dt\right| < \sum_{i=0}^{p+1} \epsilon/2(p+2) + \epsilon/2 = \epsilon,$$

which proves the lemma.

To prove Theorem 5, let $x(t)$, $t_0 \leqq t \leqq t_1$, be the F, N, Φ extremal of the theorem corresponding to $(\hat{f}, \hat{\phi}) \in F \times \Phi$. That is,

$$\begin{aligned} \dot{x}(t) &= \hat{f}(x(t), x(\omega(t)), t) \quad \text{for} \quad t \in [t_0, t_1], \\ x(t) &= \hat{\phi}(t) \qquad\qquad\qquad\quad \text{for} \quad t \in [\omega_0, t_0]. \end{aligned} \tag{4.3}$$

Let X be a fixed compact convex subset of G chosen so that each $x(t)$, $\omega_0 \leqq t \leqq t_1$, is an interior point of X, i.e., $x(t) \in X^\circ$. Denote by $[F]$ the closed convex hull of F and by $[F] - \hat{f}$ the set of elements of the form $f - \hat{f}$, where $f \in [F]$. Note that $[F] - \hat{f} = [F - \hat{f}]$ is a convex set. Let $\delta f = \sum_{i=1}^{k} \alpha_i f_i - \hat{f}$ represent an arbitrary element of $[F] - \hat{f}$ (i.e., $\alpha \in P^k$, $\{f_i\}_1^k$ in F). Let $\bar{m}$ be the $L_1(I')$ function of 3(i) in the definition of quasi-convexity for X and $f_1, \cdots, f_k, \hat{f}$.

Let $m \in L_1(I')$ be such that

$$\begin{aligned} |f_i(x, y, t)| &\leqq m(t), & |\hat{f}(x, y, t)| &\leqq m(t), \\ |f_{ix}(x, y, t)| &\leqq m(t), & |\hat{f}_x(x, y, t)| &\leqq m(t), \\ |f_{iy}(x, y, t)| &\leqq m(t), & |\hat{f}_y(x, y, t)| &\leqq m(t), \\ & & & i = 1, 2, \cdots, k, \end{aligned}$$

for all $(x, y) \in X \times X$ and $t \in I'$.

Since F is quasi-convex, it follows from the definition and Lemma 4.1 that given ϵ, $0 \leqq \epsilon \leqq 1$, there exists $g_\epsilon(x, y, t)$ defined on $G \times G \times I'$ such that:

$$(\hat{f} + \epsilon\delta f + g_\epsilon) \in F; \tag{4.4}$$

$$\begin{aligned} |g_\epsilon(x, y, t)| &< \bar{m}(t), \\ |g_{\epsilon x}(x, y, t)| &< \bar{m}(t), \\ |g_{\epsilon y}(x, y, t)| &< \bar{m}(t) \end{aligned} \tag{4.5}$$

for all $(x, y) \in X \times X$ and $t \in I'$;

$$\left|\int_{\tau_1}^{\tau_2} g_\epsilon(z(t), z(\omega(t)), t)\, dt\right| < \epsilon^2 \tag{4.6}$$

for every solution $z(t)$ of

$$\dot{z}(t) = \hat{f}(z(t),z(\omega(t)),t) + \epsilon\delta f(z(t),z(\omega(t)),t) + g_\epsilon(z(t),z(\omega(t)),t)$$

that is sufficiently near $x(t)$ (so that $z(t), z(\omega(t)) \in X$) and all $[\tau_1, \tau_2] \subset I'$ on which $z(t), z(\omega(t))$ are defined. (We shall show later that such solutions actually exist.)

Hereafter, we assume that for each $\delta f \in ([F] - \hat{f})$ and each ϵ, $0 \leqq \epsilon \leqq 1$, such a $g_\epsilon(x, y, t)$ has been chosen.

Let M be an arbitrary but fixed positive function in $L_1(\omega_0, t_0)$. Define $D(M)$ to be the subset of A. C. n-vector functions $\delta\phi$ on $[\omega_0, t_0]$ into R^n satisfying $|\delta\dot{\phi}(t)| \leqq M(t)$ a.e. on $[\omega_0, t_0]$. The topology in $D(M)$ will be taken as the one induced by the sup norm in $AC([\omega_0, t_0], R^n)$; that is, $\|\delta\phi\| = \sup\{|\delta\phi(t)| : t \in [\omega_0, t_0]\}$. Thus,

$$D(M) = \{\delta\phi \in AC([\omega_0, t_0], R^n) : |\delta\dot{\phi}(t)| \leqq M(t) \quad \text{a.e. on} \quad [\omega_0, t_0]\}.$$

Note that $D(M)$ is a convex subset of $AC([\omega_0, t_0], R^n)$.

Recall that $\hat{\phi} \in \Phi$, hence $\hat{\phi}(t)$, $t \in [\omega_0, t_0]$, is in G. Therefore, for a given $\delta\phi \in D(M)$ we have $(\hat{\phi} + \epsilon\delta\phi) \in \Phi$ for ϵ sufficiently small. In fact, since $x(t)$, $\omega_0 \leqq t \leqq t_1$, is in X°, we have $\hat{\phi}(t)$, $\omega_0 \leqq t \leqq t_0$, in X°. Thus given $\delta\phi \in D(M)$, $\hat{\phi}(t) + \epsilon\delta\phi(t)$, $\omega_0 \leqq t \leqq t_0$, is in X° for ϵ sufficiently small.

We now "perturb" system (4.3). Let $\delta t \in R^1$, $\delta\phi \in D(M)$, $\delta f \in ([F] - \hat{f})$ be arbitrary. Then consider, for $0 \leqq \epsilon \leqq 1$, the system

$$\begin{aligned} \dot{z}(t, \epsilon) &= \hat{f}(z(t, \epsilon), z(\omega(t), \epsilon), t) + \epsilon\delta f(z(t, \epsilon), z(\omega(t), \epsilon), t) \\ &\qquad + g_\epsilon(z(t, \epsilon), z(\omega(t), \epsilon), t) \quad \text{for} \quad t > t_0, \\ z(t, \epsilon) &= \hat{\phi}(t) + \epsilon\delta\phi(t) \qquad \text{for} \quad t \in [\omega_0, t_0]. \end{aligned} \tag{4.7}$$

Note that it follows easily from previous definitions that there exists $m^* \in L_1(I)$ such that $|\dot{z}(t, \epsilon)| \leqq m^*(t)$ for solutions sufficiently close to $x(t)$.

LEMMA 4.2. *Let $x(t)$ be the solution to*

$$\begin{aligned} \dot{x}(t) &= \hat{f}(x(t), x(\omega(t)), t) \quad \textit{for} \quad t \in [t_0, t_1], \\ x(t) &= \hat{\phi}(t) \qquad \textit{for} \quad t \in [\omega_0, t_0]. \end{aligned}$$

Let $\delta t \in R^1$ be given and let X be a compact convex set containing $x(t)$, $\omega_0 \leqq t \leqq t_1$, in its interior. Suppose $Y(x, y, t)$ and $Z(x, y, t; \epsilon)$ are defined for $(x, y) \in X \times X$, $t \in I'$, $0 \leqq \epsilon \leqq 1$, and satisfy (4.5) for some $\bar{m} \in L_1(I')$. Let Z satisfy (4.6) with g_ϵ and δf replaced by Z and Y. Let $\Delta \in D(M)$.

If for each ϵ, $0 \leqq \epsilon < \epsilon_0 \leqq 1$, the solution $z(t, \epsilon)$ to

$$\begin{aligned} \dot{z}(t, \epsilon) &= \hat{f}(z(t, \epsilon), z(\omega(t), \epsilon), t) + \epsilon Y(z(t, \epsilon), z(\omega(t), \epsilon), t) \\ &\qquad + Z(z(t, \epsilon), z(\omega(t), \epsilon), t; \epsilon) \quad \textit{for} \quad t > t_0, \\ z(t, \epsilon) &= x(t) + \epsilon\Delta(t) \qquad \textit{for} \quad t \in [\omega_0, t_0], \end{aligned}$$

exists on $[t_0, \tau_\epsilon]$ and is interior to X on $[\omega_0, \tau_\epsilon]$, then there exists $\epsilon_2 > 0$ such that for each ϵ, $0 \leqq \epsilon < \epsilon_2 \leqq \epsilon_0$, $z(t, \epsilon)$ exists on $[t_0, t_1 + \epsilon_2 |\delta t|]$, and is in X°.

Proof. Since $x(t)$, $t_0 \leqq t \leqq t_1$, is in X°, it follows from slight modifications of Theorems 2.1 and 2.8 in [4] that given $\delta t \in R^1$, there exists $\epsilon_0' > 0$ such that $x(t)$ may be extended to $[t_0, t_1 + \epsilon_0' |\, \delta t\,|]$ with $x(t)$, $t_0 \leqq t \leqq t_1 + \epsilon_0' |\, \delta t\,|$, in X°. Let $\epsilon_1 = \min\{\epsilon_0, \epsilon_0'\}$, where ϵ_0 is as stated in the hypotheses of the lemma. Then for each ϵ, $0 \leqq \epsilon < \epsilon_1$, we have $z(t, \epsilon)$ exists on $[t_0, \tau_\epsilon]$ and is in X°.

For $t \in [t_0, \tau_\epsilon]$ we have

$$\begin{aligned} z(t,\epsilon) - x(t) = \epsilon\Delta(t_0) &+ \int_{t_0}^{t} \{\hat{f}(z(s,\epsilon), z(\omega(s),\epsilon), s) \\ &- \hat{f}(x(s), x(\omega(s)), s)\}\, ds \\ &+ \epsilon \int_{t_0}^{t} Y(z(s,\epsilon), z(\omega(s),\epsilon), s)\, ds \\ &+ \int_{t_0}^{t} Z(z(s,\epsilon), z(\omega(s),\epsilon), s; \epsilon)\, ds. \end{aligned}$$

Hence,

$$\begin{aligned} |\, z(t,\epsilon) - x(t)| \leqq \epsilon\, |\, \Delta(t_0)| &+ \int_{t_0}^{t} m(s)\{|\, z(s,\epsilon) - x(s)| \\ &+ |\, z(\omega(s),\epsilon) - x(\omega(s))|\}\, ds + \epsilon \int_{t_0}^{t} \bar{m}(s)\, ds + \epsilon^2. \end{aligned} \tag{4.8}$$

Let $\rho(\epsilon) = \sup\{|\, z(s,\epsilon) - x(s)\,| : \omega_0 \leqq s \leqq \tau_\epsilon\}$. Then (4.8) becomes

$$\begin{aligned} |\, z(t,\epsilon) - x(t)| &\leqq \epsilon\, |\, \Delta(t_0)| + \rho(\epsilon) \int_{t_0}^{t} 2m(s)\, ds + \epsilon \int_{t_0}^{t} \bar{m}(s)\, ds + \epsilon^2 \\ &\leqq \epsilon \,\|\, \Delta \,\| + \rho(\epsilon) 2\tilde{M}(t) + \epsilon\tilde{M}(t) + \epsilon^2, \end{aligned}$$

where $\|\Delta\| = \sup\{|\Delta(s)| : \omega_0 \leqq s \leqq t_0\}$ and $\tilde{M}(t) \equiv 0$ for $t < t_0$, $\tilde{M}(t) \equiv \int_{t_0}^{t} \tilde{m}(s)\, ds$ for $t \geqq t_0$, with $\tilde{m} \equiv m + \bar{m}$.

Since $z(t, \epsilon) = x(t) + \epsilon\Delta(t)$ for $t \in [\omega_0, t_0]$ and since $\tilde{M}$ is nondecreasing, this last estimate also holds for $|\, z(\omega(t), \epsilon) - x(\omega(t))\,|$ whenever $t \in [t_0, \tau_\epsilon]$. Using these estimates in (4.8) gives

$$\begin{aligned} |\, z(t,\epsilon) - x(t)\,| &\leqq \epsilon\, |\, \Delta(t_0)\,| + \epsilon\tilde{M}(t) + \epsilon^2 \\ &\quad + \int_{t_0}^{t} m(s) 2\{\epsilon \,\|\, \Delta \,\| + \rho(\epsilon) 2\tilde{M}(s) + \epsilon 2\tilde{M}(s) + \epsilon^2\}\, ds \\ &\leqq \epsilon \,\|\, \Delta \,\|\{1 + 2\tilde{M}(t)\} + \epsilon\{2\tilde{M}(t) + [2\tilde{M}(t)]^2/2\} \\ &\quad + \epsilon^2\{1 + 2\tilde{M}(t)\} + \rho(\epsilon)[2\tilde{M}(t)]^2/2. \end{aligned}$$

Again, this estimate also holds for $|\, z(\omega(t),\epsilon) - x(\omega(t))\,|$ whenever

$t \in [t_0, \tau_\epsilon]$. Using these again in (4.8) yields

$$|z(t,\epsilon) - x(t)| \leqq \epsilon \|\Delta\| \{1 + 2\tilde{M}(t) + [2\tilde{M}(t)]^2/2\} + \epsilon\{2\tilde{M}(t) + [2\tilde{M}(t)]^2/2 + [2\tilde{M}(t)]^3/3!\} + \epsilon^2\{1 + 2\tilde{M}(t) + [2\tilde{M}(t)]^2/2\} + \rho(\epsilon)[2\tilde{M}(t)]^3/3!.$$

It follows that

$$|z(t,\epsilon) - x(t)| \leqq \epsilon \|\Delta\| e^{2\tilde{M}(t)} + \epsilon e^{2\tilde{M}(t)} + \epsilon^2 e^{2\tilde{M}(t)} + \rho(\epsilon)[2\tilde{M}(t)]^k/k! \tag{4.9}$$

for all integers $k = 1, 2, \cdots$. For each fixed ϵ, we let $k \to \infty$ in (4.9) which gives for $t \in [t_0, \tau_\epsilon]$,

$$|z(t,\epsilon) - x(t)| \leqq \{\epsilon(1 + \|\Delta\|) + \epsilon^2\} e^{2B} \tag{4.10}$$

whenever $\tau_\epsilon \leqq t_1 + \epsilon_1 |\delta t|$, where $B \equiv \tilde{M}(t_1 + \epsilon_1 |\delta t|)$.

Now define $C = \{x(t) : \omega_0 \leqq t \leqq t_1 + \epsilon_1 |\delta t|\}$ and let $d =$ distance $(C, \partial X)$. Then $d > 0$. Choose $\epsilon_2 \leqq \epsilon_1$ such that $0 < \epsilon < \epsilon_2$ implies $\{\epsilon(1 + \|\Delta\|) + \epsilon^2\} e^{2B} < d/2$. Note that ϵ_2 is uniform in Δ and δt as long as Δ, δt range over bounded sets.

It follows that for $0 \leqq \epsilon < \epsilon_2$, $z(t,\epsilon)$ exists on $[t_0, t_1 + \epsilon_2 |\delta t|]$ where it is in X°. Suppose not. Then there exists ϵ^*, $0 < \epsilon^* < \epsilon_2$, such that $z(t, \epsilon^*)$ exists on $[t_0, \tau_{\epsilon^*}]$ with $z(\tau_{\epsilon^*}, \epsilon^*) \in \partial X$ and $\tau_{\epsilon^*} \leqq t_1 + \epsilon_2 |\delta t|$. From (4.10) we have $|z(t, \epsilon^*) - x(t)| < d/2$ for $t_0 \leqq t \leqq \tau_{\epsilon^*}$. In particular, $|z(\tau_{\epsilon^*}, \epsilon^*) - x(\tau_{\epsilon^*})| < d/2$. But distance$(x(\tau_{\epsilon^*}), \partial X) \geqq d$ since $\tau_{\epsilon^*} \leqq t_1 + \epsilon_2 |\delta t| \leqq t_1 + \epsilon_1 |\delta t|$. This implies distance$(z(\tau_{\epsilon^*}, \epsilon^*), \partial X) \geqq d/2$ which is a contradiction.

LEMMA 4.3. *For $\epsilon > 0$ sufficiently small the solution $z(t,\epsilon)$ to (4.7) exists on $[t_0, t_1 + \epsilon\delta t]$ and has the form*

$$z(t,\epsilon) = x(t) + \epsilon\delta z(t) + o(\epsilon), \tag{4.11}$$

where δz satisfies the linear variational (differential-difference) equation

$$\begin{aligned} \delta\dot{z}(t) &= \hat{f}_x(x(t), x(\omega(t)), t)\delta z(t) + \hat{f}_y(x(t), x(\omega(t)), t)\delta z(\omega(t)) \\ &\quad + \delta f(x(t), x(\omega(t)), t) \qquad \text{for} \quad t \in [t_0, t_1 + \epsilon\delta t], \\ \delta z(t) &= \delta\phi(t) \qquad \text{for} \quad t \in [\omega_0, t_0]. \end{aligned} \tag{4.12}$$

Proof. Consider first the equation

$$\dot{y}(t,\epsilon) = \hat{f}(y(t,\epsilon), y(\omega(t),\epsilon), t) + \epsilon\delta f(y(t,\epsilon), y(\omega(t),\epsilon), t) \quad \text{for} \quad t > t_0,$$

$$y(t,\epsilon) = \hat{\phi}(t) + \epsilon\delta\phi(t) \quad \text{for} \quad t \in [\omega_0, t_0],$$

for $0 \leqq \epsilon \leqq 1$. From Theorem 2.15 in [4], there exists $\epsilon_0 > 0$ such that for

each ϵ, $0 \leqq \epsilon < \epsilon_0$, the solution $y(t, \epsilon)$ exists and is in X° on $[t_0, \tau_\epsilon]$. Applying Lemma 4.2 with $Y = \delta f$, $Z = 0$, and $\Delta = \delta\phi$, we obtain that there exists $\epsilon_2 > 0$ such that for $0 \leqq \epsilon < \epsilon_2$, $y(t, \epsilon)$ exists on $[t_0, t_1 + \epsilon_2 \mid \delta t \mid]$ and is in X°. Put $W(x, y, t; \gamma) = \hat{f}(x, y, t) + \gamma\delta f(x, y, t)$, where x and y are n-vectors, and t and γ are scalars. Then if $\chi(t, t_0, \phi, \gamma)$ is the solution to

$$\begin{aligned} \dot{\chi}(t) &= W(\chi(t), \chi(\omega(t)), t; \gamma) \quad \text{for} \quad t > t_0, \\ \chi(t) &= \phi(t) \qquad\qquad\qquad\qquad \text{for} \quad t \in [\omega_0, t_0], \end{aligned}$$

we have $y(t, \epsilon) = \chi(t, t_0, \hat{\phi} + \epsilon\delta\phi, \epsilon)$ and $x(t) = \chi(t, t_0, \hat{\phi}, 0)$. Note that W is C^∞ in γ and has the same smoothness w.r.t. (x, y, t) as the members of F.

Consider next

$$\begin{aligned} y(t, \epsilon) - x(t) &= \chi(t, t_0, \hat{\phi} + \epsilon\delta\phi, \epsilon) - \chi(t, t_0, \hat{\phi}, 0) \\ &= \{\chi(t, t_0, \hat{\phi} + \epsilon\delta\phi, \epsilon) - \chi(t, t_0, \hat{\phi}, \epsilon)\} \\ &\quad + \{\chi(t, t_0, \hat{\phi}, \epsilon) - \chi(t, t_0, \hat{\phi}, 0)\}. \end{aligned} \tag{4.13}$$

From extensions of well-known theorems (see [7, Chap. 9]) it follows that χ is Gâteaux differentiable w.r.t. ϕ and $d\chi[t, t_0, \hat{\phi} + s\delta\phi, \gamma; \delta\phi]$ is continuous in t, s, γ, where $d\chi$ is the Gâteaux derivative of χ w.r.t. ϕ in the direction of $\delta\phi$. Furthermore, as a function of t, $z(t) = d\chi[t, t_0, \hat{\phi}, \gamma; \delta\phi]$ satisfies

$$\begin{aligned} \dot{z}(t) &= DW[\chi(t, t_0, \hat{\phi}, \gamma), \chi(\omega(t), t_0, \hat{\phi}, \gamma), t; \gamma : z] \\ & \qquad\qquad\qquad\qquad \text{for} \quad t > t_0, \\ z(t) &= \delta\phi(t) \qquad\qquad\qquad\qquad \text{for} \quad t \in [\omega_0, t_0], \end{aligned} \tag{4.14}$$

where DW is the Fréchet differential of W w.r.t. x (considering $W(x(t), x(\omega(t)), t; \gamma)$ as a functional in x on $C[\omega_0, t_1 + \epsilon_2 \mid \delta t \mid]$).

Hence, considering the first term in (4.13) we obtain

$$\begin{aligned} &\chi(t, t_0, \hat{\phi} + \epsilon\delta\phi, \epsilon) - \chi(t, t_0, \hat{\phi}, \epsilon) \\ &\qquad = \epsilon\{d\chi[t, t_0, \hat{\phi}, \epsilon; \delta\phi] + o(1)\} = \epsilon d\chi[t, t_0, \hat{\phi}, \epsilon; \delta\phi] + o(\epsilon). \end{aligned}$$

From the continuity of $d\chi$ it follows that the term $o(\epsilon)$ is uniform in t on $[t_0, t_1 + \epsilon_2 \mid \delta t \mid]$. That $d\chi$ is homogeneous of degree one in $\delta\phi$ follows from the definition of Gâteaux derivative. In fact, examining (4.14) one sees that $d\chi[t, t_0, \hat{\phi}, \gamma; \delta\phi]$ is actually linear in $\delta\phi$. Thus it follows that if $\delta\phi$ were not fixed but allowed to range over $[\delta\phi_1, \cdots, \delta\phi_l] \subset D(M)$, the convex hull of a finite collection of elements in $D(M)$, then the term $o(\epsilon)$ above would be uniform in $\delta\phi$ in this set.

Now

$$\begin{aligned} \epsilon d\chi[t, t_0, \hat{\phi}, \epsilon; \delta\phi] &= \epsilon d\chi[t, t_0, \hat{\phi}, 0; \delta\phi] + \epsilon\{d\chi[t, t_0, \hat{\phi}, \epsilon; \delta\phi] \\ &\qquad - d\chi[t, t_0, \hat{\phi}, 0; \delta\phi]\}, \end{aligned}$$

where the second term is again uniformly $o(\epsilon)$ for t in $[t_0, t_1 + \epsilon_2 \mid \delta t \mid]$ and $\delta\phi \in [\delta\phi_1, \cdots, \delta\phi_l]$.

Thus we obtain

$$\chi(t, t_0, \hat{\phi} + \epsilon\delta\phi, \epsilon) - \chi(t, t_0, \hat{\phi}, \epsilon) = \epsilon d\chi[t, t_0, \hat{\phi}, 0; \delta\phi] + o(\epsilon), \tag{4.15}$$

where as a function of t, $\delta x(t) = d\chi[t, t_0, \hat{\phi}, 0; \delta\phi]$ satisfies

$$\begin{aligned} \delta\dot{x}(t) &= DW[x(t), x(\omega(t)), t; 0{:}\,\delta x] \quad \text{for} \quad t \in [t_0, t_1 + \epsilon_2 \mid \delta t \mid], \\ \delta x(t) &= \delta\phi(t) \qquad\qquad\qquad\qquad\quad \text{for} \quad t \in [\omega_0, t_0]. \end{aligned} \tag{4.16}$$

Returning to (4.13) we next consider the term $\chi(t, t_0, \hat{\phi}, \epsilon) - \chi(t, t_0, \hat{\phi}, 0)$. Again from extensions of well-known theorems (see [7]), we get that $\partial\chi(t, t_0, \hat{\phi}, \epsilon)/\partial\gamma$ exists (in fact, is continuous in t and ϵ) and satisfies (as a function of t) at $\epsilon = 0$,

$$\begin{aligned} \delta\dot{x}(t) &= DW[x(t), x(\omega(t)), t; 0{:}\,\delta x] + \frac{\partial W}{\partial\gamma}(x(t), x(\omega(t)), t; 0) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{for} \quad t \in [t_0, t_1 + \epsilon_2 \mid \delta t \mid] \\ \delta x(t) &= 0 \qquad\qquad\qquad\qquad\qquad\quad \text{for} \quad t \in [\omega_0, t_0]. \end{aligned} \tag{4.17}$$

Thus we have

$$\chi(t, t_0, \hat{\phi}, \epsilon) - \chi(t, t_0, \hat{\phi}, 0) = \epsilon\frac{\partial\chi}{\partial\gamma}(t, t_0, \hat{\phi}, 0) + o(\epsilon), \tag{4.18}$$

where $o(\epsilon)$ is uniform in t on $[t_0, t_1 + \epsilon_2 \mid \delta t \mid]$.

In fact, if one considers the previous discussions with δf not fixed, but ranging over the convex hull of a finite collection in $[F] - \hat{f}$, i.e., $\delta f \in [\delta f_1, \cdots, \delta f_m]$, then one can show that the terms $o(\epsilon)$ are uniform w.r.t. δf in this set.

Combining (4.13), (4.15) and (4.18) we have

$$\begin{aligned} y(t, \epsilon) - x(t) &= \epsilon\left\{d\chi[t, t_0, \hat{\phi}, 0; \delta\phi] + \frac{\partial\chi}{\partial\gamma}(t, t_0, \phi, 0)\right\} + o(\epsilon) \\ &= \epsilon\delta z(t) + o(\epsilon), \end{aligned}$$

where $\delta z(t)$ satisfies

$$\begin{aligned} \delta\dot{z}(t) &= DW[x(t), x(\omega(t)), t; 0{:}\,\delta z] + \frac{\partial W}{\partial\gamma}(x(t), x(\omega(t)), t; 0) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{for} \quad t \in [t_0, t_1 + \epsilon_2 \mid \delta t \mid], \\ \delta z(t) &= \delta\phi(t) \qquad\qquad\qquad\qquad\quad \text{for} \quad t \in [\omega_0, t_0]. \end{aligned}$$

It is easy to show that

$$\begin{aligned} &DW[x(t), x(\omega(t)), t; 0{:}\,\eta] \\ &\qquad = \hat{f}_x(x(t), x(\omega(t)), t)\eta(t) + \hat{f}_y(x(t), x(\omega(t)), t)\eta(\omega(t)) \end{aligned}$$

and

$$\frac{\partial W}{\partial \gamma}(x(t), x(\omega(t)), t; 0) = \delta f(x(t), x(\omega(t)), t).$$

Thus we have that there exists $\epsilon_2 > 0$ such that for each ϵ, $0 \leqq \epsilon < \epsilon_2$, $y(t, \epsilon)$ exists on $[t_0, t_1 + \epsilon_2 \mid \delta t \mid]$, is in X°, and has the form (4.11) where δz satisfies (4.12) and $o(\epsilon)$ is uniform in t on $[t_0, t_1 + \epsilon_2 \mid \delta t \mid]$ and in $\delta\phi$, δf as described above.

For $0 \leqq \epsilon \leqq 1$, we consider next the equation

$$\dot{z}(t, \epsilon) = \hat{f}(z(t, \epsilon), z(\omega(t), \epsilon), t) + \epsilon\delta f(z(t, \epsilon), z(\omega(t), \epsilon), t) + g_\epsilon(z(t, \epsilon), z(\omega(t), \epsilon), t) \quad \text{for} \quad t > t_0,$$

$$z(t, \epsilon) = \hat{\phi}(t) + \epsilon\delta\phi(t) \quad \text{for} \quad t \in [\omega_0, t_0].$$

Again, by Theorem 2.15 in [4], there exists $\bar{\epsilon}_0 > 0$ such that for each ϵ, $0 \leqq \epsilon < \bar{\epsilon}_0$, the solution $z(t, \epsilon)$ exists and is in X° on $[t_0, \tau_\epsilon]$. Applying Lemma 4.2 with $Y = \delta f$, $Z = g_\epsilon$, and $\Delta = \delta\phi$, we get that there exists $\epsilon_3 > 0$ such that for each ϵ, $0 \leqq \epsilon < \epsilon_3$, $z(t, \epsilon)$ exists on $[t_0, t_1 + \epsilon_3 \mid \delta t \mid]$ and is in X°. We take $\epsilon_3 \leqq \epsilon_2$. Next, for any $t \in [t_0, t_1 + \epsilon_3 \mid \delta t \mid]$, consider, for $0 \leqq \epsilon < \epsilon_3$,

$$z(t, \epsilon) - y(t, \epsilon) = \int_{t_0}^{t} \{\hat{f}(z(s, \epsilon), z(\omega(s), \epsilon), s) - \hat{f}(y(s, \epsilon), y(\omega(s), \epsilon), s)\}\, ds$$

$$+ \epsilon \int_{t_0}^{t} \{\delta f(z(s, \epsilon), z(\omega(s), \epsilon), s) - \delta f(y(s, \epsilon), y(\omega(s), \epsilon), s)\}\, ds$$

$$+ \int_{t_0}^{t} g_\epsilon(z(s, \epsilon), z(\omega(s), \epsilon), s)\, ds.$$

The last term above is $o(\epsilon)$ uniformly in t and in $\delta\phi$, δf, δt as long as δt is in some bounded set and $\delta f \in [\delta f_1, \cdots, \delta f_m]$ (see the discussion following this proof). Hence we have

$$|z(t,\epsilon) - y(t,\epsilon)| \leqq \int_{t_0}^{t} \{m(s)|z(s,\epsilon) - y(s,\epsilon)| + m(s)|z(\omega(s),\epsilon) - y(\omega(s),\epsilon)|\}\, ds$$

$$+ 2\epsilon \int_{t_0}^{t} \{m(s)|z(s,\epsilon) - y(s,\epsilon)| + m(s)|z(\omega(s),\epsilon) - y(\omega(s),\epsilon)|\}\, ds + o(\epsilon).$$

Thus, for $t \in [t_0, t_1 + \epsilon_3 \mid \delta t \mid]$, we have

$$|z(t, \epsilon) - y(t, \epsilon)| \leqq (1 + 2\epsilon) \int_{t_0}^{t} m(s)\{|z(s, \epsilon) - y(s, \epsilon)| + |z(\omega(s), \epsilon) - y(\omega(s), \epsilon)|\}\, ds + o(\epsilon). \tag{4.19}$$

Recall that $m(t)$ depends on $f_1, f_2, \cdots, f_k, \hat{f}$, where $\delta f = \sum_1^k \alpha_i f_i - \hat{f}$. Hence if $\delta f \in [\delta f_1, \cdots, \delta f_m]$ for $\delta f_1, \cdots, \delta f_m$ fixed, then $m(t)$ will be independent of the particular δf in this set. Also recall that $z(t, \epsilon) = y(t, \epsilon)$ on $[\omega_0, t_0]$. Put $\rho(\epsilon) = \sup \{|z(s, \epsilon) - y(s, \epsilon)| : t_0 \leqq s \leqq t_1 + \epsilon_3 |\delta_t|\}$ and $\mu(t) = \int_{t_0}^{t} m(s)\, ds$ for $t > t_0$, $\mu(t) = 0$ for $t \leqq t_0$. Then (4.19) becomes

$$|z(t, \epsilon) - y(t, \epsilon)| \leqq (1 + 2\epsilon)\rho(\epsilon)2\mu(t) + o(\epsilon)$$

for $t \in [t_0, t_1 + \epsilon_3 |\delta t|]$.

Since μ is nondecreasing and $z = y$ on $[\omega_0, t_0]$, the above estimate also holds for $|z(\omega(t), \epsilon) - y(\omega(t), \epsilon)|$ whenever $t \in [t_0, t_1 + \epsilon_3 |\delta t|]$. Substitution of these estimates in (4.19) yields

$$|z(t,\epsilon) - y(t,\epsilon)| \leqq (1 + 2\epsilon) \int_{t_0}^{t} 2m(s)\{(1 + 2\epsilon)\rho(\epsilon)2\mu(s) + o(\epsilon)\}\, ds$$

$$= (1 + 2\epsilon)^2(2)^2\rho(\epsilon)[\mu(t)]^2/2 + o(\epsilon).$$

Again this estimate holds for $|z(\omega(t), \epsilon) - y(\omega(t), \epsilon)|$ and the term $o(\epsilon)$ is uniform as described previously. Repeated application of these arguments gives

$$|z(t, \epsilon) - y(t, \epsilon)| \leqq \rho(\epsilon)\{2(1 + 2\epsilon)\}^k[\mu(t)]^k/k! + o(\epsilon)$$

for $k = 1, 2, \cdots$, and all t in $[t_0, t_1 + \epsilon_3 |\delta t|]$. Since $\mu(t) \leqq \mu(t_1 + \epsilon_3 |\delta t|) \leqq D$, where D is a constant, for δt in a bounded set, it follows that $z(t, \epsilon) - y(t, \epsilon) = o(\epsilon)$ for $t \in [t_0, t_1 + \epsilon_3 |\delta t|]$, where the $o(\epsilon)$ is uniform in t, $\delta\phi$, and δf as described above.

Combining this with the previous results for $y(t, \epsilon)$ gives the desired results and completes the proof of Lemma 4.3.

Remark. If δt is allowed to vary over some bounded subset of R^1, then the $o(\epsilon)$ term is also uniform in δt.

Suppose that $\delta f_1, \cdots, \delta f_m$ are fixed elements of $[F] - \hat{f}$. Then there exist $f_j \in F, j = 1, 2, \cdots, k$, and vectors $\alpha^i = (\alpha_{i1}, \cdots, \alpha_{ik}) \in P^k$, $i = 1, 2, \cdots, m$, such that $\delta f_i = \sum_{j=1}^k \alpha_{ij} f_j - \hat{f}$. For $\beta = (\beta_1, \cdots, \beta_m)$ in P^m, let

$$\delta f^\beta = \sum_{i=1}^m \beta_i \delta f_i = \sum_{j=1}^k \left(\sum_{i=1}^m \beta_i \alpha_{ij}\right) f_j - \hat{f}.$$

By the quasi-convexity of F and Lemma 4.1 we have:

For every fixed ϵ, $0 \leqq \epsilon \leqq 1$, and $\beta \in P^m$, there exists $g_\epsilon(x, y, t; \beta)$ defined on $G \times G \times I'$, C^1 in x, y, and measurable in t, such that:

1. $\hat{f} + \epsilon\delta f^\beta + g_\epsilon \in F$.

2. $|g_\epsilon(x, y, t; \beta)| < \bar{m}(t)$,

$$\left|\frac{\partial g_\epsilon}{\partial x}(x, y, t; \beta)\right| < \bar{m}(t),$$

$$\left|\frac{\partial g_\epsilon}{\partial y}(x, y, t; \beta)\right| < \bar{m}(t)$$

for every $t \in I'$ and all (x, y) in $X \times X$, where $\bar{m} \in L_1(I')$ and depends on $X, f_1, \cdots, f_k, \hat{f}$.

3. $\left|\int_{\tau_1}^{\tau_2} g_\epsilon(z(t), z(\omega(t)), t; \beta)\, dt\right| < \epsilon^2$ for every solution $z(t)$ of $\dot{z} = \hat{f} + \epsilon\delta f + g_\epsilon$ sufficiently near $x(t)$ and every interval $[\tau_1, \tau_2] \subset I'$ on which $z(t), z(\omega(t))$ are defined.
4. If $\{\beta^j\}$ is a sequence in P^m such that $\beta^j \to \bar{\beta} \in P^m$ as $j \to \infty$, then $\{g_\epsilon(x, y, t; \beta^j)\}$ converges in measure on I' to $g_\epsilon(x, y, t; \bar{\beta})$ for every fixed $(x, y) \in X \times X$ and ϵ, $0 \leqq \epsilon \leqq 1$.

For any given functions $\delta f_1, \cdots, \delta f_m$ in $[F] - \hat{f}$, we assume that in choosing the g_ϵ (for every $\delta f \in [F] - \hat{f}$ and ϵ in $[0, 1]$) such that (4.4), (4.5) and (4.6) hold, we choose the function $g_\epsilon(x, y, t; \beta)$ as described above whenever δf is in $[\delta f_1, \cdots, \delta f_m]$.

Now consider again the proofs of Lemmas 4.2 and 4.3 for $(\delta t, \delta\phi, \delta f)$ in $E = B \times [\delta\phi_1, \cdots, \delta\phi_l] \times [\delta f_1, \cdots, \delta f_m]$, where B is a bounded subset of R^1. Studying Theorem 2.15 in [4] and Lemma 4.3, one sees that the solution to (4.7) exists and is in X° on $[t_0, \tau_\epsilon]$ for $0 \leqq \epsilon < \bar{\epsilon}_0$, where $\bar{\epsilon}_0$ and τ_ϵ depend on E, but not on the individual members $(\delta t, \delta\phi, \delta f)$ in E. Furthermore, in the proof of Lemma 4.2, the function $\tilde{M}(t)$ depends on E and not $(\delta t, \delta\phi, \delta f) \in E$. Hence the proofs of Lemmas 4.2 and 4.3 can be carried out uniformly in $(\delta t, \delta\phi, \delta f) \in E$. From this it follows that the term $o(\epsilon)$ in (4.11) is uniform in t and in $(\delta t, \delta\phi, \delta f)$ on E, when E is as described above.

For t in I', let $\Lambda(s, t)$ be a matrix solution to

$$\Lambda(s, t) = 0 \quad \text{for} \quad s > t,$$

$$\Lambda(t, t) = I,$$

$$\frac{\partial \Lambda}{\partial s}(s, t) + \Lambda(s, t)\hat{f}_x(x(s), x(\omega(s)), s) = 0 \quad \text{for} \quad s \in [\omega(t), t], \tag{4.20}$$

$$\frac{\partial \Lambda}{\partial s}(s, t) + \Lambda(s, t)\hat{f}_x(x(s), x(\omega(s)), s)$$

$$+ \Lambda(r(s), t)\hat{f}_y(x(r(s)), x(s), r(s))\dot{r}(s) = 0 \quad \text{for} \quad s \in [t_0, \omega(t)].$$

Note that it follows from Lemma 1 and the hypotheses on ω and $r = \omega^{-1}$ that the coefficients in (4.20) are measurable. Existence of a solution to

(4.20), which is an advanced differential-difference system, is guaranteed by Corollary 2.2.

Applying Theorem 3, we get that the solution $\delta z(t)$ of (4.12) has the form

$$\begin{aligned}\delta z(t) &= \int_{\omega_0}^{t_0} \Lambda(r(s), t)\hat{f}_y(x(r(s)), x(s), r(s))\delta\phi(s)\dot{r}(s)\, ds \\ &\quad + \Lambda(t_0, t)\delta\phi(t_0) + \int_{t_0}^{t} \Lambda(s, t)\delta f(x(s), x(\omega(s)), s)\, ds\end{aligned} \tag{4.21}$$

for $t > t_0$. From (4.11), which holds on $[t_0, t_1 + \epsilon\delta t]$, we obtain

$$z(t_1 + \epsilon\delta t, \epsilon) = x(t_1 + \epsilon\delta t) + \epsilon\delta z(t_1 + \epsilon\delta t) + o(\epsilon). \tag{4.22}$$

Now

$$\begin{aligned}x(t_1 + \epsilon\delta t) &= x(t_1) + \int_{t_1}^{t_1+\epsilon\delta t} \hat{f}(x(t), x(\omega(t)), t)\, dt \\ &= x(t_1) + \hat{f}(x(t_1), x(\omega(t_1)), t_1)\epsilon\delta t + o(\epsilon) \\ &= x(t_1) + \hat{f}_1\epsilon\delta t + o(\epsilon)\end{aligned}$$

since t_1 is a regular point for $\hat{f}(x(t), x(\omega(t)), t)$. Note that the $o(\epsilon)$ term is independent of t, $\delta\phi$, δf and is uniform in δt in a bounded subset of R^1. From (4.12) it follows from standard theorems on dependence of solutions on initial data and parameters that δz is uniformly continuous in $\delta\phi$ and δf whenever $(\delta\phi, \delta f) \in [\delta\phi_1, \cdots, \delta\phi_l] \times [\delta f_1, \cdots, \delta f_m]$. Hence $\delta z(t_1 + \epsilon\delta t) - \delta z(t_1) \to 0$ as $\epsilon \to 0$ uniformly for $(\delta t, \delta\phi, \delta f)$ in $B \times [\delta\phi_1, \cdots, \delta\phi_l] \times [\delta f_1, \cdots, \delta f_m]$, where B is a bounded subset of R^1. Therefore,

$$\epsilon\delta z(t_1 + \epsilon\delta t) = \epsilon\delta z(t_1) + \epsilon\{\delta z(t_1 + \epsilon\delta t) - \delta z(t_1)\} = \epsilon\delta z(t_1) + o(\epsilon),$$

where the term $o(\epsilon)$ is again uniform in $(\delta t, \delta\phi, \delta f)$ as described above.

Hence, (4.22) becomes

$$z(t_1 + \epsilon\delta t, \epsilon) = x(t_1) + \epsilon\{\delta z(t_1) + \delta t\hat{f}_1\} + o(\epsilon). \tag{4.23}$$

Since the right side of the first equation in (4.7) is in F and since $\hat{\phi} + \epsilon\delta\phi \in \Phi$ for ϵ sufficiently small, we have that $q_z = (z(t_0, \epsilon), z(t_1 + \epsilon\delta t, \epsilon), t_1 + \epsilon\delta t)$ is in Q. Furthermore,

$$\begin{aligned}q_z &= (\hat{\phi}(t_0) + \epsilon\delta\phi(t_0), x(t_1) + \epsilon\{\delta z_1 + \delta t\hat{f}_1\} + o(\epsilon), t_1 + \epsilon\delta t) \\ &= (x(t_0), x(t_1), t_1) + \epsilon(\delta\phi(t_0), \delta z_1 + \delta t\hat{f}_1, \delta t) + o(\epsilon);\end{aligned}$$

thus we see that

$$q_z = q_x + \epsilon(\delta z(t_0), \delta z(t_1) + \delta t\hat{f}_1, \delta t) + o(\epsilon), \tag{4.24}$$

where $\delta z(t_0) = \delta\phi(t_0)$, δz is given by (4.21), and $\hat{f}_1 = \hat{f}(x(t_1), x(\omega(t_1)), t_1)$.

Now consider the convex set $A \equiv R^1 \times D(M) \times ([F] - \hat{f})$, i.e.,

$$A = \{(\delta t, \delta\phi, \delta f) : \delta t \in R^1, \delta\phi \in D(M), \delta f \in [F] - \hat{f}\}.$$

If $\{\delta\phi_1, \cdots, \delta\phi_l\}$ and $\{\delta f_1, \cdots, \delta f_m\}$ are finite subsets of $D(M)$ and $[F] - \hat{f}$, then we define a subset of A as follows:

$$\begin{aligned} A(\{\delta\phi_i\}_1^l, \{\delta f_j\}_1^m) &\equiv R^1 \times [\delta\phi_1, \cdots, \delta\phi_l] \times [\delta f_1, \cdots, \delta f_m] \\ &= \{(\delta t, \delta\phi, \delta f) : \delta t \in R^1, \delta\phi \in [\delta\phi_1, \cdots, \delta\phi_l], \\ &\qquad \delta f \in [\delta f_1, \cdots, \delta f_m]\}. \end{aligned}$$

Then $A(\{\delta\phi_i\}_1^l, \{\delta f_j\}_1^m)$ is a convex subset of A which can be identified with $R^1 \times P^l \times P^m$. Thus, in the future when referring to a topology on $A(\{\delta\phi_i\}, \{\delta f_j\})$, we shall mean the topology induced by the Euclidean topology in R^{1+l+m} under the abovementioned identification.

From previous remarks, we have that given $(\delta t, \delta\phi, \delta f)$ in A, there exists $\epsilon_0 > 0$ such that the solution to (4.7) exists for $0 \leqq \epsilon \leqq \epsilon_0$ and is such that (4.24) holds. This gives us a mapping from A to R^{2n+1} depending on ϵ defined by

$$h_\epsilon(\delta t, \delta\phi, \delta f) = (q_z - q_x)/\epsilon = L(\delta t, \delta\phi, \delta f) + \Delta(\delta t, \delta\phi, \delta f, \epsilon), \tag{4.25}$$

where $L(\delta t, \delta\phi, \delta f) = (\delta z_0, \delta z_1 + \delta t\hat{f}_1, \delta t)$ and $\Delta \to 0$ as $\epsilon \to 0$. Note that L is linear and independent of ϵ.

LEMMA 4.4. *For every compact $C \subset A(\{\delta\phi_i\}_1^l, \{\delta f_j\}_1^m)$ there exists an $\epsilon_0 > 0$ such that for each fixed ϵ, $0 \leqq \epsilon < \epsilon_0$, the mapping h_ϵ given in* (4.25) *is defined and continuous on C.*

Proof. Given $C \subset A(\{\delta\phi_i\}, \{\delta f_j\})$, C compact, there exists $\epsilon_0 > 0$ such that for each fixed ϵ, $0 \leqq \epsilon < \epsilon_0$, the solution $z(t, \epsilon)$ of

$$\begin{aligned} \dot{z} &= \hat{f} + \epsilon\delta f + g_\epsilon \quad \text{for} \quad t \in [t_0, t_1 + \epsilon\delta t], \\ z &= \hat{\phi} + \epsilon\delta\phi \qquad \text{for} \quad t \in [\omega_0, t_0], \end{aligned} \tag{4.26}$$

exists and is in X° for each $(\delta t, \delta\phi, \delta f)$ in C. (This follows from Lemmas 4.2 and 4.3 and the remarks following these lemmas.)

So let ϵ be fixed, $0 < \epsilon < \epsilon_0$. Let $z_1(t, \epsilon) = z_1(t)$ and $z_2(t, \epsilon) = z_2(t)$ be solutions of (4.26) corresponding to $(\delta t^1, \delta\phi^1, \delta f^1)$ and $(\delta t^2, \delta\phi^2, \delta f^2)$ in C. Then from (4.24) we have

$$q_{z_i} = q_x + \epsilon(\delta\phi^i(t_0), \delta z^i(t_1) + \delta t^i\hat{f}_1, \delta t^i) + o(\epsilon)$$

for $i = 1, 2$. Since $|h_\epsilon(\delta t^1, \delta\phi^1, \delta f^1) - h_\epsilon(\delta t^2, \delta\phi^2, \delta f^2)| = |q_{z_1} - q_{z_2}|/\epsilon$, we may make the following observation: From the way in which the solutions depend on δt^1, δt^2, to show h_ϵ continuous on C it is sufficient to show that solutions of (4.26) corresponding to $(\delta t, \delta\phi^1, \delta f^1)$ and $(\delta t, \delta\phi^2, \delta f^2)$, with δt fixed, can be made arbitrarily close by taking $(\delta t, \delta\phi^1, \delta f^1)$ and

$(\delta t, \delta\phi^2, \delta f^2)$ sufficiently close in C. That is, if

$$\delta\phi^\eta = \sum_{i=1}^{l} \eta_i \delta\phi_i, \quad \delta\phi^\zeta = \sum_{i=1}^{l} \zeta_i \delta\phi_i, \qquad \eta, \zeta \in P^l,$$
$$\delta f^\gamma = \sum_{j=1}^{m} \gamma_j \delta f_j, \quad \delta f^\beta = \sum_{j=1}^{m} \beta_j \delta f_j, \qquad \gamma, \beta \in P^m,$$

where $\delta f_j = \sum_{n=1}^{k} \alpha_{jn} f_n - \hat{f}$, $\alpha_j = (\alpha_{j1}, \cdots, \alpha_{jk}) \in P^k$, $j = 1, 2, \cdots, m$, then it is sufficient to show that for fixed ϵ and fixed δt,

$$z(t, \epsilon; \delta t, \delta\phi^\eta, \delta f^\gamma) \to z(t, \epsilon; \delta t, \delta\phi^\zeta, \delta f^\beta)$$

uniformly on $[t_0, t_1 + \epsilon\delta t]$ as $\eta \to \zeta$, $\gamma \to \beta$, for $(\delta t, \delta\phi^\eta, \delta f^\gamma)$ and $(\delta t, \delta\phi^\zeta, \delta f^\beta)$ in C.

To show this, let $z(t) = z(t, \epsilon)$ and $y(t) = y(t, \epsilon)$ be solutions corresponding to $(\delta t, \delta\phi^\zeta, \delta f^\beta)$ and $(\delta t, \delta\phi^\eta, \delta f^\gamma)$, where (ζ, β) and (η, γ) are in $\bar{C}$, some compact subset of $P^l \times P^m$. Then $z(t)$, $\omega_0 \leqq t \leqq t_1 + \epsilon\delta t$, and $y(t)$, $\omega_0 \leqq t \leqq t_1 + \epsilon\delta t$, are in X°. For $t_0 \leqq \tau \leqq t_1 + \epsilon\delta t$ we have

$$\begin{aligned} z(\tau) - y(\tau) &= \epsilon\{\delta\phi^\zeta(t_0) - \delta\phi^\eta(t_0)\} \\ &\quad + \int_{t_0}^{\tau} \{\hat{f}(z(t), z(\omega(t)), t) - \hat{f}(y(t), y(\omega(t)), t)\}\, dt \\ &\quad + \epsilon \int_{t_0}^{\tau} \{\delta f^\beta(z(t), z(\omega(t)), t) - \delta f^\gamma(y(t), y(\omega(t)), t)\}\, dt \\ &\quad + \int_{t_0}^{\tau} \{g_\epsilon(z(t), z(\omega(t)), t; \beta) - g_\epsilon(y(t), y(\omega(t)), t; \gamma)\}\, dt \\ &= \epsilon\left\{\sum_{i=1}^{l} (\zeta_i - \eta_i)\delta\phi_i(t_0)\right\} + I_1(\tau) + I_2(\tau) + I_3(\tau). \end{aligned} \tag{4.27}$$

Now

$$\begin{aligned} |I_1(\tau)| &= \left| \int_{t_0}^{\tau} \{\hat{f}(z(t), z(\omega(t)), t) - \hat{f}(y(t), y(\omega(t)), t)\}\, dt \right| \\ &\leqq \int_{t_0}^{\tau} m(t)\{|z(t) - y(t)| + |z(\omega(t)) - y(\omega(t))|\}\, dt, \end{aligned}$$

where m is an L_1 function depending on X, $\hat{f}$, and $f_1, \cdots, f_k$. Furthermore, it is not hard to show that there exist constants K_1 and K_2 such that

$$\begin{aligned} |I_2(\tau)| &\leqq K_1 |\beta - \gamma| \\ &\quad + K_2 \int_{t_0}^{\tau} m(t)\{|z(t) - y(t)| + |z(\omega(t)) - y(\omega(t))|\}\, dt. \end{aligned}$$

Next consider

$$|I_3(\tau)| = \left| \int_{t_0}^{\tau} \{g_\epsilon(z(t), z(\omega(t)), t; \beta) - g_\epsilon(y(t), y(\omega(t)), t; \gamma)\}\, dt \right|.$$

For fixed (x, y) in $X \times X$ and ϵ in $[0, 1]$, we have $g_\epsilon(x, y, t; \beta^i) \to g_\epsilon(x, y, t; \bar{\beta})$ in measure as a function of t on I' if $\beta^i \to \bar{\beta}$ in P^m. From this and the fact that each g_ϵ is dominated by an integrable function, we obtain

$$\lim_{i\to\infty} \int_{E^*} |\, g_\epsilon(x, y, t; \beta^i) - g_\epsilon(x, y, t; \bar{\beta})|\, dt = 0$$

for $E^* \subset I'$, E^* of finite measure. Thus, for each fixed $(x, y) \in X \times X$ and $0 < \epsilon < 1$ and $\tau \in [t_0, t_1 + \epsilon\delta t]$, we have

$$\lim_{\gamma\to\beta} \int_{t_0}^{\tau} |\, g_\epsilon(x, y, t; \beta) - g_\epsilon(x, y, t; \gamma)|\, dt = 0.$$

In a proof similar to that of Lemma 4.1, one can show that this implies

$$\lim_{\gamma\to\beta} \int_{t_0}^{\tau} |\, g_\epsilon(\xi(t), \xi(\omega(t)), t; \beta) - g_\epsilon(\xi(t), \xi(\omega(t)), t; \gamma)|\, dt = 0$$

for any solution $\xi(t)$ of (4.7) which is sufficiently near $x(t)$.

Hence, in particular, the above holds with $\xi(t)$ replaced by $z(t)$. Thus,

$$\begin{aligned}
|\, I_3(\tau)| &\leqq \int_{t_0}^{\tau} |g_\epsilon(z(t), z(\omega(t)), t; \beta) - g_\epsilon(z(t), z(\omega(t)), t; \gamma)|\, dt \\
&\quad + \int_{t_0}^{\tau} |\, g_\epsilon(z(t), z(\omega(t)), t; \gamma) - g_\epsilon(y(t), y(\omega(t)), t; \gamma)|\, dt \\
&\leqq \int_{t_0}^{t_1+\epsilon\delta t} |\, g_\epsilon(z(t), z(\omega(t)), t; \beta) - g_\epsilon(z(t), z(\omega(t)), t; \gamma)|\, dt \\
&\quad + \int_{t_0}^{\tau} \bar{m}(t)\{|\, z(t) - y(t)| + |\, z(\omega(t)) - y(\omega(t))|\}\, dt,
\end{aligned}$$

where $\bar{m}$ depends on X and $f_1, \cdots, f_k, \hat{f}$. Therefore,

$$|\, I_3(\tau)| \leqq E(\beta, \gamma) + \int_{t_0}^{\tau} \bar{m}(t)\{|\, z(t) - y(t)| + |\, z(\omega(t)) - y(\omega(t))|\}\, dt,$$

where $E(\beta, \gamma) \to 0$ as $\gamma \to \beta$.

Combining these estimates on I_1, I_2 and I_3 yields from (4.27),

$$\begin{aligned}
|\, z(\tau) - y(\tau)| &\leqq \epsilon_0 \sum_{i=1}^{l} |\, \zeta_i - \eta_i\, |\, |\, \delta\phi_i(t_0)| \\
&\quad + \int_{t_0}^{\tau} m(t)\{|\, z(t) - y(t)| + |\, z(\omega(t)) - y(\omega(t))|\}\, dt \\
&\quad + K_1 |\, \beta - \gamma\, | + K_2 \int_{t_0}^{\tau} m(t)\{|\, z(t) - y(t)| \\
&\qquad + |\, z(\omega(t)) - y(\omega(t))|\}\, dt
\end{aligned}$$

$$+ E(\beta, \gamma) + \int_{t_0}^{\tau} \bar{m}(t)\{|z(t) - y(t)| + |z(\omega(t)) - y(\omega(t))|\}\, dt$$

for any $\tau \in [t_0, t_1 + \epsilon\delta t]$.

This may be written

$$|z(\tau) - y(\tau)| \leqq H(\beta, \gamma) + \int_{t_0}^{\tau} \tilde{m}(t)\{|z(t) - y(t)| + |z(\omega(t)) - y(\omega(t))|\}\, dt + K_3 |\zeta - \eta|, \tag{4.28}$$

where $H(\beta, \gamma) \to 0$ as $\gamma \to \beta$ and $\tilde{m}$ is a nonnegative integrable function. Letting $\tilde{\mu}(t) = \int_{t_0}^{t} \tilde{m}(s)ds$ for $t > t_0$, $\tilde{\mu}(t) = 0$ for $t \leqq t_0$, and $\rho = \sup\{|z(t) - y(t)| : t \in [\omega_0, t_1 + \epsilon\delta t]\}$, we have that (4.28) becomes

$$|z(\tau) - y(\tau)| \leqq H(\beta, \gamma) + K_3 |\zeta - \eta| + 2\rho\tilde{\mu}(\tau).$$

This estimate also holds for $|z(\omega(\tau)) - y(\omega(\tau))|$.

Using these in (4.28) and repeating the procedure as in other proofs in this paper, one finds

$$|z(\tau) - y(\tau)| \leqq \{H(\beta, \gamma) + K_3 |\zeta - \eta|\} e^{2\tilde{\mu}(\tau)} + \rho\{2\tilde{\mu}(\tau)\}^{K+1}/(K + 1)!$$

for $K = 1, 2, \cdots$. This estimate holds for any $\tau \in [t_0, t_1 + \epsilon\delta t]$. Since there exists a constant B such that $2\tilde{\mu}(\tau) \leqq B$ on this interval, it follows that $|z(\tau) - y(\tau)| \to 0$ as $\gamma \to \beta$, $\eta \to \zeta$, uniformly on $[t_0, t_1 + \epsilon\delta t]$, which completes the proof of Lemma 4.4.

Note that h_ϵ is not necessarily continuous in ϵ since $z(t, \epsilon)$, the solution of (4.7), is not.

It is easy to see that the mapping L, the linear part of h_ϵ, which is independent of ϵ, is continuous on $A(\{\delta\phi_i\}_1^{\,l}, \{\delta f_j\}_1^{\,m})$. Furthermore, for fixed $(\delta t, \delta\phi, \delta f)$, we had that $\Delta(\delta t, \delta\phi, \delta f, \epsilon) \to 0$ as $\epsilon \to 0$. From previous remarks it follows that $\Delta(\delta t, \delta\phi, \delta f, \epsilon) \to 0$ as $\epsilon \to 0$ uniformly in $(\delta t, \delta\phi, \delta f)$ on any given compact subset of $A(\{\delta\phi_i\}_1^{\,l}, \{\delta f_j\}_1^{\,m})$. Also, for fixed ϵ sufficiently small, Δ is continuous on such subsets.

Now let $K = L(A)$. (K is the set of first order terms from the endpoints of perturbed solutions.) Since A is convex and L is linear, we get that K is a convex subset of R^{2n+1} which contains the origin (corresponding to $(\delta t, \delta\phi, \delta f) = (0, 0, 0)$ in A).

In reference to the manifold N with boundary $\tilde{M}$ given in Theorem 5, we shall denote by N_T and M_T, respectively, the sets $N_T(q_x)$ and $M_T(q_x)$. Then M_T is the edge of N_T. Consider next the subset of R^{2n+1} given by $N_T - q_x = \{p \in R^{2n+1} : p = p^* - q_x$, where $p^* \in N_T\}$, which is a half-plane with its edge through the origin.

Then we have under consideration two convex subsets of R^{2n+1}, namely, K and $N_T - q_x$, which both contain the origin.

LEMMA 4.5. *The sets K and $N_T - q_x$ can be separated by a hyperplane through the origin.*

Proof. Since N is a C^1 manifold, there is a homeomorphism h^* from a neighborhood U_{q_x} of q_x in N_T onto a neighborhood of q_x in N of the form $h^*(y) = y + \Delta^*(y)$, where $\Delta^*(y)/|y - q_x| \to 0$ as $y \to q_x$ in U_{q_x}. Now suppose the lemma false. Then the carrier planes of K and $N_T - q_x$ are in general position (i.e., the union of the carrier planes is not contained in any hyperplane), and there is a point q' which is a relative interior point of both $N_T - q_x$ and K. Let $U_{q'}$ be a neighborhood of q' with compact closure $\bar{U}_{q'}$ in the carrier of $N_T - q_x$ such that $\bar{U}_{q'} \subset (N_T - q_x)^\circ$, the interior of $N_T - q_x$. Let m = dimension K, $0 \leqq m \leqq 2n + 1$. Then there exists a simplex K_m of dimension m such that $q' \in K_m^\circ$ and $K_m \subset K^\circ$. Let $p_1, \cdots, p_m, p_{m+1}$ be the vertices of K_m, i.e., $K_m = [p_1, \cdots, p_{m+1}]$. Since $p_i \in K$, there exist $\gamma_i \in A$, $\gamma_i = (\delta t_i, \delta\phi_i, \delta f_i)$, $i = 1, 2, \cdots, m + 1$, such that $L(\gamma_i) = p_i$. Putting $S_m = [\gamma_1, \cdots, \gamma_{m+1}]$ we have that S_m is an m-simplex, $S_m \subset A(\{\delta\phi_i\}_1^{m+1}, \{\delta f_i\}_1^{m+1})$. S_m is compact and $K_m = L(S_m)$ by the linearity of L. We assume the functions g_ϵ corresponding to $\delta f \in [\delta f_1, \cdots, \delta f_{m+1}]$ are chosen in the manner indicated previously. Then from Lemma 4.4 it follows that h_ϵ, restricted to S_m, is defined and continuous for fixed $\epsilon > 0$, ϵ sufficiently small. Furthermore, for ϵ sufficiently small, $q_x + \epsilon\bar{U}_{q'} \subset U_{q_x}$. Hence, for fixed $\epsilon > 0$ sufficiently small, we can define a continuous mapping π on $S = S_m \times \bar{U}_{q'}$ into R^{2n+1} by

$$\begin{aligned}\pi(s, q; \epsilon) &= h_\epsilon(s) - \frac{1}{\epsilon} h^*(q_x + \epsilon q) + \frac{1}{\epsilon} q_x \\ &= L(s) - q + \Delta(s, \epsilon) - \frac{1}{\epsilon}\Delta^*(q_x + \epsilon q)\end{aligned}$$

for $s \in S_m$, $q \in \bar{U}_{q'}$; or $\pi(\sigma; \epsilon) = \tilde{L}(\sigma) + \tilde{\Delta}(\sigma, \epsilon)$, where $\sigma = (s, q)$, $\tilde{L}(\sigma) = L(s) - q$, and $\tilde{\Delta}(\sigma, \epsilon) = \Delta(s, \epsilon) - \Delta^*(q_x + \epsilon q)/\epsilon$.

From the properties of Δ and Δ^*, it follows that $\tilde{\Delta}(\sigma, \epsilon) \to 0$ as $\epsilon \to 0$ uniformly on S.

Since the carriers of K and $N_T - q_x$ are in general position and since $q' \in L(S_m) \cap U_{q'}$, the image of $S = S_m \times \bar{U}_{q'}$ under $\tilde{L}(\sigma) = L(s) - q$ contains a sphere B_ρ about the origin in R^{2n+1}. That is, $B_\rho \subset \tilde{L}(S)$.

From the linearity of $\tilde{L}$ we have that the kernel of $\tilde{L}$ is a linear manifold and hence it follows that there is a compact convex subset $S^* \subset S$ such that $\tilde{L}$ restricted to S^* is a homeomorphism from S^* onto $\tilde{L}(S^*)$ and such that $\tilde{L}(S^*)$ contains a sphere $B_{\rho'}$ in R^{2n+1}. We denote this restriction by $\tilde{L}^*$.

We then consider $\pi(\sigma, \epsilon)$ on S^* and show that for each fixed ϵ sufficiently small, there exists $\sigma \in S^* \subset S$ such that $\pi(\sigma, \epsilon) = 0$.

On S^*, $\pi(\sigma, \epsilon) = \tilde{L}^*(\sigma) + \tilde{\Delta}(\sigma, \epsilon)$. Choose $\epsilon_1 > 0$ such that $\epsilon < \epsilon_1$ implies $|\tilde{\Delta}(\sigma, \epsilon)| < \rho'$ for $\sigma \in S$. Define for $p \in \tilde{L}(S^*)$ and $\epsilon < \epsilon_1$, ϵ fixed, the continuous mapping H_ϵ of $\tilde{L}(S^*)$ into R^{2n+1} by $H_\epsilon(p) = -\pi(\tilde{L}^{*-1}(p), \epsilon) + p$. Then

$$|H_\epsilon(p)| = |-\pi(\tilde{L}^{*-1}(p), \epsilon) + \tilde{L}^*(\tilde{L}^{*-1}(p))| = |-\tilde{\Delta}(\tilde{L}^{*-1}(p), \epsilon)| < \rho',$$

so that H_ϵ maps the convex compact $\tilde{L}(S^*)$ into $B_{\rho'} \subset \tilde{L}(S^*)$. Then by a well-known theorem (Schauder-Tychonoff), H_ϵ has a fixed point. That is, there exists $\bar{p} \in \tilde{L}(S^*)$ such that $H_\epsilon(\bar{p}) = \bar{p}$ or $-\pi(\tilde{L}^{*-1}(\bar{p}), \epsilon) + \bar{p} = \bar{p}$, which gives $\pi(\bar{\sigma}, \epsilon) = 0$, where $\bar{\sigma} = \tilde{L}^{*-1}(\bar{p}) \in S^*$. Hence, for $\epsilon > 0$ sufficiently small, there exist $s \in S_m$ and $q \in \bar{U}_{q'}$ (depending on ϵ) such that

$$\pi(s, q; \epsilon) = 0$$

or

$$L(s) + \Delta(s, \epsilon) = q + \frac{1}{\epsilon}\Delta^*(q_x + \epsilon q).$$

Hence, there exists $q_z \in Q$ such that (see (4.25))

$$\frac{q_z - q_x}{\epsilon} = q + \frac{1}{\epsilon}\Delta^*(q_x + \epsilon q)$$

or

$$\begin{aligned} q_z &= q_x + \epsilon q + \Delta^*(q_x + \epsilon q) \\ &= h^*(q_x + \epsilon q). \end{aligned}$$

Recall that $q \in \bar{U}_{q'} \subset (N_T - q_x)^\circ$ so that $q_x + \epsilon q \in N_T{}^\circ$. Since h^* is a homeomorphism, there exists $q_z \in Q \cap N^\circ$, q_z arbitrarily close to q_x, which contradicts the fact that x is an F, N, Φ extremal. Therefore, K and $N_T - q_x$ can be separated.

Thus, there exists a hyperplane Γ of dimension $2n$ through 0 such that K is contained in one of the closed half-spaces defined by Γ and $N_T - q_x$ is contained in the other.

Let η be a nonzero normal vector to Γ such that

$$\eta \cdot p \leqq 0 \leqq \eta \cdot w \tag{4.29}$$

for $p \in K$ and $w \in N_T - q_x$.

It is easily seen that $M_T - q_x \subset \Gamma$. For let $\xi \in M_T - q_x$. Since $M_T - q_x \subset N_T - q_x$, we have $\eta \cdot \xi \geqq 0$. But $M_T - q_x$ is a plane through the origin. Hence, $-\xi \in M_T - q_x$, and so $\eta \cdot (-\xi) \geqq 0$. Thus, $\eta \cdot \xi = 0$ when $\xi \in M_T - q_x$ and therefore, $\xi \in \Gamma$ or $M_T - q_x \subset \Gamma$.

This gives η orthogonal to $M_T - q_x$ or η orthogonal to $\tilde{M}$ at q_x.

Let $\eta = (b_0, b_1, a_1)$ be the normal in R^{2n+1}, where b_0, b_1 are n-vectors and a_1 is a scalar. From the definitions of K and η, it follows that

$$b_0 \cdot \delta z_0 + b_1 \cdot (\delta z_1 + \delta t \hat{f}_1) + a_1 \delta t \leqq 0$$

for all $p = (\delta z_0, \delta z_1 + \delta t \hat{f}_1, \delta t) \in K$. This may be written (see 4.21)

$$\begin{aligned} b_0 \cdot \delta\phi(t_0) + b_1 \cdot \Big\{ & \int_{\omega_0}^{t_0} \Lambda(r(s), t_1) \hat{f}_y(x(r(s)), x(s), r(s)) \delta\phi(s) \dot{r}(s)\, ds \\ & + \Lambda(t_0, t_1) \delta\phi(t_0) + \int_{t_0}^{t_1} \Lambda(s, t_1) \delta f(x(s), x(\omega(s)), s)\, ds \Big\} \\ & + \delta t (b_1 \cdot \hat{f}_1 + a_1) \leqq 0 \end{aligned} \tag{4.30}$$

for arbitrary $(\delta t, \delta\phi, \delta f) \in A = R^1 \times D(M) \times ([F] - \hat{f})$. Since δt, $\delta\phi$, δf are arbitrary and independent of each other, (4.30) gives

$$\begin{aligned} & \{b_0 + b_1 \Lambda(t_0, t_1)\} \cdot \delta\phi(t_0) \\ & + \int_{\omega_0}^{t_0} b_1 \Lambda(r(s), t_1) \hat{f}_y(x(r(s)), x(s), r(s)) \delta\phi(s) \dot{r}(s)\, ds \leqq 0 \end{aligned} \tag{4.31}$$

for arbitrary $\delta\phi \in D(M)$;

$$a_1 + b_1 \cdot \hat{f}_1 = 0, \tag{4.32}$$

$$\int_{t_0}^{t_1} b_1 \Lambda(s, t_1) \delta f(x(s), x(\omega(s)), s)\, ds \leqq 0 \tag{4.33}$$

for arbitrary $\delta f \in [F] - \hat{f}$.

At this point, we return to the case of ν lags $\omega_j(t) = t - \theta_j(t)$, $j = 1, 2, \cdots, \nu$. As was pointed out previously, all of the preceding arguments can be carried out for multiple lags. Statement (4.32) remains the same. The inequality (4.33) becomes

$$\int_{t_0}^{t_1} b_1 \Lambda(s, t_1) \delta f(x(s), x(\omega_1(s)), \cdots, x(\omega_\nu(s)), s)\, ds \leqq 0 \tag{4.34}$$

for arbitrary $\delta f \in [F] - \hat{f}$. However, now $\Lambda(s, t)$ satisfies the system (1.3) (see Theorem 3) with

$$A(s) = \hat{f}_x(x(s), x(\omega_1(s)), \cdots, x(\omega_\nu(s)), s) = \hat{f}_x(s)$$

and

$$B_j(s) = \hat{f}_{y_j}(x(s), x(\omega_1(s)), \cdots, x(\omega_\nu(s)), s) = \hat{f}_{y_j}(s).$$

Also, $\delta z(t)$ satisfies (see Lemma 4.3) the equation

$$\delta\dot{z}(t) = \hat{f}_x(t)\delta z(t) + \sum_{j=1}^{\nu} \hat{f}_{y_j}(t)\delta z(\omega_j(t)) + \delta f(x(t), x(\omega_1(t)), \cdots, x(\omega_\nu(t)), t) \quad \text{for} \quad t \in [t_0, t_1 + \epsilon_2 | \delta t |],$$

$$\delta z(t) = \delta\phi(t) \quad \text{for} \quad t \in [\omega_\nu(t_0), t_0].$$

Hence, using Theorem 3, we see that statement (4.31) becomes (using the appropriate representation for δz in place of (4.21))

$$\{b_0 + b_1\Lambda(t_0, t_1)\}\cdot\delta\phi(t_0) + \sum_{j=1}^{\nu} \int_{\omega_j(t_0)}^{t_0} b_1\Lambda(r_j(s), t_1)\hat{f}_{y_j}(r_j(s))\delta\phi(s)\dot{r}_j(s)\, ds \leqq 0 \tag{4.35}$$

for all $\delta\phi \in D(M)$, where now $D(M) = \{\delta\phi \in AC([\omega_\nu(t_0), t_0], R^n)$: $|\delta\dot{\phi}(t)| \leqq M(t)$ a.e. on $[\omega_\nu(t_0), t_0]\}$. Now define

$$\psi(s) = b_1\Lambda(s, t_1), \qquad s \in [t_0, t_1]. \tag{4.36}$$

Then (4.35) becomes

$$\{b_0 + \psi(t_0)\}\cdot\delta\phi(t_0) + \sum_{j=1}^{\nu} \int_{\omega_j(t_0)}^{t_0} \psi(r_j(s))\hat{f}_{y_j}(r_j(s))\delta\phi(s)\dot{r}_j(s)\, ds \leqq 0 \tag{4.37}$$

for all $\delta\phi \in D(M)$. This gives

$$\sum_{j=1}^{\nu} \int_{\omega_j(t_0)}^{t_0} \psi(r_j(s))\hat{f}_{y_j}(r_j(s))\delta\phi(s)\dot{r}_j(s)\, ds \leqq 0 \tag{4.38}$$

for all $\delta\phi \in D(M)$ with $\delta\phi(t_0) = 0$. Taking $\delta\phi \in D(M)$ with $\delta\phi(t) = 0$ on $[\omega_{\nu-1}(t_0), t_0]$ and applying Lemma 2 of §1, this gives

$$\psi(r_\nu(s))\hat{f}_{y_\nu}(r_\nu(s))\dot{r}_\nu(s) = 0 \tag{4.39}$$

almost everywhere on $[\omega_\nu(t_0), \omega_{\nu-1}(t_0)]$.

Taking $\delta\phi \in D(M)$ with $\delta\phi(t) = 0$ on $[\omega_{\nu-2}(t_0), t_0]$ and applying Lemma 2 again (and using (4.38) and (4.39)), we obtain

$$\psi(r_\nu(s))\hat{f}_{y_\nu}(r_\nu(s))\dot{r}_\nu(s) + \psi(r_{\nu-1}(s))\hat{f}_{y_{\nu-1}}(r_{\nu-1}(s))\dot{r}_{\nu-1}(s) = 0$$

a.e. on $[\omega_{\nu-1}(t_0), \omega_{\nu-2}(t_0)]$.

Continuing this procedure gives statement (iii) of Theorem 5. It follows that (4.37) reduces to

$$\{b_0 + \psi(t_0)\}\cdot\delta\phi(t_0) \leqq 0$$

for $\delta\phi \in D(M)$, which implies

$$b_0 + \psi(t_0) = 0. \tag{4.40}$$

Condition (ii) of Theorem 5 follows immediately from (4.34) by choosing $\delta f = f - \hat{f}$ for any $f \in F$. Since b_1 is constant, it follows from the equation for Λ that ψ satisfies (i) of Theorem 5. Furthermore, $\psi(t_1) \not\equiv 0$. For $\psi(t_1) = 0$ implies $b_1\Lambda(t_1, t_1) = b_1 = 0$, which gives $a_1 = 0$ and $b_0 = 0$ from (4.32) and (4.40). This implies $\eta = (b_0, b_1, a_1) = 0$, which is a contradiction. Hence ψ is nontrivial. (In fact, $\psi(t) \not\equiv 0$ for t in $[\omega_1(t_1), t_1]$ since $\Lambda(t, t_1)$ is nonsingular on this interval. Even if $\omega_1(t_1) = t_1$, we have $\psi(t_1) \not\equiv 0$; hence by continuity of ψ, $\psi(t) \not\equiv 0$ on $[a, t_1]$ for some $a < t_1$. Hence, the condition (ii) of Theorem 5 is nontrivial.)

Finally, from (4.40) we have $\psi(t_0) = -b_0$. Also, $\Lambda(t_1, t_1) = I$ gives $\psi(t_1) = b_1$. And from (4.32) we get $a_1 = -b_1 \cdot \hat{f}_1 = -\psi(t_1) \cdot \hat{f}_1$. Thus, we have $(-\psi(t_0), \psi(t_1), -\psi(t_1) \cdot \hat{f}_1) = (b_0, b_1, a_1) = \eta$ is orthogonal to $\tilde{M}$ at q_x, which is condition (iv) of Theorem 5. This completes the proof of the theorem.

5. Control problems as extremal problems. In this section we show that the control problem formulated in §2 can be considered as an extremal problem as given in §3. Theorem 4 will then follow from Theorem 5.

Assume that we have the control problem as formulated in §2. Recall that T is a C^1 manifold in $\bar{G} \times \bar{G} \times I'$, $\bar{x}, \bar{y}_1, \cdots, \bar{y}_\nu, \bar{g}$ are $(n-1)$-dimensional vectors, and $\bar{g} = \bar{g}(\bar{x}, \bar{Y}, u, t) = \bar{g}(\bar{x}, \bar{y}_1, \cdots, \bar{y}_\nu, u, t)$, where u is an υ-vector.

Let $\bar{\Omega}$ be the class of elements of the form $(\bar{\phi}, u, \bar{z}, t_1)$, where

(i) $u \in \Omega$, $\bar{\phi} \in \bar{\Phi}$,

(ii) $t_1 \in R^1$, $t_1 > t_0$,

(iii) $\bar{z}: [t_0, t_1] \to \bar{G} \subset R^{n-1}$ is a solution to

$$\begin{aligned} \dot{\bar{z}}(t) &= \bar{g}(\bar{z}(t), \bar{z}(\omega_1(t)), \cdots, \bar{z}(\omega_\nu(t)), u(t), t) \quad \text{for} \quad t \in [t_0, t_1], \\ \bar{z}(t) &= \bar{\phi}(t) \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{for} \quad t \in [\omega_\nu(t_0), t_0] \end{aligned} \tag{5.1}$$

satisfying $(\bar{z}(t_0), \bar{z}(t_1), t_1) \in T$.

The problem then becomes: Find $(\bar{\phi}^*, u^*, \bar{x}^*, t_1^*) \in \bar{\Omega}$ such that $J[\bar{\phi}^*, u^*, \bar{x}^*, t_1^*] \leqq J[\bar{\phi}, u, \bar{z}, t_1]$ for all $(\bar{\phi}, u, \bar{z}, t_1) \in \bar{\Omega}$. That is, minimize J over $\bar{\Omega}$.

Let $(\bar{\phi}^*, u^*, \bar{x}^*, t_1^*)$ be a solution to this problem. Put

$$x^*(t) \equiv \begin{pmatrix} x^{0*}(t) \\ \bar{x}^*(t) \end{pmatrix}, \qquad \omega_\nu(t_0) \leqq t \leqq t_1^*,$$

where

$$x^{0*}(t) \equiv \int_{t_0}^{t} g^0(\bar{x}^*(s), \bar{x}^*(\omega_1(s)), \cdots, \bar{x}^*(\omega_\nu(s)), u^*(s), s)\, ds \quad \text{for} \quad t \in [t_0, t_1^*],$$

$$x^{0*}(t) \equiv 0 \quad \text{for} \quad t \in [\omega_\nu(t_0), t_0].$$

Define

$$\tilde{g}(x, Y, u, t) \equiv \begin{pmatrix} g^0(\bar{x}, \bar{Y}, u, t) \\ \bar{g}(\bar{x}, \bar{Y}, u, t) \end{pmatrix},$$

where

$$x = (x^0, \bar{x}) \in R^1 \times \bar{G} \equiv G,$$
$$y_j = (y^0, \bar{y}_j) \in R^1 \times \bar{G}, \qquad j = 1, 2, \cdots, \nu,$$
$$Y = (y_1, \cdots, y_\nu).$$

Then the system (5.1) becomes the n-system

$$\begin{aligned} \dot{x}(t) &= \tilde{g}(x(t), x(\omega_1(t)), \cdots, x(\omega_\nu(t)), u(t), t) \quad \text{for} \quad t \geqq t_0, \\ x(t) &= \phi(t) \quad \text{for} \quad t \in [\omega_\nu(t_0), t_0], \end{aligned} \tag{5.2}$$

where $\phi = (0, \bar{\phi})$. Defining $\Phi = \{\phi : \phi = (\phi^0, \bar{\phi}), \bar{\phi} \in \bar{\Phi}, \phi^0 \in A\,C([\omega_\nu(t_0), t_0], R^1)\}$, we have $\phi \in \Phi$ implies $\phi \in A\,C([\omega_\nu(t_0), t_0], G)$. By the above definition, we have $\tilde{g}$ is an n-vector function defined on $G \times G^\nu \times R_U \times I'$.

Define $F = \{f(x, Y, t) : f(x, Y, t) = \tilde{g}(x, Y, u(t), t), u \in \Omega\}$. It follows from the Borel measurability of $\tilde{g}$ in u, t that each f in F is measurable in t on I'.

Let R^{2n+1} be considered as the space of parameters $(\eta^0, \eta, \xi^0, \xi, \tau)$ where η^0, ξ^0, τ are scalars and η, ξ are $(n-1)$-vectors. Then define $N \subset R^{2n+1}$ to be all $(\eta^0, \eta, \xi^0, \xi, \tau)$ with (η, ξ, τ) near $(\bar{x}^*(t_0), \bar{x}^*(t_1^*), t_1^*)$ satisfying

$$(\eta, \xi, \tau) \in T, \quad \eta^0 = 0, \qquad \xi^0 \leqq x^{0*}(t_1^*).$$

Define $\tilde{M}$ to be the above set with the last inequality replaced by equality. Then N is a C^1 manifold with boundary $\tilde{M}$.

Under the above definitions and assumptions we have the following lemma.

LEMMA 5.1. *x^* is an F, N, Φ extremal.*

Proof. We have

$$\begin{aligned} q_{x^*} &= (x^*(t_0), x^*(t_1^*), t_1^*) \\ &= (x^{0*}(t_0), \bar{x}^*(t_0), x^{0*}(t_1^*), \bar{x}^*(t_1^*), t_1^*). \end{aligned}$$

Clearly, $q_{x^*} \in \tilde{M}$. Furthermore, there is a neighborhood V of q_{x^*} such that $Q \cap N \cap V \subset \tilde{M}$, where Q is the set of all endpoints q_x corresponding to solutions of (5.2) for $u \in \Omega$, $\phi \in \Phi$. Suppose not. Then in any neighborhood of q_{x^*} there exists $q_{x'} \in Q$ such that $q_{x'}$ is in the interior of N. That is, there exists (ϕ', u', x', t_1'), where $x' = (x^{0'}, \bar{x}')$, with $(\bar{\phi}', u', \bar{x}', t_1') \in \bar{\Omega}$, such that x' is a solution of (5.2) corresponding to $u' \in \Omega$, $\phi' \in \Phi$ with $(x'(t_0), x'(t_1'), t_1')$ in the interior of N. This implies that $x^{0'}(t_1') < x^{0*}(t_1^*)$, which implies $(\bar{\phi}^*, u^*, \bar{x}^*, t_1^*)$ is not minimizing over $\bar{\Omega}$, a contradiction.

Thus x^* is an F, N, Φ extremal.

To obtain Theorem 4 from Theorem 5, it remains only to show that F defined above is quasi-convex. Taking $f_i(z, t) = f_i(x, Y, t) = \tilde{g}(x, Y, u_i(t), t)$, where $u_i \in \Omega$ and $z = (x, Y)$ is in the compact convex metric space $Z = X \times X^\nu$, and applying a lemma of Gamkrelidze (see [1, Lemma 4.1]), it is not hard to show that F is quasi-convex.

Thus, we have Theorem 5 holds with x^* the F, N, Φ extremal on $[t_0, t_1^*]$ corresponding to (f^*, ϕ^*) in $F \times \Phi$, where

$$\hat{f} = f^*(x, Y, t) = \tilde{g}(x, Y, u^*(t), t),$$

$$\hat{\phi} = \phi^* = (0, \bar{\phi}^*).$$

We show now that Theorem 4 follows from Theorem 5. Taking $\psi(t)$ from Theorem 5 and calling it $\tilde{\psi}(t) = (\psi^0(t), \psi^1(t), \cdots, \psi^{n-1}(t))$, we see that $\psi^0 = 0$ since $\partial\tilde{g}/\partial x^0 = \partial\tilde{g}/\partial y_j^0 = 0$. Hence $\psi^0(t) = \text{const}$. It also follows from (i) in Theorem 5 that $\psi^1(t), \cdots, \psi^{n-1}(t)$ satisfy the equations in (i) of Theorem 4. Statement (ii) of Theorem 4 follows immediately from (ii) of Theorem 5. Furthermore, since $\partial\tilde{g}/\partial y_j^0 = 0, j = 1, \cdots, \nu$, (iii) of Theorem 4 follows from (iii) of Theorem 5. Under the above definitions, (iv) of Theorem 5 becomes:

$$(5.3) \qquad (-\tilde{\psi}(t_0), \tilde{\psi}(t_1^*), -\tilde{\psi}(t_1^*)\cdot\tilde{g}_1^*)$$

is orthogonal to $\tilde{M}$ *at*

$$q_{x^*} = (x^*(t_0), x^*(t_1^*), t_1^*) = (0, \bar{x}^*(t_0), x^{0*}(t_1^*), \bar{x}^*(t_1^*), t_1^*),$$

where $\tilde{\psi} = (\psi^0, \bar{\psi})$.

From the definition of N and $\tilde{M}$, it is clear that whenever (η_T, ξ_T, τ_T) is any tangent vector to T at $(\bar{x}^*(t_0), \bar{x}^*(t_1^*), t_1^*)$ in R^{2n-1}, then vector $(0, \eta_T, 0, \xi_T, \tau_T)$ will be tangent to $\tilde{M}$ at $(0, \bar{x}^*(t_0), x^{0*}(t_1^*), \bar{x}^*(t_1^*), t_1^*)$. This taken with (5.3) gives that $(-\bar{\psi}(t_0), \bar{\psi}(t_1^*), -\bar{\psi}(t_1^*)\cdot\tilde{g}_1^*)$ is orthogonal to any (η_T, ξ_T, τ_T) tangent to T at $(\bar{x}^*(t_0), \bar{x}^*(t_1^*), t_1^*)$ which is (iv) of Theorem 4.

To show that $\psi^0 \leqq 0$, we recall that $\tilde{\psi}(t) = b_1\Lambda(t, t_1^*)$, where $\eta = (b_0, b_1, a_1)$ was the normal to the separating hyperplane Γ in the proof

of Theorem 5. Since $\Lambda(t_1^*, t_1^*) = I$, we get $\psi^0 = \psi^0(t_1^*) = b_1^1$, where $b_1 = (b_1^1, \cdots, b_1^n)$.

Recall also that η was chosen so that

$$\eta \cdot w \geqq 0 \tag{5.4}$$

for all w in $N_T - q_{x^*}$. From the definition of N in our application, we see that the curve $(x^*(t_0), x^{0*}(t_1^*) - \sigma, \bar{x}^*(t_1^*), t_1^*)$, $0 \leqq \sigma < \infty$, is in N, and starts at q_{x^*}. Hence, $(x^*(t_0), x^{0*}(t_1^*) - 1, \bar{x}^*(t_1^*), t_1^*)$ is in N_T or $(0_n, -1, 0_{n-1}, 0)$ is in $N_T - q_{x^*}$. Using this and (5.4) gives $b_1^1(-1) \geqq 0$ or $b_1^1 = \psi^0 \leqq 0$.

To study the zeros of $\tilde{\psi}(t)$ on $[t_0, t_1^*]$, we recall that $\tilde{\psi}(t) \not\equiv 0$ on $[\omega_1(t_1^*), t_1^*]$ and that $\tilde{\psi}$ satisfies an advanced differential-difference equation on $[t_0, \omega_1(t_1^*)]$. Now consider the example (where $y = (y^1, y^2)$):

$$\dot{y}(t) = y(t+1), \quad t \in [-1, 0],$$

$$y(t) = (1 + t, 0), \quad t \in [0, 1].$$

This has solution $y(t) = (1 + 2t + t^2/2, 0)$ for $-1 \leqq t \leqq 0$, which has [a] zero in $(-1, 0)$ at $t = -2 + \sqrt{2}$. Hence, for linear systems, initial functions that vanish nowhere do not necessarily give nowhere vanishing solutions to advanced equations. Furthermore, if we consider the above system with initial function $x(t) = (0, -1 - t)$ on $[0, 1]$, we then get a solution $x(t) = (0, -1 + 2t + t^2/2)$ on $[-1, 0]$. Thus we can make the following statement about "fundamental matrix" solutions: There are simple examples of linear advanced systems such that linearly independent initial functions do not give linearly independent solutions. In particular, $\Lambda(s, t_1^*)$ can be nonsingular on $[\omega_1(t_1^*), t_1^*]$ and yet be singular at points in $[t_0, \omega_1(t_1^*)]$. Thus it is possible that $\tilde{\psi}$ has zeros in $[t_0, \omega_1(t_1^*)]$ even though $\tilde{\psi}$ is nonzero on $[\omega_1(t_1^*), t_1^*]$. (As was pointed out in §4, even if $\omega_1(t_1^*) = t_1^*$, we have $\tilde{\psi} \not\equiv 0$ on $[a, t_1^*]$ for some $a < t_1^*$. This guarantees that the condition (ii) of Theorem 4 is nontrivial.)

We complete the proof of statement (v) of Theorem 4 by proving the lemma below. This will also complete the proof of Theorem 4.

LEMMA 5.2. *Suppose*

(i) $\theta_\nu(t) > \theta_{\nu-1}(t) > \cdots > \theta_1(t) > 0$ *on* $[t_0, t_1^*]$,

(ii) $\bar{g}_{\bar{\nu}_j}(\bar{x}^*(t), \bar{x}^*(\omega_1(t)), \cdots, \bar{x}^*(\omega_\nu(t)), u^*(t), t)$ *is nonsingular for* $t \in [t_0, t_1^*]$, $j = 1, 2, \cdots, \nu$.

Then $\psi^0 \not\equiv 0$ *and hence* $\tilde{\psi}$ *is nonzero on* $[t_0, t_1^*]$.

Proof. We give the proof for the case $\nu = 2$. The same type of arguments will also prove the lemma for cases $\nu > 2$. (The arguments for $\nu = 1$ are even simpler than those given below.) So assume $\nu = 2$. Denote by $\bar{g}_{\bar{\nu}_j}(s)$ the matrix functions $\bar{g}_{\bar{\nu}_j}(\bar{x}^*(s), \bar{x}^*(\omega_1(s)), \bar{x}^*(\omega_2(s)), u^*(s), s)$, $j = 1, 2$.

Since (i) holds, we have

$$\omega_2(t) < \omega_1(t) < t,$$

so that

$$r_2(t) > r_1(t) > t.$$

Since r_1 and r_2 are continuous, there exists a $\delta > 0$ such that $r_2(\tau) - r_1(\tau) > \delta$. Hence, for any τ, the interval $[\tau, r_2(\omega_1(\tau))] = [r_1(\omega_1(\tau)), r_2(\omega_1(\tau))]$ has length $> \delta$. From (iii) of Theorem 4 we have

$$\tilde{\psi}(r_2(s))\tilde{g}_{\tilde{y}_2}(r_2(s))\dot{r}_2(s) = 0 \quad \text{a.e. on} \quad [\omega_2(t_0), \omega_1(t_0)] \tag{5.5}$$

and

$$\tilde{\psi}(r_1(s))\tilde{g}_{\tilde{y}_1}(r_1(s))\dot{r}_1(s) + \tilde{\psi}(r_2(s))\tilde{g}_{\tilde{y}_2}(r_2(s))\dot{r}_2(s) = 0 \quad \text{a.e. on} \quad [\omega_1(t_0), t_0]. \tag{5.6}$$

We have also that $\tilde{\psi}$ satisfies

$$\dot{\tilde{\psi}}(t) + \tilde{\psi}(t)\tilde{g}_{\tilde{x}}(t) = 0 \quad \text{on} \quad [\omega_1(t_1^*), t_1^*], \tag{5.7}$$

$$\dot{\tilde{\psi}}(t) + \tilde{\psi}(t)\tilde{g}_{\tilde{x}}(t) + \tilde{\psi}(r_1(t))\tilde{g}_{\tilde{y}_1}(r_1(t))\dot{r}_1(t) = 0 \quad \text{on} \quad [\omega_2(t_1^*), \omega_1(t_1^*)], \tag{5.8}$$

$$\dot{\tilde{\psi}}(t) + \tilde{\psi}(t)\tilde{g}_{\tilde{x}}(t) + \tilde{\psi}(r_1(t))\tilde{g}_{\tilde{y}_1}(r_1(t))\dot{r}_1(t) + \tilde{\psi}(r_2(t))\tilde{g}_{\tilde{y}_2}(r_2(t))\dot{r}_2(t) = 0 \quad \text{on} \quad [t_0, \omega_2(t_1^*)]. \tag{5.9}$$

Recall that $\tilde{\psi}(t) = (\psi^0, \bar{\psi}(t)) \not\equiv 0$ on $[\omega_1(t_1^*), t_1^*]$. To prove the lemma, we suppose that $\psi^0 = 0$ and show that this implies $\bar{\psi}(t)$ has zeros in $[\omega_1(t_1^*), t_1^*]$, which gives a contradiction.

Assume $\psi^0 = 0$. From (5.5) we then get $\bar{\psi}(r_2(s))\bar{g}_{\bar{y}_2}(r_2(s))\dot{r}_2(s) = 0$ almost everywhere on $[\omega_2(t_0), \omega_1(t_0)]$. Recall that $\dot{r}_2(s) > 0$. Using this and the nonsingularity of $\bar{g}_{\bar{y}_2}$ gives $\bar{\psi}(r_2(s)) = 0$ almost everywhere, hence everywhere, on $[\omega_2(t_0), \omega_1(t_0)]$. This implies $\bar{\psi}(s) = 0$ on $[t_0, r_2(\omega_1(t_0))]$.

Consider the two following cases.

Case 1. $r_1(t_0) \leqq r_2(\omega_1(t_0))$. Then $\bar{\psi}(s) = 0$ on $[t_0, r_1(t_0)]$ or $\bar{\psi}(r_1(s)) = 0$ on $[\omega_1(t_0), t_0]$. Using this in (5.6) gives $\bar{\psi}(r_2(s)) = 0$ on $[\omega_1(t_0), t_0]$ or $\bar{\psi}(s) = 0$ on $[r_2(\omega_1(t_0)), r_2(t_0)]$. Hence, $\bar{\psi}(s) = 0$ on $[t_0, r_2(t_0)]$.

Case 2. $r_1(t_0) > r_2(\omega_1(t_0))$. Then $\bar{\psi}(s) = 0$ on $[t_0, r_2(\omega_1(t_0))]$ implies $\bar{\psi}(r_1(s)) = 0$ on $[\omega_1(t_0), \omega_1 r_2\omega_1(t_0)] \subset [\omega_1(t_0), t_0]$. Using this in (5.6) gives $\bar{\psi}(r_2(s)) = 0$ on $[\omega_1(t_0), \omega_1 r_2\omega_1(t_0)]$ or $\bar{\psi}(s) = 0$ on $[r_2\omega_1(t_0), r_2\omega_1 r_2\omega_1(t_0)]$. Thus, $\bar{\psi}(s) = 0$ on $[t_0, r_2\omega_1 r_2\omega_1(t_0)]$. If $r_2(t_0) > r_2\omega_1 r_2\omega_1(t_0) \geqq r_1(t_0)$, then one can use the arguments in Case 1 to get $\bar{\psi}(s) = 0$ on $[t_0, r_2(t_0)]$. If $r_2\omega_1 r_2\omega_1(t_0) < r_1(t_0)$, then repeating the above arguments one finds $\bar{\psi}(s) = 0$ on $[t_0,$

$r_2\omega_1 r_2\omega_1 r_2\omega_1(t_0)]$. Repeating this a finite number of times, one eventually has $\bar{\psi}(s) = 0$ on an interval containing $[t_0, r_1(t_0)]$ so that the arguments in Case 1 give $\bar{\psi}(s) = 0$ on $[t_0, r_2(t_0)]$.

Hence, in either case, we have $\bar{\psi}(s) = 0$ on $[t_0, r_2(t_0)]$.

If $r_2(t_0) \geqq \omega_1(t_1^*)$, we have a contradiction.

If $\omega_1(t_1^*) > r_2(t_0) > \omega_2(t_1^*)$, then we have $\bar{\psi}(s) = 0$ on $[\omega_2(t_1^*), a] \subset [\omega_2(t_1^*), \omega_1(t_1^*)]$. Using (5.8) gives $\bar{\psi}(r_1(s))\bar{g}_{\bar{\nu}_1}(r_1(t))\dot{r}_1(t) = 0$ a.e. on $[\omega_2(t_1^*), a]$ which implies $\bar{\psi}(s) = 0$ on $[r_1\omega_2(t_1^*), r_1(a)]$. If $r_1(a) > \omega_1(t_1^*)$ we get a contradiction. If $r_1(a) \leqq \omega_1(t_1^*)$, we repeat the above argument and get $\bar{\psi}(s) = 0$ on $[r_1 r_1\omega_2(t_1^*), r_1 r_1(a)]$. After a finite number of such steps we get $\bar{\psi}(s)$ vanishing at points in $[\omega_1(t_1^*), t_1^*]$, which is a contradiction. (Note that given a, $r_1 r_1 \cdots r_1(a) > \omega_1(t_1^*)$ after a finite number of times. Also, if $\dot{r}_1(t) > \delta_1$ and $a - b = \epsilon$, then $r_1(a) - r_1(b) \geqq \delta_1\epsilon$ so that $[r_1 r_1 \cdots r_1\omega_2(t_1^*), r_1 r_1 \cdots r_1(a)]$ has positive length after a finite number of steps.)

If $r_2(t_0) \leqq \omega_2(t_1^*)$, then $\bar{\psi}(s) = 0$ on $[t_0, r_2(t_0)]$ implies (using (5.9))

$$\bar{\psi}(r_1(s))\bar{g}_{\bar{\nu}_1}(r_1(s))\dot{r}_1(s) + \bar{\psi}(r_2(s))\bar{g}_{\bar{\nu}_2}(r_2(s))\dot{r}_2(s) = 0 \tag{5.10}$$

a.e. on $[t_0, r_2(t_0)]$. Since $r_1(t_0) < r_2(t_0)$, we have $\bar{\psi}(s) = 0$ on $[r_1(t_0), r_2(t_0)]$ which implies $\bar{\psi}(r_1(s)) = 0$ on $[t_0, \omega_1 r_2(t_0)]$. Using (5.10) this gives $\bar{\psi}(r_2(s)) = 0$ on $[t_0, \omega_1 r_2(t_0)]$ which implies $\bar{\psi}(s) = 0$ on $[r_2(t_0), r_2\omega_1 r_2(t_0)]$. Repeating this argument a finite number of times gives $\bar{\psi}(s) = 0$ on a subinterval of $[\omega_2(t_1^*), \omega_1(t_1^*)]$. Then the arguments for the case $r_2(t_0) > \omega_2(t_1^*)$ can be used to get $\bar{\psi}$ vanishing at points in $[\omega_1(t_1^*), t_1^*]$, which is again a contradiction.

Thus, under assumptions (i) and (ii) of the lemma, $\psi^0 \neq 0$ and hence $\tilde{\psi}(s) = (\psi^0, \bar{\psi}(s))$ is nonzero on $[t_0, t_1^*]$.

6. Systems with a lag in the control. Although lags in the controls were not considered in the previous sections, we shall consider such problems now. However, we shall assume that the control system is linear in the controls. Then Theorem 5 will be applicable. To simplify notation, we again restrict our considerations to a single lag $\omega(t) = t - \theta(t)$ in the state variable.

Let $\zeta(t)$ satisfy the conditions put on θ in §1. Put $\lambda(t) = t - \zeta(t)$ and let I be a bounded open interval containing $[\alpha_0, t_0]$, where $\alpha_0 = \min\{\omega(t_0), \lambda(t_0)\}$.

Consider the system

$$\begin{aligned} \dot{\bar{x}}(t) &= \bar{g}(\bar{x}(t), \bar{x}(\omega(t)), t) + \bar{A}(t)u(t) + \bar{B}(t)u(\lambda(t)), \quad t > t_0, \\ \bar{x}(t) &= \bar{\phi}(t) \quad \text{on} \quad [\omega_0, t_0], \end{aligned} \tag{6.1}$$

where $\bar{x}$ is an $(n-1)$-vector, u is an ν-vector and $\bar{\phi} \in \bar{\Phi}$, with $\bar{\Phi}$ as defined in §2. We take the C^1 manifold T as defined in §2. For $1 < p < \infty$ and f

any k-vector we define

$$\| f \|_p = \{ \sum_1^k |f_i|^p \}^{1/p}.$$

Then a vector function $f(t)$ will be said to be in $L_p(I)$ if

$$\int_I \| f(t) \|_p{}^p \, dt < \infty.$$

Let $U(t) = R^v$ for each $t \in I$. Define

$$\hat{\Omega} = \left\{ u \in L_p(I) : u(t) \in R^v, \int_I \| u \|_p{}^p \leqq 1 \right\}.$$

We assume $\bar{g}$ is C^1 in x, y and measurable in t, and is dominated along with its partials by an $m \in L_1(I)$. The $(n-1) \times v$ matrices $\bar{A}$ and $\bar{B}$ are in $L_q(I)$, where $1/q + 1/p = 1$.

Let

$$J = \int_{t_0}^{t_1} \{ g^0(\bar{x}(t), \bar{x}(\omega(t)), t) + A^0(t)u(t) + B^0(t)u(\lambda(t)) \} \, dt,$$

where g^0 is a scalar function, A^0, B^0 are $1 \times v$ vectors, satisfying the same hypotheses as $\bar{g}$, $\bar{A}$, $\bar{B}$, respectively.

Then the problem is to minimize J over $(\bar{\phi}, u, t_1)$ in $\bar{\Phi} \times \hat{\Omega} \times I'$ subject to equations (6.1) and $(\bar{x}(t_0), \bar{x}(t_1), t_1) \in T$. Suppose $(\bar{\phi}^*, u^*, t_1{}^*)$ is a solution with corresponding trajectory $\bar{x}^*$.

Set

$$x^{0*}(t) = \int_{t_0}^{t} \{ g^0(\bar{x}^*(s), \bar{x}^*(\omega(s)), s) + A^0(s)u^*(s) + B^0(s)u^*(\lambda(s)) \} \, ds \text{ for } t > t_0,$$

$$x^{0*}(t) = 0 \quad \text{for} \quad t \in [\omega_0 , t_0].$$

Let

$$x = (x^0, \bar{x}), \quad \tilde{g} = (g^0, \bar{g}), \quad \tilde{A} = \begin{pmatrix} A^0 \\ \bar{A} \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} B^0 \\ \bar{B} \end{pmatrix}.$$

Define

$$F = \{ f(x, y, t) : f(x, y, t) = \tilde{g}(x, y, t) + \tilde{A}(t)u(t) + \tilde{B}(t)u(\lambda(t)), \, u \in \hat{\Omega} \}$$

and

$$\Phi = \{ (\phi^0, \bar{\phi}) : \bar{\phi} \in \bar{\Phi}, \phi^0 \in AC([\omega_0 \, t_0]R^1) \}.$$

The sets N and $\tilde{M}$ are defined exactly as in §5. Then, as before, one can show that $x^* = (x^{0*}, \bar{x}^*)$ is an F, N, Φ extremal. Furthermore, F is quasi-con-

vex. (It is convex, hence one may take $g_\epsilon = 0$ in statement 3 of Definition 3.1. Statements 1 and 2 are satisfied from the hypotheses on $\tilde{g}$, $\tilde{A}$, $\tilde{B}$, and u.) Thus, Theorem 5 is applicable and Theorem 5(ii) gives

$$\int_{t_0}^{t_1^*} \tilde{\psi}(t)\{\tilde{A}(t)u^*(t) + \tilde{B}(t)u^*(\lambda(t))\}\,dt \geqq \int_{t_0}^{t_1^*} \tilde{\psi}(t)\{\tilde{A}(t)u(t) + \tilde{B}(t)u(\lambda(t))\}\,dt \tag{6.2}$$

for all $u \in \hat{\Omega}$. This can be written

$$\int_{t_0}^{t_1^*} \tilde{\psi}(t)\tilde{A}(t)\{u^*(t) - u(t)\}\,dt + \int_{t_0}^{t_1^*} \tilde{\psi}(t)\tilde{B}(t)\{u^*(\lambda(t)) - u(\lambda(t))\}\,dt \geqq 0.$$

Letting $t = \rho(s)$ in the second integral above where $\rho = \lambda^{-1}$ we have

$$\int_{t_0}^{t_1^*} \tilde{\psi}(t)\tilde{A}(t)\{u^*(t) - u(t)\}\,dt + \int_{\lambda(t_0)}^{\lambda(t_1^*)} \tilde{\psi}(\rho(s))\tilde{B}(\rho(s))\{u^*(s) - u(s)\}\dot{\rho}(s)\,ds \geqq 0$$

for $u \in \hat{\Omega}$. We assume in the following discussion that $\lambda(t_1^*) > t_0$. Obvious changes can be made to carry out the same arguments if $\lambda(t_1^*) \leqq t_0$. Then the above may be written

$$\int_{\lambda(t_0)}^{t_0} \tilde{\psi}(\rho(t))\tilde{B}(\rho(t))\dot{\rho}(t)\{u^*(t) - u(t)\}\,dt + \int_{t_0}^{\lambda(t_1^*)} \{\tilde{\psi}(t)\tilde{A}(t) + \tilde{\psi}(\rho(t))\tilde{B}(\rho(t))\dot{\rho}(t)\}\{u^*(t) - u(t)\}\,dt + \int_{\lambda(t_1^*)}^{t_1^*} \tilde{\psi}(t)\tilde{A}(t)\{u^*(t) - u(t)\}\,dt \geqq 0$$

for all $u \in \hat{\Omega}$.

Next define

$$K(t) = \begin{cases} \dot{\rho}(t)\tilde{B}'(\rho(t))\tilde{\psi}'(\rho(t)) & \text{for } t \in [\lambda(t_0), t_0), \\ \dot{\rho}(t)\tilde{B}'(\rho(t))\tilde{\psi}'(\rho(t)) + \tilde{A}'(t)\tilde{\psi}'(t) & \text{for } t \in [t_0, \lambda(t_1^*)], \\ \tilde{A}'(t)\tilde{\psi}'(t) & \text{for } t \in (\lambda(t_1^*), t_1^*], \\ 0 & \text{for } t \in I - [\lambda(t_0), t_1^*], \end{cases}$$

where $'$ denotes the transpose. Then $K \in L_q(I)$. (We assume that $\int_I \| K \|_q^q \neq 0$. If this integral is zero, it is clear that the maximum prin-

ciple will give no information. Note that it is possible to put additional hypotheses on the problem to insure that $\int_I \| K \|_q^q \neq 0$. For example, assuming that $\tilde{A}(t)$ has full rank would be sufficient.)

From the definition of K, we see that the maximum principle becomes

$$\int_I K(t)\cdot u^*(t)\, dt \geqq \int_I K(t)\cdot u(t)\, dt$$

for all $u \in \hat{\Omega}$.

Now define $\bar{u}(t)$ by

$$\bar{u}^i(t) = \frac{\operatorname{sgn} K^i(t)|K^i(t)|^{q-1}}{\left\{\int_I \| K(t) \|_q^q\right\}^{1/p}}, \qquad i = 1, \cdots, v. \tag{6.3}$$

Then $\bar{u} \in L_p(I)$ and $\int_I \| \bar{u} \|_p^p = 1$, so that $\bar{u} \in \hat{\Omega}$.

THEOREM 6. $u^*(t) = \bar{u}(t)$ a.e. *on* I.

Proof. From the maximum principle we have

$$\begin{aligned}
\int_I \sum K^i(t)\bar{u}^i(t)\, dt &\leqq \int_I \sum K^i(t)u^{i*}(t)\, dt \\
&\leqq \int_I \sum |K^i(t)|\, |u^{i*}(t)|\, dt \\
&\leqq \int_I \{\sum |K^i(t)|^q\}^{1/q}\{\sum |u^{i*}(t)|^p\}^{1/p}\, dt \\
&= \int_I \|K(t)\|_q \|u^*(t)\|_p\, dt \\
&\leqq \left\{\int_I \|K(t)\|_q^q\right\}^{1/q}\left\{\int_I \|u^*(t)\|_p^p\right\}^{1/p} \\
&\leqq \left\{\int_I \|K(t)\|_q^q\right\}^{1/q},
\end{aligned}$$

where we have used the Hölder inequality for sums and for integrals.

An easy calculation shows that

$$\int_I \sum K^i(t)\bar{u}^i(t)\, dt = \left\{\int_I \|K(t)\|_q^q\right\}^{1/q}$$

so that equality holds throughout the above inequalities. In particular, equality holds in the Hölder inequality for sums. It follows that

$$|u^{i*}(t)\,| = \beta |\, K^i(t)\,|^{q-1}$$

a.e. on I. But

$$\int_I \sum |K^i(t)|\,|u^{i*}(t)|\,dt = \left\{\int_I \|K(t)\|_q^q\right\}^{1/q}$$

implies $\beta = \left\{\int_I \|K(t)\|_q^q\right\}^{-1/p}$. Hence we get that $|u^{i*}(t)| = |\bar{u}^i(t)|$ a.e. on I, $i = 1, \cdots, v$. It follows from the maximum principle and the fact that $K^j(t)\bar{u}^j(t) \geqq 0$ on I that sgn $u^{i*}(t) =$ sgn $\bar{u}^i(t)$ a.e. on I which completes the proof.

Acknowledgment. The author wishes to express his sincere appreciation to Professor L. D. Berkovitz of Purdue University for many helpful suggestions concerning this paper which forms part of the author's doctoral dissertation.

REFERENCES

[1] R. V. Gamkrelidze, *On some extremal problems in the theory of differential equations with applications to the theory of optimal control*, this Journal, 3 (1965), pp. 106–128.

[2] L. W. Neustadt, *An abstract variational theory with applications to a broad class of optimization problems*, this Journal, 4 (1966), pp. 505–527.

[3] Avner Friedman, *Optimal control for hereditary processes*, Arch. Rational Mech. Anal., 15 (1964), pp. 396–416.

[4] M. N. Oguztoreli, *Time-Lag Control Systems*, Academic Press, New York, 1966.

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

[6] N. Dunford and J. T. Schwartz, *Linear Operators I*, Interscience, New York, 1958.

[7] E. J. McShane, *Integration*, Princeton University Press, Princeton, 1944.

[8] I. P. Natanson, *Theory of Functions of a Real Variable*, Ungar, New York, 1955.

[9] V. Volterra, *Theory of Functionals*, Blackie, London, 1930.

[10] S. Lang, *Introduction to Differentiable Manifolds*, Interscience, New York, 1962.

[11] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

INSTABILITY OF PERIODICALLY TIME-VARYING LINEAR CONTROL SYSTEMS*

THEODORE A. BICKART†

1. Introduction. Consider the control system of Fig. 1. As indicated, the signals q and r are related by the differential equation

$$\sum_{i=0}^{d} a_i \frac{d^i r}{dt^i} = \sum_{i=0}^{n} b_i \frac{d^i q}{dt^i}$$

and the algebraic equation

$$q = k(t)[s(t) - r].$$

It is assumed that $n \leqq d$, $a_0, a_1, \cdots, a_d$ and $b_0, b_1, \cdots, b_n$ are real constants, $a_d \neq 0$, $b_n \neq 0$, and $s(t)$ and $k(t)$ are n-times differentiable for $t \geqq t_0$. Then, it is known [1] that for $t \geqq t_0$ there exists a unique d-times differentiable function $r(t)$ which satisfies the system equations above and the initial conditions $d^i r/dt^i \,|_{t=t_0} = \rho_i$, $i = 0, 1, \cdots, d-1$. Further, it is seen that $q(t)$ is an n-times differentiable function. Thus, the first set of conditions in the following useful theorem [2] are satisfied.

THEOREM. *Let $a_0, a_1, \cdots, a_d$ and $b_0, b_1, \cdots, b_n$ be real constants such that $a_d \neq 0$ and $b_n \neq 0$. Let $q(t)$ be an n-times differentiable complex-valued function of t defined on $[t_0, \infty)$. Let $p(t)$ satisfy $\sum_{i=0}^{d} a_i(d^i p/dt^i) = q$ for $t \in [t_0, \infty)$. Then $p(t)$ is $(n + d)$-times differentiable on $[t_0, \infty)$ and $r = \sum_{i=0}^{n} b_i(d^i p/dt^i)$ satisfies*

$$\sum_{i=0}^{d} a_i \frac{d^i r}{dt^i} = \sum_{i=0}^{n} b_i \frac{d^i q}{dt^i}$$

for $t \in [t_0, \infty)$. Suppose further that $\sum_{i=0}^{d} a_i\lambda^i$ and $\sum_{i=0}^{n} b_i\lambda^i$ are relatively prime polynomials in λ. Let $\rho_0, \rho_1, \cdots, \rho_{d-1}$ be complex constants. Then there exist unique complex constants $\xi_0, \xi_1, \cdots, \xi_{d-1}$ such that, if $d^i p/dt^i \,|_{t=t_0} = \xi_i$, $i = 0, 1, \cdots, d-1$, then $d^i r/dt^i \,|_{t=t_0} = \rho_i$, $i = 0, 1, \cdots, d-1$.

Therefore, with the added assumption that $\sum_{i=0}^{d} a_i\lambda^i$ and $\sum_{i=0}^{n} b_i\lambda^i$ are relatively prime, this theorem implies that the system of Fig. 1 is equivalent to that of Fig. 2 in the sense: Let $s(t)$ and ρ_i, $i = 0, 1, \cdots, d-1$, be given; then the ξ_i, $i = 0, 1, \cdots, d-1$, exist such that the response of the system of Fig. 2 is identical to the response of the system of Fig. 1.

* Received by the editors May 29, 1967, and in revised form September 28, 1967.

† Department of Electrical Engineering, Syracuse University, Syracuse, New York 13210. This research was supported by the United States Air Force Rome Air Development Center under Contract AF 30(602)-3538, and by the Syracuse University Research Corporation.

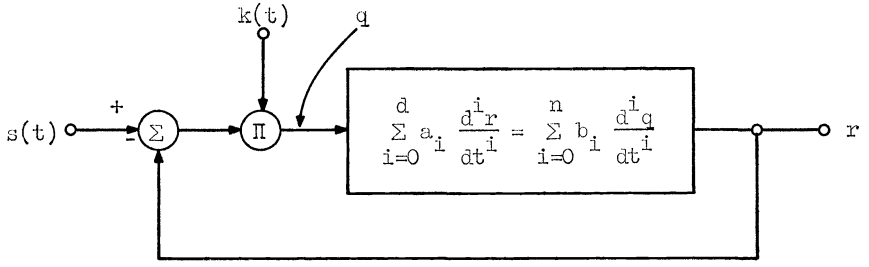


FIG. 1. *Time-varying control system with rational function filter*

Further, the uniqueness of the constants ξ_i, $i = 0, 1, \cdots, d - 1$, guarantees that there is a unique $p(t)$ such that

$$r = \sum_{i=0}^{n} b_i \frac{d^i p}{dt^i},$$

where p satisfies the differential equation

$$\sum_{i=0}^{d} [a_i + k(t)b_i] \frac{d^i p}{dt^i} = k(t)s(t)$$

with $b_i = 0$, $i = n + 1, n + 2, \cdots, d$.

Next, assume that $k(t)$ is periodic of period T and that $a_d + k(t)b_d > 0$. Then, it is known that the set of fundamental solutions to the differential equation for r and the set of fundamental solutions to the differential equation for p are of the Floquet form [1, pp. 60–62]. In each case, the characteristic exponents are determined modulo $j\,2\pi/T$. Thus, with the preceding results, it is evident that the characteristic exponents associated with both fundamental sets are equal modulo $j\,2\pi/T$. It is also evident that, if p (or r) contains a nonzero term with the multiplicative factor $e^{\{\mathrm{Re}[\lambda]\}t}$, where λ is a characteristic exponent, then r (or p) contains a nonzero term with the same multiplicative factor.

Now, the system will be called stable if $r(t) \to 0$ as $t \to \infty$ when $s(t) = 0$ for $t \geqq t_0$, t_0 an arbitrary real constant, and when the ρ_i, $i = 0, 1, \cdots, d - 1$, are arbitrary complex constants. The system will be called unstable if it is not stable. The foregoing discussion justifies the statement: The system is stable if $p(t) \to 0$ as $t \to \infty$ when $s(t) = 0$ for $t \geqq t_0$, t_0 an arbitrary real constant, and when the ξ_i, $i = 0, 1, \cdots, d - 1$, are arbitrary complex constants, and the system is unstable if $p(t) \not\to 0$ as $t \to \infty$ when $s(t) = 0$ for $t \geqq t_0$, t_0 a real constant, and for some set of complex constants ξ_i, $i = 0, 1, \cdots, d - 1$.

Sandberg [2] has established a set of sufficiency conditions which guarantee that the system of Fig. 2 is stable when $k(t)$ is periodic. It is also

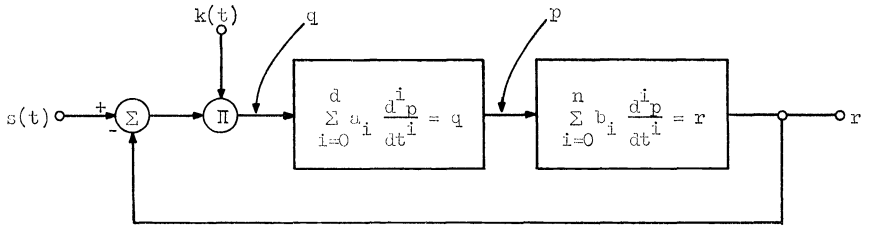


FIG. 2. *Time-varying control system*

of some interest to establish sufficiency conditions which guarantee that the system is unstable when $k(t)$ is periodic; such a set of conditions are presented in §2.

2. Principal result. Consider the differential equation

$$\sum_{i=0}^{d} [a_i + k(t)b_i] \frac{d^i p}{dt^i} = 0. \tag{1}$$

In all that follows it is assumed that:

(A1) the a_i and b_i, $i = 0, 1, \cdots, d$, are real constants with $a_d \neq 0$;

(A2) t is a real variable;

(A3) $k(t)$ is a real-valued, piecewise continuous, periodic function of period T, where T is a real, positive constant;

(A4) $[a_d + k(t)b_d] > 0$ for $t \in [0, T]$.

Let $\mathfrak{D} \subset \mathfrak{R}$, where $\mathfrak{R}$ denotes the real line, be the set of points at which $k(t)$ is discontinuous. Any complex-valued function which is $(d-1)$-times continuously differentiable on $\mathfrak{R}$ and satisfies (1) on $\mathfrak{R} - \mathfrak{D}$ is called a solution of (1).

In that which follows it is convenient to use the following definitions. Let

(D1) ν be a real variable,

(D2) $\omega_0 = 2\pi/T$,

(D3) k_0 be any constant such that $[a_d + k_0 b_d] > 0$,

(D4) $\hat{k}(t) = k(t) - k_0$,

(D5) $K = \sup_{t \in [\mathfrak{R} - \mathfrak{D}]} |\hat{k}(t)|$,

(D6) $$H(\omega) = \frac{\sum_{i=0}^{d} b_i (j\omega)^i}{\sum_{i=0}^{d} [a_i + k_0 b_i](j\omega)^i} \quad \text{with} \quad j = \sqrt{-1},$$

(D7) $$\Theta(\mathfrak{M}) = \sup_{|\nu| \leq \pi/T} \left\{ \sup_{m \in \mathfrak{M}} |H(m\omega_0 + \nu)|^2 k^2 + \frac{1}{T} \int_0^T |\hat{k}(t)|^2 \, dt \sum_{m \in [\mathfrak{J} - \mathfrak{M}]} |H(m\omega_0 + \nu)|^2 \right\},$$

where $\mathfrak{M}$ is a subset of the set of integers $\mathfrak{I}$ and the supremum over the empty set is interpreted as zero. The primary result reported in this paper is the following theorem.

THEOREM. *If there exists a* k_0 *such that*

(H1) $\sum_{i=0}^{d} [a_i + k_0 b_i] s^i \neq 0$ *for* $\mathrm{Re}[s] = 0$,

(H2) *one or more zeros of* $\sum_{i=0}^{d} [a_i + k_0 b_i] s^i$ *have positive real part, and*

(H3) *there exists an* $\mathfrak{S} \subseteq \mathfrak{I}$ *such that* $\Theta(\mathfrak{S}) < 1$,

then at least one solution of (1) *does not approach zero as* $t \to \infty$.

It is of interest to note that in the corresponding stability theorem, given by Sandberg, (H2) is replaced by:

(H2*) *no zeros of* $\sum_{i=0}^{d} [a_i + k_0 b_i] s^i$ *have positive real part,*

and the conclusion is replaced by:

then all solutions of (1) *approach zero as* $t \to \infty$.

The proof for the above instability theorem was derived, with only a few changes, from that given by Sandberg for his stability theorem.

Proof. Consider

$$\sum_{i=0}^{d} [a_i + k_0 b_i + \delta \hat{k}(t) b_i] \frac{d^i x}{dt^i} = 0$$

for $\delta \in [0, 1]$.

By (A4), (D3) and (D4), $[a_d + k_0 b_d + \delta \hat{k}(t) b_d] > 0$ for all $(t, \delta) \in [0, T] \times [0, 1]$. Thus, for any $\delta \in [0, 1]$ the solution space of (2) is spanned by a set of solutions of the Floquet form. Further, it is known [3, p. 21] that the characteristic exponents are continuous functions of δ in $[0, 1]$.

The proof will be by contradiction. Thus, suppose all solutions of (2) for $\delta = 1$ approach zerr as $t \to \infty$. Then, for $\delta = 1$ each of the characteristic exponents must have negative real part. By (H2), one or more of the characteristic exponents have positive real part for $\delta = 0$. Note that by (H1) no characteristic exponents have zero real part for $\delta = 0$. Therefore, a $\delta_0 \in (0, 1)$ must exist such that (2) possesses a solution of the form $y(t)e^{j\nu t}$, where ν is a real constant, $y(t)$ is a complex-valued function of period T, and $y(t) \not\equiv 0$ on $[0, T]$. Observe that, without loss of generality, it may be assumed that $|\nu| \leqq \pi/T$. The proof by contradiction will be accomplished by showing that $y(t) \equiv 0$ on $[0, T]$ for all $\delta_0 \in (0, 1)$ and $|\nu| \leqq \pi/T$, and, hence, that all solutions of (2) for $\delta = 1$ do not approach zero as $t \to \infty$.

Because $y(t)e^{j\nu t}$ is a solution of (2), it is $(d - 1)$-times continuously differentiable on $\mathfrak{R}$. Further, by (2) and (A4), it has a bounded, continuous dth derivative on $\mathfrak{R} - \mathfrak{D}$. This implies $y(t)$ is continuous and piecewise smooth. Hence, $y(t)$ possesses a Fourier series which converges pointwise, as well as in the mean. Thus, set

$$y(t) = \sum_{m \in \mathfrak{I}} y_m e^{jm\omega_0 t},$$

where

$$y_m = \frac{1}{T}\int_0^T y(t)e^{-jm\omega_0 t}\,dt.$$

Then, since $y(t)e^{j\nu t}$ is a solution of (2) for $\delta = \delta_0$,

$$x(t) = \sum_{m\in\mathcal{J}} y_m e^{j(m\omega_0+\nu)t} \tag{3}$$

and

$$\frac{d^i x}{dt^i} = \sum_{m\in\mathcal{J}} y_m[j(m\omega_0 + \nu)]^i e^{j(m\omega_0+\nu)t} \tag{4}$$

for $i = 1, 2, \cdots, d$.

Note. The differentiation term-by-term is valid, since $x(t) = y(t)e^{j\nu t}$ is a solution of (2) and any solution of (2) is $(d-1)$-times continuously differentiable on $\mathfrak{R}$ and has a bounded, continuous dth derivative on $\mathfrak{R} - \mathfrak{D}$.

Put

$$A(\omega) = \sum_{i=0}^{d} [a_i + k_0 b_i](j\omega)^i,$$

$$B(\omega) = \sum_{i=0}^{d} b_i(j\omega)^i,$$

$$u(t) = \sum_{i=0}^{d} [a_i + k_0 b_i]\frac{d^i x}{dt^i},$$

and

$$v(t) = \sum_{i=0}^{d} b_i \frac{d^i x}{dt^i}.$$

By (2),

$$u + \delta_0 \hat{k}(t)v = 0. \tag{5}$$

Also, by (3) and (4),

$$u(t) = \sum_{m\in\mathcal{J}} A(m\omega_0 + \nu)y_m e^{j(m\omega_0+\nu)t}$$

and

$$v(t) = \sum_{m\in\mathcal{J}} B(m\omega_0 + \nu)y_m e^{j(m\omega_0+\nu)t}.$$

Let

$$u_m = \frac{1}{T}\int_0^T u(t)e^{-j(m\omega_0+\nu)t}\,dt = A(m\omega_0 + \nu)y_m \tag{6}$$

and

$$v_m = \frac{1}{T}\int_0^T v(t)e^{-j(m\omega_0+\nu)t}\,dt = B(m\omega_0+\nu)y_m\,. \tag{7}$$

Then, examination of (6) and (7) discloses

$$v_m = u_m \frac{B(m\omega_0+\nu)}{A(m\omega_0+\nu)},$$

which by (D6) becomes

$$v_m = u_m H(m\omega_0+\nu).$$

Consider the following identity:

$$\begin{aligned}\frac{1}{T}\int_0^T |\,v(t)\,|^2\,dt &= \sum_{m\in\mathfrak{J}} |\,v_m\,|^2 \\ &= \sum_{m\in\mathfrak{M}} |\,u_m H(m\omega_0+\nu)\,|^2 + \sum_{m\in[\mathfrak{J}-\mathfrak{M}]} |\,u_m H(m\omega_0+\nu)\,|^2.\end{aligned} \tag{8}$$

Note that

$$\begin{aligned}\sum_{m\in\mathfrak{M}} |\,u_m H(m\omega_0+\nu)\,|^2 &\leqq \sup_{m\in\mathfrak{M}} |\,H(m\omega_0+\nu)\,|^2 \sum_{m\in\mathfrak{M}} |\,u_m\,|^2 \\ &\leqq \delta_0{}^2 \sup_{m\in\mathfrak{M}} |\,H(m\omega_0+\nu)\,|^2 K^2 \frac{1}{T}\int_0^T |\,v(t)\,|^2\,dt\end{aligned} \tag{9}$$

since, by (5) and (D5),

$$\begin{aligned}\sum_{m\in\mathfrak{M}} |\,u_m\,|^2 &\leqq \frac{1}{T}\int_0^T |\,u(t)\,|^2\,dt \\ &\leqq \delta_0{}^2 \frac{1}{T}\int_0^T |\,\hat{k}(t)v(t)\,|^2\,dt \\ &\leqq \delta_0{}^2 K^2 \frac{1}{T}\int_0^T |\,v(t)\,|^2\,dt.\end{aligned}$$

Let

$$\hat{k}(t) = \sum_{m\in\mathfrak{J}} \hat{k}_m e^{jm\omega_0 t},$$

where

$$\hat{k}_m = \frac{1}{T}\int_0^T \hat{k}(t)e^{-jm\omega_0 t}\,dt.$$

Then, using (5), it is easily shown that

$$u_m = -\delta_0 \sum_{l\in\mathfrak{J}} \hat{k}_{m-l}\, v_l$$

and

$$(10)\qquad \begin{aligned} |u_m|^2 &\leqq \delta_0^2\Big[\sum_{l\in\mathcal{G}} |\hat{k}_{m-l}|^2\Big]\Big[\sum_{l\in\mathcal{G}} |v_l|^2\Big] \\ &\leqq \delta_0^2 \frac{1}{T^2}\int_0^T |\hat{k}(t)|^2\,dt \int_0^T |v(t)|^2\,dt. \end{aligned}$$

Therefore,

$$(11)\qquad \begin{aligned} \sum_{m\in[\mathcal{G}-\mathfrak{M}]} |u_m H(m\omega_0+\nu)|^2 &\leqq \delta_0^2 \sum_{m\in[\mathcal{G}-\mathfrak{M}]} |H(m\omega_0+\nu)|^2 \\ &\quad\cdot \frac{1}{T^2}\int_0^T |\hat{k}(t)|^2\,dt \int_0^T |v(t)|^2\,dt. \end{aligned}$$

Now, by (8) with (9) and (11),

$$(12)\qquad \begin{aligned} \frac{1}{T}\int_0^T |v(t)|^2\,dt &\leqq \delta_0^2\Big\{ \sup_{m\in\mathfrak{M}} |H(m\omega_0+\nu)|^2K^2 \\ &\quad + \sum_{m\in[\mathcal{G}-\mathfrak{M}]} |H(m\omega_0+\nu)|^2 \frac{1}{T}\int_0^T |\hat{k}(t)|^2\,dt\Big\} \frac{1}{T}\int_0^T |v(t)|^2\,dt. \end{aligned}$$

Taking the supremum of the right-hand side of (12) with respect to $|\nu| \leqq \pi/T$ yields, upon taking note of (D7),

$$\int_0^T |v(t)|^2\,dt \leqq \delta_0^2\Theta(\mathfrak{M}) \int_0^T |v(t)|^2\,dt.$$

Now, by (H3) and the fact that $\delta_0 < 1$, $\delta_0^2\Theta(\mathcal{S}) < 1$ and, hence, $\int_0^T |v(t)|^2\,dt = 0$. It follows by (10) that $u_m = 0$ for all $m \in \mathcal{G}$. Since $A(\omega) \neq 0$ for all real ω by (H1), then, by (6), it is clear that $y_m = 0$ for all $m \in \mathcal{G}$. Since $y(t)e^{j\nu t}$ must have the properties of a solution of (2) for $\delta = \delta_0$, $y(t) \equiv 0$ on $[0, T]$ for all $\delta_0 \in (0, 1)$ and $|\nu| \leqq \pi/T$. This completes the proof of the theorem.

3. Auxiliary result. A result of somewhat greater practical value than the theorem will be presented as a corollary to the theorem. In the corollary, conditions predicting instability for a class of systems will be given. Now, consider the following corollary.

COROLLARY. *Let α and β be real constants such that $\alpha \leqq k(t) \leqq \beta$ and let $G(\omega) = \sum_{i=0}^{d} b_i(j\omega)^i / \sum_{i=0}^{d} a_i(j\omega)^i$. If*

(H4) $\sum_{i=0}^{d} [a_i + \frac{1}{2}(\alpha+\beta)b_i]s^i \neq 0$ *for* $\mathrm{Re}[s] = 0$,

(H5) *one or more zeros of* $\sum_{i=0}^{d} [a_i + \frac{1}{2}(\alpha+\beta)b_i]s^i$ *have positive real part, and*

(H6) $\inf_{\omega\in\mathfrak{R}} |G(\omega)^{-1} + \frac{1}{2}(\alpha+\beta)| > \frac{1}{2}(\beta-\alpha)$,

then at least one solution of (1) *does not approach zero as* $t \to \infty$.

Proof. Since assumption (A4) is now interpreted as valid for all $k(t)$ such that $\alpha \leqq k(t) \leqq \beta$, take $k_0 = \frac{1}{2}(\alpha + \beta)$. Note that the condition on k_0 in (D3) is satisfied. Now,

$$H(\omega) = \frac{1}{G(\omega)^{-1} + \frac{1}{2}(\alpha + \beta)};$$

therefore, (H6) implies

$$\sup_{\omega \in \mathcal{R}} | H(\omega) | \tfrac{1}{2}(\beta - \alpha) < 1.$$

Since

$$\sup_{\omega \in \mathcal{R}} | H(\omega) | = \sup_{|\nu| \leqq \pi | T} \left\{ \sup_{m \in \mathcal{J}} | H(m\omega_0 + \nu) | \right\}$$

and $K \leqq \frac{1}{2}(\beta - \alpha)$, it is clear that $\Theta(\mathcal{J}) < 1$ and (H3) is satisfied. Next, (H4) and (H5) imply (H1) and (H2), respectively. Hence, by the theorem, the corollary is proven valid.

In the corresponding corollary on stability, given by Sandberg, (H5) is replaced by:

(H5*) *no zeros of* $\sum_{i=0}^{d} [a_i + \frac{1}{2}(\alpha + \beta)b_i]s^i$ *have positive real part,*

and the conclusion is replaced by:

then all solutions of (1) *approach zero as* $t \to \infty$.

It is of interest to note that it has been shown [2] that selecting $k_0 = \frac{1}{2}(\alpha + \beta)$ is optimal (in a sense defined in [2]).

Lastly, it is obvious that (H6) implies the curve $G(\omega)^{-1}$ be bounded away from the disk, centered at the point $(-\frac{1}{2}(\alpha + \beta), 0)$ in the complex plane, of radius $\frac{1}{2}(\beta - \alpha)$. In particular, (H6) implies $G(\omega)^{-1} + \frac{1}{2}(\alpha + \beta) \neq 0$ for any $\omega \in \mathcal{R}$, and, hence, (H6) implies (H4).

4. Concluding remarks. Consider the system of Fig. 1 under the assumptions of §1. Then the corollary may be used to determine if the system of Fig. 1 is unstable.

Let Z denote the number of zeros of $\sum_{i=0}^{d} b_i s^i$ with positive real part and let $\hat{G}(s) = [\sum_{i=0}^{d} b_i s^i]/[\sum_{i=0}^{d} a_i s^i]$. Note that $\hat{G}(j\omega) = G(\omega)$. Then, the control system analyst interprets (H5) thus: The inverse Nyquist diagram [4, pp. 340–342] of the filter with rational transfer function $\hat{G}(s)$ must encircle the point $(-\frac{1}{2}(\alpha + \beta), 0)$, in the complex plane, $(1 - Z)$ or more times in the clockwise direction. Further, the inverse Nyquist diagram coincides with $G(\omega)^{-1}$ whenever $| G(\omega)^{-1} | < \infty$; therefore, (H4), (H5) and (H6) are satisfied if the inverse Nyquist diagram is bounded away from the disk and encircles the disk $1 - Z$ or more times in the clockwise direction.

Now, let P be the number of zeros of $\sum_{i=0}^{d} a_i s^i$ with positive real part.

Then, another interpretation of (H5) is: The Nyquist diagram [4, pp. 324–333] of the filter with rational transfer function $\hat{G}(s)$ must encircle the point $(-2/(\alpha + \beta), 0)$, in the complex plane, $1 - P$ or more times in the clockwise direction. Without loss of generality it may be assumed that β is positive. Now, for $\beta > 0$, it is easily shown that the statement:

> for $\alpha > 0$, the Nyquist diagram is bounded away from the disk, centered at $(-\frac{1}{2}(1/\alpha + 1/\beta), 0)$ in the complex plane, of radius $\frac{1}{2}(1/\alpha - 1/\beta)$,
>
> for $\alpha = 0$, the Nyquist diagram is bounded and contained in the half-plane for which the real part is greater than $-1/\beta$, and
>
> for $\alpha < 0$, the Nyquist diagram is contained in the interior of the disk, centered at $(-\frac{1}{2}(1/\alpha + 1/\beta), 0)$ in the complex plane, of radius $\frac{1}{2} \mid 1/\alpha - 1/\beta \mid$,

is equivalent to the statement:

> the inverse Nyquist diagram is bounded away from the disk, centered at $(-\frac{1}{2}(\alpha + \beta), 0)$ in the complex plane, of radius $\frac{1}{2}(\beta - \alpha)$.

Thus, if (H6) is satisfied, it is clear that the Nyquist diagram encircles the point $(-2/(\alpha + \beta), 0)$ zero times when $\alpha \leqq 0$. Hence, if (H5) is to be validated when $\alpha \leqq 0$, then P must be greater than or equal to one—the open loop system must be unstable. Note that using Sandberg's corollary on stability, it may be shown, for $\alpha \leqq 0$, that the system is stable if (H6) is satisfied and the open loop system is stable.

If the system is unstable, the number of characteristic exponents with positive real part can be determined easily. Recall that the theorem proof focused on showing that no nonzero solution of (2) has an imaginary characteristic exponent for $\delta \in [0, 1]$. Consequently, the number of characteristic exponents with positive real part is the same for all $\delta \in [0, 1]$ and, in particular, for $\delta = 0$ and $\delta = 1$. Now, the number of characteristic exponents with positive real part for $\delta = 0$ is equal to the number of zeros of $\sum_{i=0}^{d} [a_i + k_0 b_i] s^i$ with positive real part. The latter is easily determined.

Let $N^- > -Z$ be the number of times the inverse Nyquist diagram associated with $\hat{G}(s)$ encircles the point $(-k_0, 0)$ in the complex plane in the clockwise direction. Then, $\sum_{i=0}^{d} [a_i + k_0 b_i] s^i$ has $N^- + Z$ zeros with positive real part. Next, let $N^+ > -P$ be the number of times the Nyquist diagram associated with $\hat{G}(s)$ encircles the point $(-1/k_0, 0)$ in the clockwise direction. Then, $\sum_{i=0}^{d} [a_i + k_0 b_i] s^i$ has $N^+ + P$ zeros with positive real part.

Since the system equation (1) is the same as (2) with $\delta = 1$, it may be concluded by the above comments that the system has $N^- + Z = N^+ + P$ characteristic exponents with positive real part.

The last observation to be made is: The hypotheses of the corollary are independent of T and the proofs of the corollary and of its underlying theorem are independent of the value of T. This gives substance to the speculation that the conclusions of the corollary are valid when $k(t)$ is not periodic. This fact has been substantiated recently by Brockett and Lee [5]; that part of their results which corresponds to the results of the corollary is presented in Theorem 1 of their paper. It is interesting to note, however, that the proof given by Brockett and Lee is significantly longer than that presented in this paper, where the added assumption of periodicity is imposed. Further comparison of the results developed in this paper and those reported by Brockett and Lee are to be found in [6].

5. Acknowledgments. The author gratefully acknowledges the discussions with I. W. Sandberg which led to the realization that the proof for the instability results presented here could be built upon his work on stability.

REFERENCES

[1] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[2] I. W. Sandberg, *On the stability of solutions of linear differential equations with periodic coefficients*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 487–496.

[3] J. K. Hale, *Oscillations in Nonlinear Systems*, McGraw-Hill, New York, 1963.

[4] J. J. D'Azzo and C. H. Houpis, *Feedback Control System Analysis and Synthesis*, 2nd ed., McGraw-Hill, New York, 1960.

[5] R. W. Brockett and H. B. Lee, *Frequency-domain instability criteria for time-varying and nonlinear systems*, Proc. IEEE, 55 (1967), pp. 604–619.

[6] T. A. Bickart, *Periodically time-varying system: An instability criterion*, Proc. IEEE, to appear.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

ERRATA: ON THE PROBLEM OF APPROXIMATE SYNTHESIS OF OPTIMAL CONTROLS*

T. F. BRIDGLAND, JR.

Theorems 2 and 3 of the paper are probably not true without an additional hypothesis such as:

(*) For each $(t_0, x_0) \in B_\gamma$, $\lambda(\cdot)$ is continuous on $P_1(t_0, x_0)$.

Continuity is to be understood here in the sense implied by the supremum norm topology on the space of continuous functions from I_T into E^n. With (*) as an additional hypothesis the proofs of these theorems may be carried out as indicated in the paper. Incorporation of (*) in the statements of Theorems 2 and 3 has no bearing on the material of §4 since the corollary to Theorem 4 is an independent existence theorem whose proof clearly involves no assumption of continuity for $\lambda(\cdot)$.

The third sentence of the proof of Theorem 4 makes no sense and should be changed to read:

"Moreover, if $\zeta(x, p) \in \partial R(\psi(x), x)$ be a point nearest $g_p^\epsilon(\psi(x), x, p)$, then $\| g_p^\epsilon(\psi(x), x, p) - \zeta(x, p)\| < \epsilon$, and from this inequality and (vii) one obtains easily . . . ".

* This Journal, 5 (1967), pp. 326–344. Received by the editors October 9, 1967.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

THE SOLUTION OF SOME OPTIMAL CONTROL PROBLEMS*

V. F. DEM'YANOV†

Below we shall consider some (linear and nonlinear) problems in the theory of optimal control. One can turn one's attention to two aspects of similar problems. First of all, it is of interest to establish necessary conditions which must be satisfied by an optimal control. The classic results in this direction were obtained by L. S. Pontryagin and his collaborators [1].

References [2]–[5] have been devoted to the derivation of necessary conditions for an extremum in other problems. The second aspect which is both of theoretical and, primarily, of practical interest is the question of constructing algorithms for finding optimal controls (or at least controls that satisfy the necessary condition for a minimum—so-called "stationary controls").

Many works have been devoted to the solution of this problem. A good survey of some approaches is found in [6]. Some methods for solving time-optimal problems have been developed in [7]–[10]. Some other problems have been considered in [11]. An extensive bibliography on this problem is found in [6].

Below we shall set forth a general approach to the solution of a series of extremal problems. The methods we shall consider lead to stationary controls (to local extremum points), and if the functional being considered has no other local extrema, then these methods yield the possibility of solving the problem completely.

In this article only fixed time, free endpoint problems are considered. This approach makes it possible to solve such problems as well as others—provided that the indicated methods are suitably modified.

1. Classes of admissible controls. Given is the interval $[0, T]$, where $T > 0$ is fixed and $T < \infty$. Let us describe the classes of admissible controls U. The control, the r-dimensional vector function $u(t) = (u^1(t), \cdots, u^r(t))$, is square integrable on $[0, T]$ and satisfies one of the following constraints:

$$|u^i(t)| \leqq \alpha_i(t), \quad i = 1, \cdots, r, \quad t \in [0, T], \tag{1.1}$$

* Received by the editors April 19, 1967, and in revised form October 12, 1967. This translation into English has been prepared by Lucien W. Neustadt. The translation was supported in part through a grant-in-aid by the National Science Foundation.

† Computing Center of Leningrad State University, Leningrad, USSR. Presently visiting the Department of Electrical Engineering, University of Southern California Los Angeles, California 90007.

where the $\alpha_i(t)$, $i = 1, \cdots, r$, are nonnegative piecewise continuous functions bounded on $[0, T]$;

$$u^*(t)N(t)u(t) \leqq \beta(t), \quad t \in [0, T], \tag{1.2}$$

where $\beta(t) \geqq 0$ is a piecewise continuous function, bounded on $[0, T]$, $N(t)$ is a symmetric $r \times r$ matrix which is positive definite on $[0, T]$, with piecewise continuous elements that are bounded on $[0, T]$, and * indicates the transpose.

Special cases of (1.2) are (1.3)-(1.11):

$$u^*(t)N(t)u(t) \leqq 1, \quad t \in [0, T], \tag{1.3}$$

where N is a real, symmetric, positive definite $r \times r$ matrix;

$$\sum_{i=1}^{r} u^{i2}(t) \leqq 1, \quad t \in [0, T]; \tag{1.4}$$

$$| u^i(t) | = 1, \quad i = 1, \cdots, r, \quad t \in [0, T], \tag{1.5}$$

$$| u^i(t) | \in \{a_{i1}, a_{i2}, \cdots, a_{ip_i}\}, \quad t \in [0, T], \quad p_i \leqq p < \infty, \quad i = 1, 2, \cdots, r, \tag{1.6}$$

where the a_{ij} are given finite nonnegative numbers;

$$\int_0^T u^{i2}(t)\, dt \leqq c_i, \quad i = 1, \cdots, r, \quad 0 < c_i < \infty, \tag{1.7}$$

$$\int_0^T u^*(t)N(t)u(t)\, dt \leqq c, \quad 0 < c < \infty, \tag{1.8}$$

where $N(t)$ is a symmetric $r \times r$ matrix, positive definite on $[0, T]$, with elements integrable on $[0, T]$;

$$\int_0^T \sum_{i=1}^{r} u^{i2}(t)\, dt \leqq 1, \tag{1.9}$$

$$\int_0^T \sum_{r=1}^{r} | u^i(t) |^p\, dt \leqq 1, \quad 1 < p < \infty, \tag{1.10}$$

(here $| u^i(t) |$ is supposed to be summable on $[0, T]$ when raised to the pth power);

the control function $u(t)$ simultaneously satisfies the two constraints

$$| u^i(t) | \leqq 1, \quad t \in [0, T], \quad i = 1, \cdots, r, \qquad \int_0^T u^{i2}(t)\, dt \leqq c_i, \quad i = 1, \cdots, r, \quad 0 < c_i < \infty. \tag{1.11}$$

The classes of controls satisfying one of the constraints (1.1)–(1.11)

will be denoted, respectively, by U_1-U_{11}. We note that the classes U_1-U_4 and U_7-U_{11} are convex, bounded and weakly closed. The classes U_5 and U_6 are not convex, although they are bounded.

2. Systems of differential equations. On $[0, T]$, where $T > 0$ is fixed, we shall consider one of the four following systems of ordinary differential equations.

$$\frac{dX(t)}{dt} \equiv \dot{X}(t) = A(t)X(t) + \sum_{i=1}^{r} B_i(t)u^i(t) + F(t), \tag{2.1}$$

$$X(0) = X_0, \tag{2.2}$$

where $A(t)$ is an $n \times n$ matrix, $X(t)$, $F(t)$ and the $B_i(t)$ are n-vectors, $i = 1, \cdots, r$. The elements of the matrix A and the components of the vectors $F(t)$ and $B_i(t)$, $i = 1, \cdots, r$, are assumed to be real piecewise continuous functions bounded on $[0, T]$.

$$\dot{X}(t) = f(X(t), u(t), t), \tag{2.3}$$

$$X(0) = X_0, \tag{2.4}$$

where $X(t) = (x^1(t), \cdots, x^n(t))$ is an n-vector, $f = (f^1, \cdots, f^n)$ is an n-vector-valued function and the control $u(t) = (u^1(t), \cdots, u^r(t))$ is an r-vector-valued function (belonging to one of the above-described classes of admissible controls). The function in the right-hand side of (2.3) is assumed to be continuously differentiable with respect to x^i and u^j, $i = 1, \cdots, n, j = 1, \cdots, r$, in the region of admissible values of x^i and u^j defined by the class of controls U, the system (2.3) and the initial conditions (2.4), and to be continuous in t on $[0, T]$.

$$\dot{X}(t) = f(X(t), X(t - h_1), u(t), t), \tag{2.5}$$

$$X(t) = X_0(t) \quad \text{for} \quad t \in [-h_1, 0], \quad 0 < h_1 < \infty. \tag{2.6}$$

Here, $X(t) = (x^1(t), \cdots, x^n(t))$ is an n-vector-valued function, the control $u(t) = (u^1(t), \cdots, u^r(t))$ is an r-vector-valued function, subject to being chosen from the above-described class of controls U; $f(X, Y, u, t) = (f^i, \cdots, f^n)$ is a real n-vector-valued function continuous in x^i, y^j, u^k, t and continuously differentiable with respect to x^i, y^j, u^k, $i, j = 1, \cdots, n$, $k = 1, \cdots, r$, in the region of admissible values for x^i, y^j, u^k, defined by the class of controls U, the system (2.5) and the given continuous initial vector function $X_0(t)$.

$$\dot{X}(t) = f(X(t), X(t - h_1(t)), u(t), t), \tag{2.7}$$

$$X(t) = X_0(t) \quad \text{for} \quad t \in [-h(0), 0]. \tag{2.8}$$

The function $\nu(t) = t - h_1(t)$ is real, strictly increasing and continu-

ously differentiable on $[0, T]$,

$$0 < h_1(t) < \infty \quad \text{for} \quad t \in [0, T], \ \min_{t \in [0,T]} h_1(t) > 0.$$

Then there exists an inverse function $t = r_i(\nu)$, which is also a strictly increasing, continuously differentiable real-valued function on $[-h_1(0), T - h_1(T)]$; $X(t) = (x^1(t), \cdots, x^n(t))$ is an n-vector-valued function, the control $u(t) = (u^1, \cdots, u^r)$ is an r-vector-valued function subject to being chosen from one of the control classes U described in §1, $f(X, Y, u, t)$ is a real n-vector-valued function continuous with respect to x^i, y^j, u^k, t and continuously differentiable with respect to x^i, y^j, u^k, $i, j = 1, 2, \cdots, n$, $k = 1, \cdots, r$, in the region of admissible values for x^i, y^j, u^k, defined by the class U, the system (2.7) and the initial vector-valued function $X_0(t)$, which is given and continuous on $[-h(0), 0]$. The region of admissible values for t is the interval $[0, T]$.

It is clear that systems (2.1), (2.3) and (2.5) are special cases of system (2.7). We shall denote by $X(t, u)$ the solution of systems (2.1), (2.3), (2.5) and (2.7) for a given $u \in U$. We shall suppose that the function f is such that, for any $u \in U$, there exists a unique solution of the systems (2.1), (2.3), (2.5) and (2.7) for the initial conditions (2.2), (2.4), (2.6) and (2.8), respectively, on the entire interval $[0, T]$. For system (2.1), the solution $X(t, u)$ is given by the Cauchy formula

$$X(t) = Y(t)X_0 + \int_0^t \sum_{i=1}^{r} Y(t)Y^{-1}(\tau)B_i(\tau)u^i(\tau)\, d\tau + \int_0^t Y(t)Y^{-1}(\tau)F(\tau)\, d\tau, \tag{2.9}$$

where $Y(t)$ is the fundamental matrix of the homogeneous part of system (2.1):

$$\dot{Y}(t) = A(t)Y(t), \tag{2.10}$$

$$Y(0) = E. \tag{2.11}$$

The solution $X(t, u)$ of the system (2.3) satisfies the integral equation

$$X(t) = X_0 + \int_0^t f(X(\tau), u(\tau), \tau)\, d\tau, \tag{2.12}$$

and the respective solutions $X(t, u)$ of the systems (2.5), (2.7) satisfy the integral equations

$$X(t) = X_0 + \int_0^t f(X(\tau), X(\tau - h_1), u(\tau), \tau)\, d\tau, \tag{2.13}$$

$$X(t) = X_0 + \int_0^t f(X(\tau), X(\tau - h_1(\tau)), u(\tau), \tau)\, d\tau, \tag{2.14}$$

where $X_0 = X_0(0)$.

The attainable set $R(t)$ of systems (2.1), (2.3), (2.5) and (2.7) at the time $t \in [0, T]$ is defined as follows:

$$z \in R(t) \quad \text{if there exists a} \quad u \in U \quad \text{such that} \quad X(t, u) = z.$$

For the system (2.1), at any time t, the attainable set is convex, closed, and bounded.

3. The functionals to be considered. For the solutions of systems (2.1), (2.3), (2.5) and (2.7) we shall consider the following four functionals:

$$J(u) = F(X(T, u)), \tag{3.1}$$

where the function $F(X)$ is scalar-valued, real, and continuously differentiable on the attainable set of the system at the time T;

$$J(u) = \int_0^T g(X(t, u), u(t), t)\, dt, \tag{3.2}$$

where $g(X, u, t)$ is a scalar-valued function that is continuous on $[0, T]$ and continuously differentiable with respect to x^i and u^j, $i = 1, \cdots, n$, $j = 1, \cdots, r$, in the region of admissible values for x^i and u^j;

$$J(u) = \int_0^T g(X(t, u), u(t), t)\, dt + F(X(T, u)), \tag{3.3}$$

where the functions $g(X, u, t)$ and $F(X)$ are as described above;

$$J(u) = \int_0^T g(X(t, u), X(t - h_2(t), u), u(t), t)\, dt, \tag{3.4}$$

where $\nu(t) = t - h_2(t)$ is a real function strictly increasing and continuously differentiable on $[0, T]$, $0 \leqq h_2(t) < \infty$ for $t \in [0, T]$. Let $t = r_2(\nu)$ be the inverse function of $\nu(t)$ (it is also strictly increasing and continuously differentiable on $[-h_2(0), T - h_2(T)]$), $g(X, Y, u, t)$ is a real scalar-valued function which is continuous in t and continuously differentiable with respect to $x^i, y^j, u^k, i, j = 1, \cdots, n$, $k = 1, \cdots, r$, in the region of admissible values of x^i, y^j, and u^k.

For the systems (2.5) and (2.7), we shall assume that if $h_2(0) > h_1(0)$, then $X(t)$ is given and continuous on $[-h_2(0), -h_1(0)]$ (on $[-h_1(0), 0]$, $X(t)$ is given, respectively, by relations (2.6) and (2.8)).

Since

$$F(X(T, u)) = F(X_0) + \int_0^T \left(\frac{\partial F(X(t, u))}{\partial x}\right)^* \dot{X}(t, u)\, dt, \tag{3.5}$$

where $\dot{X}(t)$ is defined, respectively, by (2.1), (2.3), (2.5) or (2.7), then the functionals (3.1), (3.2) and (3.3) are special cases of the functional (3.4).

It is required to find a control $u \in U$, such that

$$J(u) = \min_{v \in U} J(v). \tag{3.6}$$

A control $u \in U$ which satisfies (3.6) will be called an optimal control.

Let U be a convex set of control functions. The functional $J(u)$ will be called convex if, for any $u_1, u_2 \in U$ and $\alpha \in [0, 1]$, we have

$$J(\alpha u_1 + (1 - \alpha)u_2) \leqq \alpha J(u_1) + (1 - \alpha)J(u_2). \tag{3.7}$$

In particular, if in (3.4) the integrand is convex with respect to X, Y, u, then, for the linear system (2.1), the functional $J(u)$ is a convex functional. This problem was considered in [12].

4. Necessary conditions for optimality. If the conditions that we imposed on the functional J and on the systems (2.1), (2.3), (2.5) and (2.7) are satisfied, then

$$\begin{aligned} J(v) &= J(u + (v - u)) \\ &= J(u) + \int_0^T G_u^*(\tau)(v(\tau) - u(\tau))\, d\tau + o(\| v - u \|), \end{aligned} \tag{4.1}$$

$$J(v) = J(u) + \int_0^T G^*_{u+\theta(v-u)}(\tau)(v(\tau) - u(\tau))\, d\tau, \qquad 0 \leqq \theta \leqq 1, \tag{4.2}$$

where $G_u(\tau) = (G_u^1(\tau), \cdots, G_u^r(\tau))$ is related to the gradient of the functional $J(u)$, evaluated at the point $v = u$ (this is an r-vector-valued function given on $[0, T]$), and $\| \cdot \|$ denotes the L^2 norm. Then we have the following theorem (see [13]).

THEOREM 1. *In order that the functional $J(u)$, given and bounded on U and having a continuous gradient (in the sense of Fréchet) thereon, achieve its minimum (relative to the controls of the class U) at a control u it is necessary (and, in the case when the functional $J(u)$ is convex, also sufficient) that*

$$\min_{v \in U} \int_0^T G_u^*(\tau)(v(\tau) - u(\tau))\, d\tau = 0. \tag{4.3}$$

Let us apply this theorem to our problems. Evaluating $G_u(\tau)$ and substituting into (4.3) (these calculations are carried out in [14], [15]), we obtain from Theorem 1 that the following theorem holds for the functional (3.4) and the system (2.7).

THEOREM 2. *In order that the control $u \in U$ achieve a minimum for the functional (3.4), which satisfies the previously indicated conditions, it is neces-*

sary (and, in the case when the functional (3.4) *is convex, also sufficient) that*

$$\min_{v\in U}\int_0^T \sum_{i=1}^{r}\left[\left(\frac{\partial f_u(\tau)}{\partial u^i}\right)^* \Psi_u(\tau) - \frac{\partial g_u(\tau)}{\partial u^i}\right](v^i(\tau) + u^i(\tau))\,d\tau = 0, \tag{4.4}$$

where

$$f_u(\tau) = f(X(\tau, u), X(\tau - h_1(\tau), u), u(\tau), \tau),$$

$$g_u(\tau) = g(X(\tau, u), X(\tau - h_2(\tau), u), u(\tau), \tau),$$

$$\frac{\partial \Psi_u(\tau)}{\partial \tau} = \begin{cases} -\left(\dfrac{\partial f_u(\tau)}{\partial X}\right)^* \Psi_u(\tau) - \left(\dfrac{\partial f_u(r_i(\tau))}{\partial y^i}\right)^* \dot{r}(\tau)\psi_u(r_1(\tau)) - C(\tau) \\ \qquad\qquad\qquad\qquad\qquad\qquad \textit{for} \quad \tau \in [0, T - h_1(T)], \\ -\left(\dfrac{\partial f_u(\tau)}{\partial X}\right)^* \Psi_u(\tau) - C(\tau) \quad \textit{for} \quad \tau \in [0, T - h_1(T), T], \end{cases} \tag{4.5}$$

$$\Psi_u(T) = 0, \tag{4.6}$$

$$\frac{\partial f}{\partial X} = \begin{pmatrix} \dfrac{\partial f^1}{\partial x^1} & \cdots & \dfrac{\partial f^1}{\partial x^n} \\ \cdot & \cdots & \cdot \\ \dfrac{\partial f^n}{\partial x^1} & \cdots & \dfrac{\partial f^n}{\partial x^n} \end{pmatrix}, \qquad \frac{\partial f}{\partial Y} = \begin{pmatrix} \dfrac{\partial f^1}{\partial y^1} & \cdots & \dfrac{\partial f^1}{\partial y^n} \\ \cdot & \cdots & \cdot \\ \dfrac{\partial f^n}{\partial y^1} & \cdots & \dfrac{\partial f^n}{\partial y^n} \end{pmatrix},$$

$$C(t) = \begin{cases} \dfrac{\partial g_u(t)}{\partial X} + \dfrac{\partial g_u(r_2(t))}{\partial Y}\,\dot{r}_2(t) \quad \textit{for} \quad t \in [0, T - h_2(T)], \\ \dfrac{\partial g_u(t)}{\partial X} \qquad\qquad\qquad\qquad \textit{for} \quad t \in [T - h_2(T), T], \end{cases} \tag{4.7}$$

$$\frac{\partial f}{\partial u^i} = \left(\frac{\partial f^1}{\partial u^i}, \cdots, \frac{\partial f^n}{\partial u^i}\right), \qquad \frac{\partial g}{\partial X} = \left(\frac{\partial g}{\partial x^1}, \cdots, \frac{\partial g}{\partial x^n}\right),$$

$$\frac{\partial g}{\partial Y} = \left(\frac{\partial g}{\partial y^1}, \cdots, \frac{\partial g}{\partial y^n}\right).$$

If in (3.4), $h_2(t) \equiv h_2$, and the system of differential equations has the form (2.5), then in formula (4.4) we have

$$\frac{d\Psi_u(\tau)}{d\tau} = \begin{cases} -\left(\dfrac{\partial f_u(\tau)}{\partial X}\right)^* \Psi_u(\tau) - \left(\dfrac{\partial f_u(\tau + h_1)}{\partial Y}\right)^* \Psi_u(\tau + h_1) - C(\tau) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for} \quad \tau \in [0, T - h_1], \\ -\left(\dfrac{\partial f_u(\tau)}{\partial X}\right)^* \Psi_u(\tau) - C(\tau) \qquad \text{for} \quad \tau \in [T - h_1, T], \end{cases} \tag{4.8}$$

$$\Psi_u(T) = 0, \tag{4.9}$$

where

$$C(\tau) = \begin{cases} \dfrac{\partial g_u(\tau)}{\partial X} + \dfrac{\partial g_u(\tau + h_2)}{\partial Y} & \text{for } \tau \in [0, T - h_2], \\ \dfrac{\partial g_u(\tau)}{\partial X} & \text{for } \tau \in [T - h_2, T]. \end{cases} \tag{4.10}$$

For the class of controls U_1-U_6 we obtain, from (4.4), that the following theorem holds.

THEOREM 3. *In order that the control $u \in U$ achieve a minimum for the functional* (3.4), *which satisfies the previously indicated conditions (for the classes of controls of the type of* (1.1)–(1.6)), *it is necessary that, for almost all $t \in [0, T]$, the following relation be satisfied:*

$$\min_{v(t) \in U} \sum_{i=1}^{r} \left(\left(\frac{\partial f_u(t)}{\partial u^i} \right)^* \Psi_u(t) + \frac{\partial g_u(t)}{\partial u^i} \right) (v^i(t) - u^i(t)) = 0. \tag{4.11}$$

For the functional (3.2) and the system (2.3), we obtain Theorem 4 from Theorem 2.

THEOREM 4. *In order that the control $u \in U$ achieve for the functional $J(u)$, which satisfies the previously indicated conditions, the least possible value relative to the controls of the class U, it is necessary that*

$$\min_{v \in U} \int_0^T \sum_{i=1}^{r} \left[\left(\frac{\partial f_u(\tau)}{\partial u^i} \right)^* \Psi_u(\tau) + \frac{\partial g_u(\tau)}{\partial u^i} \right] (v^i(\tau) - u^i(\tau))\, d\tau = 0, \tag{4.12}$$

where

$$\dot{\Psi}_u(\tau) = -\left(\frac{\partial f_u(\tau)}{\partial X} \right)^* \Psi_u(\tau) - \frac{\partial g_u(\tau)}{\partial X}, \tag{4.13}$$

$$\Psi_u(T) = 0, \tag{4.14}$$

$$\frac{\partial g}{\partial X} = \left(\frac{\partial g}{\partial x^1}, \cdots, \frac{\partial g}{\partial x^n} \right), \qquad \frac{\partial f}{\partial X} = \begin{pmatrix} \dfrac{\partial f^1}{\partial x^1} & \cdots & \dfrac{\partial f^1}{\partial x^n} \\ \cdot & \cdots & \cdot \\ \dfrac{\partial f^n}{\partial x^1} & \cdots & \dfrac{\partial f^n}{\partial x^n} \end{pmatrix}, \tag{4.15}$$

$$f_u(\tau) = f(X(\tau, u), u(\tau), \tau),$$

$$g_u(\tau) = g(X(\tau, u), u(\tau), \tau).$$

The function $\Psi_u(\tau)$ in Theorem 4 to within a sign change coincides with the function $\Psi(\tau)$ in [1, pp. 23–25].

For the functional (3.2), the system (2.3) and the classes of controls (1.1)–(1.4), the necessary condition (4.12) is a linearization of the Pontryagin maximum principle.

Yu. F. Kazarinov showed that, for the classes of controls U_7, U_8, U_9, for

the functional (3.2) and the system (2.3), the maximum principle has the form:

$$\min_{v\in U}\int_0^T [f(X(\tau,u),v(\tau),\tau)\Psi_u(\tau)+g(X(\tau,u),v(\tau),\tau)]\,d\tau = \int_0^T [f_u{}^*(\tau)\Psi_u(\tau)+g_u(\tau)]\,d\tau, \tag{4.16}$$

where $\Psi(\tau)$ satisfies the system of differential equations (4.13) with initial conditions (4.14). Here, we shall not concern ourselves with the proof of condition (4.16).

We can also assume that, for the classes U_7, U_8, U_9 and for the functional (3.4) and the system (2.7), the necessary condition of the maximum principle holds:

$$\min_{v\in U}\int_0^T [f^*(X_u(\tau),Y_u(\tau),v(\tau),\tau)\Psi_u(\tau)+g(X_u(\tau),Y_u(\tau),v(\tau),\tau)]\,d\tau = \int_0^T [f(X_u(\tau),Y_u(\tau),u(\tau),\tau)\Psi_u(\tau)+g(X_u(\tau),Y_u(\tau),u(\tau),\tau)]\,d\tau, \tag{4.17}$$

where $\Psi_u(\tau)$ satisfies the system (4.5) with initial conditions (4.6).

A control $u \in U$ that satisfies (4.4) will be called a *stationary control.*

For the linear system (2.1) and the functional (3.2) we have that

$$\frac{\partial f_u(\tau)}{\partial X} = A(\tau), \quad \dot{\Psi}_u(\tau) = -A^*(\tau)\Psi_u(\tau) - \frac{\partial g_u(\tau)}{\partial X},$$
$$\Psi_u(T) = 0.$$

From this, after having carried out the corresponding manipulations, we obtain that for the linear system (2.1) and the functional (3.2) the following theorem holds.

THEOREM 5. *In order that the control $u \in U$ achieve for the functional (3.2) the least possible value relative to the controls of the class U, it is necessary (and, in the case when the function $g(X, u)$ is convex with respect to X and u, also sufficient) that*

$$\min_{v\in U}\int_0^T \sum_{i=1}^{r}\left\{[\omega(t)]^*[Y^{-1}(\tau)B_i(\tau)] + \frac{\partial g_u(\tau)}{\partial u^i}\right\}\cdot(v^i(\tau)-u^i(\tau))\,d\tau = 0, \tag{4.18}$$

where

$$\omega(t) = \int_t^T Y^*(\tau)\frac{\partial g_u(\tau)}{\partial X}\,d\tau, \tag{4.19}$$

$$\dot{Y}(\tau) = A(\tau)Y(\tau), \tag{4.20}$$

$$Y(0) = 0. \tag{4.21}$$

5. Methods of successive approximations for finding stationary points. First we shall consider methods for the classes of controls U_1-U_4, U_7-U_{11} (these classes are convex, weakly closed and bounded).

Let $u \in U$; by $v[u]$ we shall denote any function $v(t) \in U$ that satisfies the condition

$$\min_{\omega \in U} \int_0^T \sum_{i=1}^r \left[\left(\frac{\partial f_u(\tau)}{\partial u^i} \right)^* \Psi_u(\tau) + \frac{\partial g_u(\tau)}{\partial u^i} \right] \omega^i(\tau)\, d\tau = \int_0^T \sum_{i=1}^r \left[\left(\frac{\partial f_u(\tau)}{\partial u^i} \right)^* \Psi_u(\tau) + \frac{\partial g_u(\tau)}{\partial u^i} \right] v^i(\tau)\, d\tau. \tag{5.1}$$

To find $v[u]$, it is necessary to minimize a linear functional.

To find the stationary controls, one can employ the algorithms set forth in [12], [15]–[18]. Let us describe some of them for the case under consideration in this section.

5.1. The conditional gradient method. For the first approximation we shall choose any admissible $u_1 \in U$. Suppose that $u_k \in U$ has been found. We shall find $v[u_k] = v_k \in U$. Let us form the linear combination

$$u_{k\alpha} = \alpha u_k + (1 - \alpha) v_k, \quad \alpha \in [0, 1], \quad u_{k\alpha} \in U,$$

and let us find $\alpha_k \in [0, 1]$ such that

$$J(u_{k\alpha_k}) = \min_{\alpha \in [0,1]} J(u_{k\alpha}).$$

We set $u_{k+1} = u_{k\alpha_k}$. It is clear that $J(u_{k+1}) \leqq J(u_k)$. The sequences thus constructed,

$$u_1, u_2, \cdots, u_k \in U, \tag{5.2}$$

$$v_1, v_2, \cdots, v_k \in U, \tag{5.3}$$

are such that

$$J(u_1) \geqq J(u_2) \geqq \cdots \geqq \cdots. \tag{5.4}$$

Since the functional $J(u)$ is bounded from below on U, the following limit exists:

$$\lim_{k \to \infty} J(u_k) = J^* > -\infty, \quad J(u_k) \geqq J^*. \tag{5.5}$$

The sequence of controls (5.2) (see [15], [17]) converges to a stationary control in the sense of the following theorem.

THEOREM 6.

$$\lim_{k \to \infty} \min_{w \in U} \int_0^T \sum_{i=1}^r \left[\left(\frac{\partial f_{u_k}(\tau)}{\partial u^i} \right)^* \Psi_{u_k}(\tau) + \frac{\partial g_{u_k}(\tau)}{\partial u^i} \right] \cdot (w^i(\tau) - u_k{}^i(\tau))\, d\tau = 0. \tag{5.6}$$

We note that for the case where the functional has the form

$$J(u) = \int_0^T g(X(t))\,dt, \tag{5.7}$$

then, for the linear system (2.1), there is no need to deal with the linear combinations $u_{k\alpha}$, but rather it is sufficient to consider the functional (5.7) only on the combination of solutions

$$X_k(t) = X(t, u_k) \quad \text{and} \quad X(t, v_k) = \bar{X}_k(t),$$

and to try to find $\alpha_k \in [0, T]$ such that

$$J(u_{k\alpha_k}) = \int_0^T g(\alpha_k X_k(t) + (1 - \alpha_k)\bar{X}_k(t))\,dt$$

$$= \min_{\alpha \in [0,1]} \int_0^T g(\alpha X_k(t) + (1 - \alpha)\bar{X}_k(t))\,dt.$$

That is, it is not necessary to repeatedly integrate the system of differential equations (2.1).

If the functional $J(u)$ is a terminal functional of the form of (3.1), then, in the linear system (2.1), to find α_k it suffices to consider the line segment $[X(T, u_k), X(T, v_k)]$ and to find $\alpha_k \in [0, 1]$ such that

$$g(\alpha_k X(T, u_k) + (1 - \alpha_k)X(T, v_k)) = \min_{\alpha \in [0,1]} g(\alpha X(T, u_k) + (1 - \alpha)X(T, v_k)).$$

In these cases (when the system is linear and the functional has the form of (3.1) or (5.7)), the application of the conditional gradient method is especially effective.

5.2. Projective methods. Let $u \in U$, where U is one of the classes U_1–U_4, U_7–U_{11}. Let us form the linear combination of functions

$$u_\alpha(t) = u(t) - \alpha G_u(t), \quad \alpha \in [0, \infty),$$

where

$$G_u(t) = (G_u^1(t), \cdots, G_u^r(t))$$

is the gradient of the functional $J(u)$.

Let us find $w_\alpha(u) \in U$ such that

$$\int_0^T (w_\alpha(t) - u_\alpha(t))^2\,dt = \min_{v \in U} \int_0^T (v(t) - u_\alpha(t))^2\,dt. \tag{5.8}$$

The control $w_\alpha(u)$ is the projection of the function u_α onto U. There is no difficulty in finding w_α for U_1–U_{10}. In the case of U_{11}, this problem can also be solved using more complicated techniques.

For the first approximation, we choose any $u_1 \in U$. Suppose that u_k has been found. Let us form $u_{k\alpha}$ and find $\alpha_k \in [0, \infty)$ such that

$$J(w_{k\alpha_k}) = \min_{\alpha \in [0,\infty]} J(w_{k\alpha}). \tag{5.9}$$

Now we set $u_{k+1} = w_{k\alpha_k}$.

It is clear that $J(u_{k+1}) \leqq J(u_k)$. We continue in an analogous manner.

Note. The finding of the α_k that satisfies (5.9) may turn out to be laborious because $w_{k\alpha_k}$ enters into the functional through the system of differential equations. One can at each step find α_k without solving the system of differential equations. One can show that, for the sequence of controls $\{u_k\}$ thus constructed, Theorem 6 also holds.

Other projective methods yield the following modification of this second method (see [16], [17]).

By $w_{k\alpha\beta}$ we denote the control

$$w_{k\alpha\beta} = \beta u_k + (1 - \beta) w_{k\alpha}, \quad \beta \in [0, 1], \quad w_{k\alpha\beta} \in U.$$

Let us find $\beta_k(\alpha) \in [0, 1]$ such that

$$J(w_{k\alpha\beta_k}) = \min_{\beta \in [0,1]} J(w_{k\alpha\beta}).$$

Now we can find u_{k+1} by using one of the following procedures.

Method 2a. Let us choose a fixed (for all k) $\alpha \in (0, \infty)$, and let us set $u_{k+1} = w_{k\alpha\beta_k}$. As $\alpha \to \infty$, Method 2a, generally speaking, turns into the conditional gradient method, for $w_{k\alpha}$, roughly speaking, becomes one of the controls $v[u_k]$ as $\alpha \to \infty$. This question was discussed in detail in [16]. This method is similar to the conditional gradient method, except that we have $w_{k\alpha}$ instead of $v[u_k]$.

Method 2b. Let us find $\alpha_k \in [0, \infty)$ such that

$$J(w_{k\alpha_k\beta_k}) = \min_{\alpha \in [0,\infty]} J(w_{k\alpha\beta_k}),$$

and let us set $u_{k+1} = w_{k\alpha_k\beta_k}$. Here, instead of "a linear convex combination" of the control u_k and the control $w_{k\alpha}$, we consider the entire projection of the "ray" $u_{k\alpha} = u_k - \alpha G_{u_k}$. One step of Method 2b is more efficient, but more laborious than one step of Method 2a.

Method 2c. We find the coefficient α_k from the condition

$$J(w_{k\alpha_k}) = \min_{\alpha \in [0,\infty)} J(w_{k\alpha}),$$

and set $u_{k+1} = w_{k\alpha_k\beta_k}$. First of all, applying Method 2b, we find the best control on the projection, and then we construct a convex combination of u_k and $w_{k\alpha_k}$, and then apply Method 2a.

Method 2d. Let us find

$$J(w_{k\alpha_k\beta_k}) = \min_{\substack{\alpha\in[0,\infty)\\ \beta\in[0,1]}} J(w_{k\alpha\beta}),$$

and set $u_{k+1} = w_{k\alpha_k\beta_k}$. One step of this method is more efficient, but also more laborious, than is a step in any of the previous methods.

Methods 2a, 2b, 2c and 2d lead to a stationary control in the sense of Theorem 6.

We note that the projective method can also be applied to the classes of controls U_5 and U_6. For these classes (as well as for the classes U_1–U_4) one can also employ yet another method. Let us illustrate it with the class U_5 as an example.

Suppose we have found $u_k \in U_5$; u_k is not a stationary control. Let us find the set $\Omega_k \subset [0, T]$, where the necessary condition (4.8) is violated (the set Ω_k is measurable), i.e.,

$$\Omega_k = \{t \mid t \in [0, T], \quad \chi_k(t) < 0\},$$

where

$$\chi_k(t) = \min_{v\in U_5} G^*_{u_k}(t)(v - u_k(t)).$$

Let us construct the sequence of sets $\omega_1^k, \omega_2^k, \cdots$, where $\omega_1^k = \Omega_k$, $\omega_i^k = \omega_{i+1}^k \cup \bar{\omega}_{i+1}^k$ for $i > 1$, meas ω_{i+1}^k = meas $\bar{\omega}_{i+1}^k = \frac{1}{2}$ meas ω_i^k; meas $[\omega_{i+1}^k \cap \bar{\omega}_{i+1}^k] = 0$; and, for every $t \in \omega_{i+1}^k$, $\chi_k(t) \leqq \inf_{\tau\in\bar{\omega}_{i+1}} \chi_k(t)$ for $i = 1, 2, \cdots$.

It is clear that $\omega_1^k \supset \omega_2^k \supset \cdots$, and that $\omega_i^k \subset \Omega_k$ for all i.

Then there exists an m_k such that if $i > m_k$, then for the control

$$u_{ki}(t) = \begin{cases} -u_k(t), & t \in \omega_i^k, \\ u_k(t), & t \notin \omega_i^k, \end{cases}$$

it turns out that $J(u_{ki}) < J(u_k)$.

Let us set $u_{k+1} = u_{ki_k}$, where

$$J(u_{ki_k}) = \min_{i>m_i} J(u_{ki}).$$

It is clear that $J(u_{k+1}) < J(u_k)$. We continue in a similar manner. The sequence of controls $\{u_k\}$, constructed in this fashion, also tends to a stationary control in the sense of Theorem 6.

6. Acknowledgment. The author would like to express his thanks to the referees and to L. W. Neustadt for their help in clarifying and improving certain parts of this article.

REFERENCES

[1] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[2] A. YA. DUBOVITSKIY AND A. A. MILYUTIN, *Extremum problems in the presence of constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453.

[3] B. N. PSHENICHNII, *Convex programming in normed spaces*, Cybernetics, 1 (1965), no. 5, pp. 46–57.

[4] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems*, this Journal, 4 (1966), pp. 505–527.

[5] V. F. DEM'YANOV AND A. M. RUBINOV, *On necessary conditions for an extremum*, Ekonomika i Matematicheskiye Metody, 2 (1966), pp. 406–416.

[6] F. M. KIRILLOVA, *Applications of functional analysis to the theory of optimal processes*, this Journal, 5 (1967), pp. 25–50.

[7] L. W. NEUSTADT, *Synthesizing time-optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–498.

[8] J. H. EATON, *An iterative solution to time-optimal control*, Ibid., 5 (1962), pp. 329–344.

[9] N. E. KIRIN, *On a numerical method in a linear time-optimal problem*, Metody Vychisl., 2 (1963), pp. 67–74.

[10] B. N. PSHENICHNII, *A numerical method for calculating the time-optimal control for linear systems*, U.S.S.R. Comput. Math. and Math. Phys., 4 (1964), pp. 71–82.

[11] I. A. KRYLOV AND F. P. CHERNOUS'KO, *On the method of successive approximations for the solution of optimal control problems*, Zh. Vychisl. Mat. i Mat. Fiz., 2 (1962), pp. 1132–1139.

[12] V. F. DEM'YANOV, *The construction of an integral-optimal programmed linear system*, J. Appl. Math. Mech., 27 (1963), pp. 829–836.

[13] V. F. DEM'YANOV AND A. M. RUBINOV, *Minimizing a smooth convex functional on a convex set*, this Journal, 5 (1967), pp. 280–294.

[14] V. F. DEM'YANOV, *On the solution of some nonlinear optimal control problems*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1966), pp. 218–228.

[15] ———, *A direct method of finding optimal control in a nonlinear automatic control system*, Applied Problems in Engineering Cybernetics, Sov. Radio, Moscow, 1966, pp. 317–331.

[16] ———, *On the minimization of functions on bounded sets*, Cybernetics, 1 (1965), no. 6, pp. 77–87.

[17] ———, *On finding optimal controls in some automatic control problems*, Vestnik Leningrad. Univ., 13 (1965), pp. 26–34.

[18] ———, *On constructing an optimal program in linear systems*, Automat. Remote Control, 25 (1964), pp. 1–21.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

MINIMIZATION OF FUNCTIONALS IN NORMED SPACES*

V. F. DEM'YANOV† AND A. M. RUBINOV‡

We examine in this paper a number of problems involving the minimization of functionals on normed spaces. In particular, we consider the minimization of a differentiable functional on a convex set, and the minimization of a directionally differentiable functional (for example, the minimization of the maximum deviation). For all the problems which will be considered, we shall give necessary conditions for an extremum and we shall discuss various algorithms for finding points which satisfy the necessary conditions. A number of applications will be indicated, mainly to optimal control problems.

We do not consider the question of existence of a minimum, because for most of the problems to be discussed this question can be resolved by means of a theorem which states that a weakly lower semicontinuous functional achieves a minimum on a weakly compact set.

1. A few useful results.

1.1. Sublinear functionals. A functional p defined on a normed space X is called sublinear if it has the following properties:

(i) subadditivity: if $x, y \in X$, then

$$p(x + y) \leqq p(x) + p(y);$$

(ii) positive homogeneity: if $\lambda > 0$, $x \in X$, then

$$p(\lambda x) = \lambda p(x);$$

(iii) continuity: if $x_n \to x$, then $p(x_n) \to p(x)$.

Since the functional p^2 is subadditive and positive homogeneous, we easily obtain that $p(0) = 0$ and $-p(-x) \leqq p(x)$. A functional p is continuous (and, consequently, sublinear) if and only if it is bounded, i.e.,

$$\sup_{\|x\| \leqq 1} | p(x) | = \| p \| < \infty.$$

* Received by the editors April 19, 1967, and in revised form November 10, 1967. This translation into English from the original Russian has been prepared by E. Polak.

The translation was supported in part by a grant-in-aid from the National Science Foundation.

† Department of Electrical Engineering, University of Southern California, University Park, Los Angeles, California 90007; on leave from Computing Center of Leningrad State University, Leningrad, USSR.

‡ Mathematics Institute of the Siberian Branch of the USSR Academy of Sciences, Novosibirsk 90, USSR.

We observe that if p is sublinear, then it is also convex.

A linear functional h will be called a support to a sublinear functional p if for all $x \in X$, $h(x) \leqq p(x)$. We shall denote by U_p the set of all supports to p. Finally, for any $x \in X$, we introduce the set

$$U_p^x = \{h \in U_p \mid h(x) = p(x)\}.$$

It follows from the Hahn-Banach theorem that for any $x \in X$, the set U_p^x (and therefore U_p) is not empty. We observe that the sets U_p and U_p^x are convex, w^*-closed and bounded. The following important relation holds:

$$p(x) = \max_{h \in U_p} h(x) \quad \text{for} \quad x \in X. \tag{1.1}$$

It follows from (1.1) that the functional p is completely defined by the set of its supports. Now, let U be a convex, w^*-closed and bounded set in X^*. For $x \in X$, let

$$p_u(x) = \max_{f \in U} f(x).$$

It can be shown that the set of supports to the functional p_u coincides with U. Furthermore, the following important result holds.

THEOREM 1.1. *The map $\phi: \phi(p) = U_p$ takes, one-to-one, the set of all sublinear functionals on X into the set of all convex, w^*-closed and bounded subsets of X^*.*

The inverse map is defined as follows: $\phi^{-1}(U) = p_u$. Furthermore,

(i) $\phi(p_1 + p_2) = \phi(p_1) + \phi(p_2)$, $\quad \phi(\lambda p) = \lambda \phi(p)$, $\quad \lambda > 0$,

(ii) if for all $x \in X$, $p_1(x) \geqq p_2(x)$, then $\phi(p_1) \supset \phi(p_2)$.

If $U_1 \supset U_2$, then for all $x \in X$,

$$[\phi^{-1}(U_1)](x) \geqq [\phi^{-1}(U_2)](x).$$

Remark. The study of sublinear functionals is not the aim of this paper. We note, however, that Theorem 1.1 holds also in a much stronger form. Consider the smallest linear spaces which contain the families of sets and functionals defined in the above theorem. One can, in a natural way, introduce a norm into these subspaces as well as a partial ordering. Extending the map ϕ to these spaces, it can be shown that it is an isometry which preserves the partial ordering.

We now give an important consequence of Theorem 1.1.

COROLLARY 1.1. *Let $x \in X$ be such that $p(x) = \sup_{\alpha \in A} p_\alpha(x)$, where p_α ($\alpha \in A$) is a sublinear functional. Then the set U_p coincides with the convex, w^*- closure of the sets U_{p_α}, $\alpha \in A$.*

We now give an important example of a sublinear functional. Let

$X = C(E)$ be the space of continuous functions, defined on a closed and bounded subset E of a finite-dimensional space; let $p(x) = \max_{t \in E_0} x(t)$, where E_0 is a closed subset of E.

Making use of Corollary 1.1, it is not difficult to show that U_p consists of all nonnegative measures μ, defined on the σ-algebra of measurable subsets E, such that $\mu(E_0) = 1$, $\mu(E \backslash E_0) = 0$.

We note that sublinear functionals were studied in detail in [1].

We shall call a functional q superlinear if $-q$ is sublinear.

Clearly, all the properties of superlinear functionals can be obtained from the corresponding properties of sublinear functionals. As an example of a superlinear functional on the space $C[0, T]$, we can take the functional

$$q(x) = \min_{t \in [0,T]} x(t) = -\max_{t \in [0,T]} (-x(t)).$$

1.2. Directional differentiability. Let f be a functional defined on an open set ξ in the space X. By the derivative of f at the point $x \in \xi$, evaluated on the element $u \in X$, we shall mean the quantity

$$f_x'(u) = \lim_{\alpha \to 0+} \frac{f(x + \alpha u) - f(x)}{\alpha}. \tag{1.2}$$

Clearly, when $f_x'(u)$ exists, so does $f_x'(\lambda u)$ with $\lambda > 0$, whereby $f_x'(\lambda u) = \lambda f_x'(u)$. We shall call the quantity

$$\frac{1}{\| u \|} f_x'(u) = f_x'\left(\frac{u}{\| u \|}\right)$$

the derivative of f at x in the direction u. We shall say that a functional f is directionally differentiable at a point x if the limit (1.2) exists for any $u \in X$. The directional differentiability of f at a point x means that there exists a functional f_x', defined on the space X, such that for any $u \in X$ and $\alpha > 0$ sufficiently small,

$$f(x + \alpha u) = f(x) + \alpha f_x'(u) + o_{x,u}(\alpha), \tag{1.3}$$

where

$$\lim_{\alpha \to 0+} \frac{o_{x,u}(\alpha)}{\alpha} = 0.$$

The functional f_x' will be called the directional derivative of the functional f at the point x, and (1.3) will be called the formula of finite increments. Clearly, f_x' is a positively homogeneous functional.

We now introduce one more definition which will be important later on. We shall say that a functional f is uniformly directionally differentiable at a point x, if for any $u \in X$ and any $\epsilon > 0$, there exist a $\delta_u > 0$ and an

$\alpha_u > 0$ such that for $\| v - u \| \leqq \delta_u$ and $0 \leqq \alpha \leqq \alpha_u$,

$$\frac{o_{x,v}(\alpha)}{\alpha} < \epsilon$$

is satisfied, where the functions $o_{x,v}(\alpha)$ are defined as in (1.3).

The following theorems hold (see [1]).

THEOREM 1.2. *A convex functional f defined on a normed space X is directionally differentiable at every point $x \in X$. Furthermore, f_x' is a positively homogeneous, subadditive functional and the functions $o_{x,u}(\alpha)$ are nonnegative for any x, $u \in X$.*

THEOREM 1.3. *A sublinear functional is uniformly directionally differentiable at any point x, its derivative p_x' is a sublinear functional, and, in addition,*

$$U_{p_x'} = U_p^x. \tag{1.4}$$

It follows from (1.1) and (1.4) that for $u \in X$,

$$p_x'(u) = \max_{h \in U_p^x} h(u).$$

Thus, the computation of the directional derivative of a sublinear functional is connected with the description of the sets U_p^x. Consider an example. Let $X = C(E)$ (where E is a closed, bounded set in R^n), $p(x) = \max_{t \in E} x(t)$. Then U_p^x consists of nonnegative measures μ such that $\mu(E) = 1$, $\mu(E \backslash E(x)) = 0$, where

$$E(x) = \{t \in E \mid x(t) = \max_{\tau \in E} x(\tau)\}.$$

Thus,

$$p_x'(u) = \max_{\mu \in U_p^x} \int_E u(t)\, d\mu.$$

It is easy to show that

$$\max_{\mu \in U_p^x} \int_E u(t)\, d\mu = \max_{t \in E(x)} u(t),$$

whence it follows that

$$p_x'(u) = \max_{t \in E(x)} u(t), \tag{1.5}$$

a formula which will be important to us later.

As customary, we shall say that a functional f is differentiable (Gateau differentiable) at a point x if f_x' is a linear functional. We shall then call the derivative f_x' the gradient of f at x, and we shall denote it by F_x. If f is differentiable at every point of a set Ω, then the gradient F can be considered as an operator from Ω into X^*.

It is also possible to define a directional derivative for an operator. Let X and Y be normed spaces, let ξ be an open set in X, and let A be an operator mapping ξ into Y. We shall say that A is directionally differentiable at a point $x \in \xi$ if there exists an operator A_x' mapping X into Y, such that for $u \in X$ and $\alpha > 0$ sufficiently small,

$$A(x + \alpha u) = Ax + \alpha A_x'(u) + o_{x,u}(\alpha),$$

where

$$\frac{o_{x,u}(\alpha)}{\alpha} \to 0 \quad \text{as} \quad \alpha \to 0+.$$

In computing the directional derivatives of functionals, it is often convenient to make use of the formula for the differentiation of a composite function. Let X and Y be normed spaces, let A be an operator mapping an open set $\xi \subset X$ into Y, and let g be a functional defined on Y. Consider the functional f defined for $x \in \xi$ by

$$f(x) = g(Ax).$$

Let $x \in \xi$. Then, if A is directionally differentiable at the point x, and g is uniformly directionally differentiable at the point Ax, and, in addition, the functional g'_{Ax} is continuous, then f is directionally differentiable at X, and for $u \in X$,

$$f_x'(u) = g'_{Ax}(A_x'u). \tag{1.6}$$

We now give a few examples which will be important later on.

Let ξ be an open set in the n-dimensional space R^n, and let E be a closed and bounded set in the m-dimensional space R^m. We are given a real-valued function $g(x, y)$ defined on $\xi \times E = \{(x, y) \mid x \in \xi, y \in E\}$, which is continuous in y and which, for any $x \in \xi$, $y \in E$, has a continuous, in y, partial derivative

$$\frac{\partial g}{\partial x} = \left(\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2}, \cdots, \frac{\partial g}{\partial x_n}\right).$$

Consider the functional f defined on the set ξ by the relation

$$f(x) = \max_{y \in E} g(x, y).$$

We shall show that f is directionally differentiable at any point $x \in \xi$ and we shall find f_x'. For this purpose we define an operator A from ξ into $C(E)$ as follows: if $x_0 \in \xi$, then $Ax_0 = z \in C(E)$, where $z(y) = g(x_0, y)$. It is easy to verify that A is differentiable at any point $x_0 \in \xi$ and that for any $u = (u_1, u_2, \cdots, u_n) \in R^n$,

$$(A'_{x_0}(u))(y) = \frac{\partial g}{\partial x}(x_0, y)(u) = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(x_0, y)u_i.$$

Making use of (1.5) and (1.6), we obtain that f_x' exists at any point $x \in \xi$ and that for any $u \in R^n$,

$$f_x'(u) = \max_{y \in E(x)} \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(x,y)u_i,$$

where

$$E(x) = \{y \in E \mid g(x, y) = \max_{t \in E} g(x, t)\}.$$

Reasoning in exactly the same manner, it is easy to show that the functional $f_1(x) = \min_{y \in E} g(x, y)$ is directionally differentiable and that

$$(f_1')_x(u) = \min_{y \in E_1(x)} \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(x, y)u_i,$$

where

$$E_1(x) = \{y \in E \mid g(x, y) = \min_{z \in E} g(x, z)\}.$$

Generally, if A is a differentiable operator mapping some normed space X into $C(E)$, and for $x \in X$,

$$f(x) = \max_{t \in E} Ax(t), \qquad (f_1(x) = \min_{t \in E} Ax(t)),$$

then f (f_1) is directionally differentiable at any point x and

$$f_x'(u) = \max_{t \in E(Ax)} (A_x'(u))(t), \qquad (f_1')_x(u) = \min_{t \in E_1(Ax)} (A_x'(u))(t),$$

where

$$E(Ax) = \{t \in E \mid Ax(t) = \max_{\tau \in E} Ax(\tau)\},$$

$$E_1(Ax) = \{t \in E \mid Ax(t) = \min_{\tau \in E} Ax(\tau)\}.$$

2. Necessary conditions for an extremum. In this section we shall give some necessary conditions for an extremum, which can be expressed in terms of cones of a special type. These cones are "linear approximations" to the sets in which we shall be interested in a neighborhood of points of interest. Recently there have been several papers devoted to necessary conditions (see, for example, [1], [2], [3]). In some of these papers the framework for obtaining necessary conditions is more general than ours. Our approach (see [4]), however, is considerably simpler, while at the same time it can be applied to many problems of practical importance. On the basis of such necessary conditions, it is possible to construct algorithms for solving these problems. In the present paper we shall only give necessary conditions for a minimum of a directionally differentiable functional.

Let X be a normed space and let Ω be a subset of X. As is usual, we shall denote the closure of Ω by $\bar{\Omega}$. We shall say that an element u is an admissible direction at the point $x \in \bar{\Omega}$, with respect to the set Ω, if there exists a number $\alpha_0 > 0$, depending on x and u, such that for $\alpha \in (0, \alpha_0)$, $x + \alpha u \in \Omega$. Clearly, the set of admissible directions is a cone, which we shall denote by $K_x(\Omega)$. It is easy to show that $K_x(\Omega) \subset C(\Omega - x)$ (where $C(\Omega - x)$ is the cone hull of the set $\Omega - x$). When Ω is convex and $x \in \Omega$, then $K_x(\Omega) = C(\Omega - x)$. Generally we shall require not the cone $K_x(\Omega)$ but its closure $\overline{K_x(\Omega)}$. Because of this, we give one more important example. Let $x \in X$ and let $\Omega = \{z \in X \mid \phi(z) \leqq \phi(x)\}$, where ϕ is a functional which is directionally differentiable at x. In this case,

$$\{u \in X \mid \phi_x'(u) \leqq 0\} \supset K_x(\Omega) \supset \{u \in X \mid \phi_x'(u) < 0\}.$$

If some natural conditions on ϕ_x' are satisfied, we can obtain that

$$\overline{K_x(\Omega)} = \{u \in X \mid \phi_x'(u) \leqq 0\}.$$

The cone $K_x(\Omega)$ of admissible directions at x, with respect to the set Ω, linearly approximates the set $\Omega - x$ in a neighborhood of the origin (or, equivalently, the cone $x + K_x(\Omega)$, with vertex at the point x, linearly approximates the set Ω in a neighborhood of x). An approximation by the cone $K_x(\Omega)$ happens to be adequate for a large class of sets, such as convex sets, sets having a nonempty interior (when the point x is in the closure of the interior), sets of the form $\phi(x) = C$, where ϕ is some functional of the "maximum type", etc.

However, for many sets Ω of practical importance, an approximation by the cone $K_x(\Omega)$ is found to be inadequate. Thus, for example, consider the case when $\Omega = \{z \in X \mid \phi(z) = \phi(x)\}$, where ϕ is a strictly convex functional. In this case $K_x(\Omega) = \{0\}$. This leads to the need to define an object which constitutes a more accurate approximation.

Let $\Omega \subset X$. The element $u \in X$ will be called an admissible, in the broad sense, direction at the point $x \in \bar{\Omega}$, with respect to the set Ω, if for any $\epsilon > 0$ there is an element u_ϵ, $\| u_\epsilon - u \| < \epsilon$, and a number α_ϵ, $0 < \alpha_\epsilon \leqq \epsilon$, such that $x + \alpha_\epsilon u_\epsilon \in \Omega$. The cone of admissible, in the broad sense, directions will be denoted by $M_x(\Omega)$. We observe that $M_x(\Omega)$ is a closed set and that $\overline{K_x(\Omega)} \subset M_x(\Omega)$. If Ω is a convex set, $x \in \Omega$, then $M_x(\Omega) = \overline{K_x(\Omega)} = \overline{C(\Omega - x)}$. If $\Omega = \{z \in X \mid \phi(z) \leqq \phi(x)\}$ and ϕ_x' is continuous, where ϕ is a functional directionally differentiable at x, then $M_x(\Omega) = \overline{K_x(\Omega)} = \{u \in X \mid \phi x'(u) \leqq 0\}$. We shall be particularly interested in the case where $\Omega = \{z \in X \mid \phi(z) = \phi(x)\}$, where ϕ is a functional which is uniformly directionally differentiable at x, and whose derivative ϕ_x' is a continuous functional. In that case,

$$M_x(\Omega) \subset \{u \in X \mid \phi_x'(u) = 0\}.$$

If ϕ_x' is a sublinear (superlinear) functional and there exists a $v \in X$ such that $\phi_x'(v) < 0$ ($\phi_x'(v) > 0$), then

$$M_x(\Omega) = \{u \in X \mid \phi_x'(u) = 0\}.$$

We now proceed with the formulation of necessary conditions.

THEOREM 2.1. *Suppose that the functional f achieves its minimum on the set Ω at the point y. If it is directionally differentiable at y, then* $\min_{u \in K_y(\Omega)} f_y'(u) = 0$.

THEOREM 2.2. *Suppose that the functional f achieves its minimum on the set Ω at the point y. If it is uniformly directionally differentiable at y and its derivative f_y' is continuous, then* $\min_{u \in M_y(\Omega)} f_y'(u) = 0$.

The following theorem shows when the above necessary conditions are also sufficient.

THEOREM 2.3. *Let f be a convex functional defined on X, let Ω be a subset of X, and suppose that there exists a point $y \in \Omega$ such that* $\min_{u \in K_y(\Omega)} f_y'(u) = 0$ ($\min_{u \in M_y(\Omega)} f_y'(u) = 0$). *Furthermore, suppose there exists a neighborhood $S_\epsilon(y) = \{x \mid \| x - y \| < \epsilon\}$ of y, such that $(y + \overline{K_y(\Omega)}) \cap \Omega \supset S_\epsilon(y) \cap \Omega$ $((y + M_y(\Omega)) \cap \Omega \supset S_\epsilon(y) \cap \Omega)$. Then f achieves a local minimum on Ω at y. If the stronger condition $y + \overline{K_y(\Omega)} \supset \Omega$ $(y + M_y(\Omega) \supset \Omega)$ is satisfied, then f achieves a global minimum on Ω at y.*

It is possible to examine certain important nonconvex problems by means of Theorems 2.1–2.3. These theorems can also be used to some extent in the study of certain optimal control problems, approximation theory problems, as well as in the study of certain nonlinear equations (see, for example, [5]). Of main interest, however, is the case when Ω is convex.

THEOREM 2.4. *Suppose that a functional f, defined on a normed space X, achieves its minimum on a convex set Ω at a point y and that it is directionally differentiable at this point. Then*

$$\min_{x \in \Omega} f_y'(x - y) = 0. \tag{2.1}$$

If, in addition, f is a convex functional, then (2.1) is a necessary and sufficient condition for f to have a global minimum on Ω at y.

We now present a few corollaries of Theorem 2.4.

COROLLARY 2.1. *Let f be Gateau differentiable at the point y, then*

$$\min_{x \in \Omega} F_y(x) = F_y(y), \tag{2.2}$$

where F_y is the gradient of f at y.

COROLLARY 2.2. *Let f be a sublinear functional. Then (2.1) is equivalent to the condition*

$$\min_{x \in \Omega} f_y'(x) = f(y).$$

COROLLARY 2.3. *Let A be an operator from the normed space X into $C(E)$, and let $f(x) = \max_{t\in E}(Ax)(t)$. Then condition* (2.1) *can be written as follows* (cf. (1.5), (1.6)):

$$\min_{x\in\Omega}\max_{t\in E(Ay)} A_y'(x-y)(t) = 0, \tag{2.3}$$

where

$$E(Ax) = \{t \in E \mid (Ax)(t) = \max_{\tau\in E}(Ax)(\tau)\}.$$

As an example, consider the following problem. Let U be a convex, bounded subset of $L_2^r(0, T)$, the space of r-dimensional real-valued functions which are square integrable on $[0, T]$. To each $u \in U$ there corresponds a continuous function $R(t, u)$, defined for $t \in [0, T]$. Furthermore, for any $v \in U$ there exists an r-dimensional vector-valued function $G(t, \tau, v)$, integrable with respect to τ on $[0, T]$, such that for $u \in U$,

$$\begin{aligned} &R(t, u) \\ &= R(t, v) + \int_0^T G^*(t, \tau, v)(u(\tau) - v(\tau))\, d\tau + w(t, u - v), \end{aligned} \tag{2.4}$$

where $w(t, u - v)/\| u - v \| \to 0$ as $\| u - v \| \to 0$, and * denotes transposition. It is required to find a $u_0 \in U$ such that

$$\max_{t\in[0,T]} R(t, u_0) = \min_{u\in U}\max_{t\in[0,T]} R(t, u).$$

To apply our necessary condition, we observe that the function $R(t, u)$ defines an operator A from $L_2^r(0, T)$ into $C([0, T])$. Furthermore, it follows from (2.4) that this operator is differentiable and that

$$A_v'(u - v)(t) = \int_0^T G^*(t, \tau, v)(u(\tau) - v(\tau))\, d\tau.$$

The relation (2.3) leads to the following necessary condition: if the point u_0 is a solution to our problem, then

$$\min_{u\in U}\max_{t\in E(Au_0)} \int_0^T G^*(t, \tau, u_0)(u(\tau) - u_0(\tau))\, d\tau = 0.$$

This necessary condition was derived in [6].

3. Algorithms for computing stationary points. Let Ω be a convex set in a normed space X, and let f be a directionally differentiable functional defined on X. Points $y \in \Omega$ at which the condition (2.1) is satisfied will be called stationary points of f on Ω. Observe that if f is a convex functional, then the stationary points are points of minimum of f on Ω. We shall now give a few algorithms for computing stationary points.

3.1. The conditional gradient method. Let Ω be a convex, weakly compact set, and let f be a Gateau differentiable functional whose gradient, as usual, we shall denote by F. In this case it is possible to make use of the necessary conditions in the form (2.2). To find stationary points (points satisfying (2.2)), one may adopt the following conditional gradient method.

1. Choose an arbitrary element $x_1 \in \Omega$.

2. Suppose that the element x_n, $n = 1, 2, \cdots$, has been found and that $F_{x_n}(x_n) > \min_{x\in\Omega} F_{x_n}(x)$. Then let

$$x_{n+1} = x_n + \alpha_n(\bar{x}_n - x_n), \tag{3.1}$$

where $\bar{x}_n$ is determined from the condition $F_{x_n}(\bar{x}_n) = \min_{x\in\Omega} F_{x_n}(x)$, and α_n is chosen according to the principle of steepest descent, i.e., it is such that

$$f(x_n + \alpha_n(\bar{x}_n - x_n)) = \min_{\alpha\in[0,1]} f(x_n + \alpha(\bar{x}_n - x_n)). \tag{3.2}$$

Without imposing any conditions on the gradient F, except for continuity, it is possible to prove convergence of this method only when Ω is compact.

THEOREM 3.1. *If Ω is a compact set and F is a continuous operator, then the limit points of the sequence* (3.1) *are stationary points.*

From now on we shall assume that the operator F satisfies on the set Ω the following Lipschitz condition: for x', $x'' \in \Omega$,

$$\| Fx' - Fx'' \| \leqq L \| x' - x'' \|. \tag{3.3}$$

In this case, we can apply not only the above described variant of the gradient method (which we shall call the first variant of the method), but also another variant, which we shall call the second variant, and which differs from the first only in the choice of α_n. In the second variant we again determine x_{n+1} according to the formula (3.1), with $\bar{x}_n$ chosen as before, but with α_n determined by

$$\alpha_n = \min\left\{1, \gamma_n \frac{F_{x_n}(x_n - \bar{x}_n)}{\| x_n - \bar{x}_n \|^2}\right\},$$

where $0 < \epsilon_1 \leqq \gamma_n \leqq (2 - \epsilon_2)/L$, $\epsilon_2 > 0$ (L is defined as in (3.3)). The conditional gradient method was used by many authors for solving specific extremal problems (see, for example, [7], [8]). The first variant of the method was discussed in a general form in [9], [10] and [11], the second variant in [12].

We now give theorems on the convergence of the first variant of the conditional gradient method.

THEOREM 3.2. *Suppose that* (3.3) *is satisfied. Then*

$$\lim_{n\to\infty} F_{x_n}(x_n - \bar{x}_n) = 0.$$

THEOREM 3.3. *Let f be a convex functional which achieves its minimum on Ω at the point y. Then*

$$0 \leqq f(x_n) - f(y) \leqq F_{x_n}(x_n - \bar{x}_n). \tag{3.4}$$

If, in particular, the conditions of Theorem 3.1 (Ω *is compact*) *or* 3.2 (F *satisfies a Lipschitz condition*) *are satisfied, then* $f(x_n) \to f(y)$.

The first variant can be modified as follows. Let $\min F_{x_n}(x - x_n) = -a_n$. Choose $\bar{x}_n$ to satisfy $-a_n \leqq F_{x_n}(\bar{x}_n - x_n) \leqq -a_n/\zeta$, where ζ is fixed, with $1 \leqq \zeta < \infty$. Now let $f(x_n) = b_n$, $\min_{\alpha \in [0,1]} f(\bar{x}_n + \alpha(x_n - \bar{x}_n)) = c_n$ and let $\alpha_n' \in [0, 1]$ be such that $f(\bar{x}_n + \alpha_n'(x_n - \bar{x}_n)) \leqq \lambda b_n + (1 - \lambda)c_n$ ($\lambda \in (0, 1)$ is fixed). As the $(n + 1)$th approximation, take the element

$$x_{n+1} = \bar{x}_n + \alpha_n'(x_n - \bar{x}_n).$$

THEOREM 3.4. *Suppose* (3.3) *is satisfied. Then*

$$\lim_{n \to \infty} \min_{x \in \Omega} F_{x_n}(x_n - x) = 0.$$

Theorem 3.4 shows that the approximate (within reasonable limits) determination of the elements $\bar{x}_n$ and the numbers α_n does not destroy the convergence of the scheme. If the space X is a Hilbert space, then the same holds for the computation of the gradient, namely, instead of the functional F_{x_n}, we can use a functional $\tilde{F}_{x_n}$ which satisfies

$$\frac{(F_{x_n}, \tilde{F}_{x_n})}{\| F_{x_n} \| \, \| \tilde{F}_{x_n} \|} \geqq a > 0 \quad \text{and} \quad 0 < m \leqq \frac{\| \tilde{F}_{x_n} \|}{\| F_{x_n} \|} \leqq M < \infty,$$

where $n = 1, 2, \cdots$, and a, m, M are fixed numbers.

It was shown in [11] that under certain restrictions on the functional, the convergence of the sequence $f(x_n)$ can be guaranteed to be at the rate of $o(1/n)$.

For the second variant of the method it is possible to prove theorems which are analogous to Theorems 3.2 and 3.3. Furthermore, if f is a convex functional, then $f(x_n) - \min_{x \in \Omega} f(x) = o(1/n)$. Under some rather strong assumptions on the functional f and the set Ω, it is possible to obtain convergence at the rate of a geometric progression.

3.2. The gradient projection method. Let H be a Hilbert space, and let Ω be a convex, closed, bounded set in H. Let $x \in H$. Then there exists a unique element $P_\Omega x \in \Omega$ such that $\| x - P_\Omega x \| = \min_{z \in \Omega} \| x - z \|$. The element $P_\Omega x$ is referred to as the projection of the point x on the set Ω. Properties of projections were studied in detail in [13]. Now let f be a differentiable functional, whose gradient F satisfies on Ω the Lipschitz condition (3.3). For $\alpha > 0$, let $w_\alpha(x) = P_\Omega(x - \alpha F_x)$. It can be shown that in the case under consideration a point $x \in \Omega$ is a stationary point (satisfies

condition (2.2)) if and only if for some $\alpha > 0$, $x = w_\alpha(x)$. Furthermore, if for at least one $\alpha_0 > 0$, $x = w_{\alpha_0}(x)$ $(x \neq w_{\alpha_0}(x))$, then for all $\alpha > 0$, $x = w_\alpha(x)$ $(x \neq w_\alpha(x))$. Making use of this fact, one can propose the following gradient projection method for the minimization of f on Ω.

1. Choose an arbitrary $x_1 \in \Omega$.

2. Suppose that the element x_n, $n = 1, 2, \cdots$, has been determined and that $x_n \neq w_\alpha(x_n)$ for any $\alpha > 0$. Then, let

$$x_{n+1} = x_n + \beta_n(w_{\alpha_n}(x_n) - x_n).$$

Different choices of α_n, β_n result in different variants of the gradient projection method. Some of these schemes are given in [13]. Observe that although all the results in [13] are formulated for finite-dimensional spaces, they remain valid in Hilbert space. We shall now describe the more important schemes given in [13].

I. $\alpha_n \in [\gamma_{1n}, \gamma_{2n}]$, where

$$\gamma_{1n} = \min\left\{\frac{\| y_n - x_n \|}{\| F_{x_n} \| \cos \phi_n}, \frac{1}{L}\epsilon' \cos \phi_n(1 - \sin \phi_n)\right\},$$

$$\gamma_{2n} = \min\left\{\frac{\| y_n - x_n \|}{\| F_{x_n} \| \cos \phi_n}, \frac{2}{L}(1 - \epsilon'') \cos \phi_n(1 - \sin \phi_n)\right\},$$

where y_n is an arbitrary point in Ω, such that $(y_n - x_n, F_{x_n}) < 0$,

$$\phi_n = \cos^{-1} \frac{(y_n - x_n, -F_{x_n})}{\| y_n - x_n \| \| F_{x_n} \|},$$

and L is determined by formula (3.3); ϵ', ϵ'' are fixed, with $0 < \epsilon', \epsilon'' \leqq 1$; $\beta_n = 1$.

II. α_n is chosen to satisfy

$$f(w_{\alpha_n}(x_n)) = \min_{0 \leqq \alpha < \infty} f(w_\alpha(x_n)), \quad \beta_n = 1.$$

III. $\alpha_n \geqq \alpha_0$, where α_0 is an arbitrary positive number, and

$$\beta_n = \min\left\{1, -\frac{(w_{\alpha_n}(x_n) - x_n, F_{x_n})}{L \| w_{\alpha_n}(x_n) - x_n \|^2}\right\}. \tag{3.5}$$

IV. $\alpha_n \geqq \alpha_0$, where α_0 is an arbitrary positive number and β_n is chosen to satisfy

$$f(x_n + \beta_n(w_{\alpha_n}(x_n) - x_n)) = \min_{\beta \in [0,1]} f(x_n + \beta(w_{\alpha_n}(x_n) - x_n)). \tag{3.6}$$

V. α_n is chosen to satisfy

$$f(x_n - \alpha_n F_{x_n}) = \min_{0 \leqq \alpha < \infty} f(x_n - \alpha F_{x_n}), \tag{3.7}$$

and β_n is chosen to satisfy (3.5).

VI. α_n is chosen according to (3.7) and β_n is chosen according to (3.6).

Apart from the ones above, there were described in [13] schemes which are based on the simultaneous minimization with respect to the parameters α, β. Observe that for $\alpha \to \infty$, the schemes III and IV, of the gradient projection method, "converge" to the schemes II and I of the conditional gradient method respectively. For each of the above stated variants of the gradient projection method, it is possible to prove a convergence theorem, analogous to Theorem 3.2. A scheme which chooses $\alpha_n \in [\epsilon_1, 2/(M + \epsilon_2)]$, where $\epsilon_1, \epsilon_2 > 0$, and $\beta_n = 1$, was studied in detail in [12].

Finally, we wish to point out that the schemes I and II above can also be used for solving minimization problems on certain nonconvex sets.

3.3. A generalized conditional gradient method. Let X be a Banach space, let Ω be a convex, weakly compact subset of X, let A be a completely continuous operator from X into the space $C(E)$, where E is a closed and bounded subset of R^n. We give below an algorithm for minimizing the functional $f(x) = \max_{t \in E}(Ax)(t)$ on the set Ω. This algorithm was described in detail in [6], [17], and is a generalization of the conditional gradient method.

Let ϵ, ρ be arbitrary positive numbers.

1. Let $x_1 \in \Omega$ be arbitrary.
2. Suppose x_n has been found, $n = 1, 2, \cdots$. Let

$$\delta_n = \min_{x \in \Omega} \max_{t \in E(Ax)} A'_{x_n}(x - x_n)(t),$$

where

$$E(Ax) = \{t \in E \mid Ax(t) = \max_{\tau \in E} Ax(\tau)\}.$$

If $\delta_n = 0$, then the necessary condition (2.3) is satisfied, and, consequently, x_n is a stationary point. Suppose $\delta_n < 0$. Then we find the least real k such that

$$\begin{aligned} \min_{x \in \Omega} \max_{t \in E_{n,\epsilon/2^k}} A'_{x_n}(x - x_n)(t) &\leqq \max_{t \in E_{n,\epsilon/2^k}} A'_{x_n}(\bar{x}_n - x_n)(t) \\ &\leqq \frac{-\rho}{2^k}, \end{aligned} \tag{3.8}$$

where

$$E_{n,\epsilon/2^k} = \left\{ t \in E \mid A(x_n)(t) - \max_{\tau \in E} A(x_n)(\tau) | < \frac{\epsilon}{2^k}. \right.$$

Since $\delta_n < 0$, there will exist the required k. Now let

$$x_{n+1} = x_n + \alpha_n(\bar{x}_n - x_n),$$

where $\bar{x}_n$ is chosen according to (3.8) and α_n is chosen to satisfy

$$\max_{t\in E} A(x_n + \alpha_n(\bar{x}_n - x_n))(t) = \min_{\alpha\in[0,1]} \max_{t\in E} A(x_n + \alpha(\bar{x}_n - x_n))(t).$$

It is possible to show that the sequence thus constructed converges to a stationary point, or, to be more precise, the following theorem holds.

THEOREM 3.5. *For any* $\epsilon > 0$,

$$\lim_{n\to\infty} \inf_{x\in\Omega} \sup_{t\in E_{n,\epsilon}} A'_{x_n}(x - x_n)(t) = 0,$$

where

$$E_{n,\epsilon} = \{t \in E \mid |Ax_n(t) - \max_{\tau\in E} Ax_n(\tau)| \leqq \epsilon\}.$$

This theorem is a strenthening of the convergence theorem in [6].

3.4. The approximating functional method. The above described methods belong to the class of methods of feasible directions. For the minimization of functionals of the form $p(Ax)$, it is possible to adapt a method based on the idea of approximating the given functional by more simple ones. In a somewhat different formulation the method we propose below was discussed in [14], [15], [16]. This method applies also to nonconvex sets. Its merit lies in the fact that it always leads to a point of global minimum, while its disadvantage lies in the difficulty of the auxiliary problem.

Let X, Y be Banach spaces, let Ω be a weakly compact subset of X, let A be a completely continuous operator from X into Y, and let p be a sublinear functional defined on Y. The approximating functional method consists of the following.

1. Let $x_1, x_2, \cdots, x_r$, $r \geqq 2$, be arbitrary points in Ω and let g_i be arbitrary functionals in $U_p{}^{Ax_i}$, $i = 1, 2, \cdots, r$. (For a definition of the sets $U_p{}^{Ax_i}$, see §1.1.)

2. Suppose that the points $x_1, x_2, \cdots, x_n$, $n \geqq r$, and functionals $g_i \in U_p{}^{Ax_i}$, $i = 1, 2, \cdots, n$, have already been found. The point x_{n+1} is then chosen to satisfy

$$\max_{i\leqq n} g_i(Ax_{n+1}) = \min_{x\in\Omega} \max_{i\leqq n} g_i(Ax).$$

For the functional g_{n+1} we take an arbitrary element in $U_p{}^{Ax_{n+1}}$. Observe that if $Y = C(E)$, $p(y) = \max_{t\in E} y(t)$, then the functional g_n can be chosen as follows: for $y \in C(E)$, $g_n(y) = y(t_n)$, where the point $t_n \in E$ is such that $Ax_n(t_n) = \max_{t\in E} Ax_n(t)$. The element x_{n+1} is then chosen

to satisfy

$$\max_{i \leqq n} (Ax_{n+1})(t_i) = \min_{x \in \Omega} \max_{i \leqq n} Ax(t_i).$$

We now let $\mu = \min_{x \in \Omega} p(Ax)$, $\mu_n = p(Ax_n) = \max_{i \leqq n} g_i(Ax_n)$,

$$\lambda_n = \max_{i \leqq n-1} g_i(Ax_n) = \min_{x \in \Omega} \max_{i \leqq n-1} g_i(Ax).$$

THEOREM 3.6. *If $x_{n+1} = x_n$, then x_n solves the problem. Otherwise, the derived extremum is achieved on the limit points of the sequence constructed, as above. Furthermore,*

$$\mu_n - \lambda_n \to 0 \quad \textit{and} \quad \lambda_n \leqq \mu \leqq \mu_n. \tag{3.9}$$

The relation (3.9) can be considered to be an a posteriori two-sided bound on the rate of convergence.

Remark. The problem of minimax in finite-dimensional spaces was considered in [17].

The above described methods can be applied for solving a number of problems of practical importance. Some of these were considered in [18] (a minimum time problem with a nonlinear plant) and in [19] (an optimal control problem with nonlinear plant and state space constraints).

REFERENCES

[1] A. YA. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems with constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453.

[2] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[3] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems, I. General theory*, this Journal, 4 (1966), pp. 505–525.

[4] V. F. DEM'YANOV AND A. M. RUBINOV, *On necessary conditions for an extremum*, Ekonomika i Matematicheskie Metody, 2 (1966), pp. 406–417.

[5] A. M. RUBINOV, *Necessary conditions for an extremum with applications to certain equations*, Dokl. Akad. Nauk SSSR, 169 (1966), pp. 533–535.

[6] V. F. DEM'YANOV, *On the minimization of the maximum deviation*, Vestnik Leningrad. Univ., 21 (1966), no. 7, pp. 21–28.

[7] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[8] V. F. DEM'YANOV, *On the construction of an optimal strategy for a linear system*, Avtomat. i Telemeh., 25 (1964), pp. 3–12.

[9] V. F. DEM'YANOV AND A. M. RUBINOV, *The minimization of a smooth convex functional on a convex set*, Vestnik Leningrad. Univ., 19 (1964), no. 19, pp. 5–17.

[10] ———, *On the problem of minimizing a smooth functional with convex constraints*, Dokl. Akad. Nauk SSSR, 160 (1965), pp. 15–17.

[11] A. M. RUBINOV, *On certain generalizations of the method of steepest descent*, Matematicheskoye Programirovanie, Moscow, 1966, pp. 90–105.

[12] E. S. LEVITIN AND B. T. POLYAK, *Methods for constrained minimization*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1966), pp. 787–823.

[13] V. F. DEM'YANOV, *The minimization of functions on convex sets*, Kibernetika, 1 (1965), no. 6, pp. 65–74.

[14] G. P. AKILOV AND A. M. RUBINOV, *The method of successive approximations for the determination of a best approximation polynomial*, Dokl. Akad. Nauk SSSR, 157 (1964), pp. 503–505.

[15] A. M. RUBINOV, *The minimization of a norm on a compact set*, Vestnik Leningrad. Univ., 20 (1965), no. 1, pp. 140–142.

[16] D. A. ZIYAUDINOVA AND A. M. RUBINOV, *The minimization of sublinear functionals on compact subsets of metrizable locally convex topological spaces*, Optimal'noye Planirovanie, issue 7, Novosibirsk, 1967.

[17] V. F. DEM'YANOV, *On the solution of certain minimax problems I*, Kibernetika, 2 (1966), no. 6, pp. 58–66.

[18] ———, *On a nonlinear extremum problem*, Zh. Vychisl. Mat. i. Mat. Fiz., 7 (1967), pp. 1060–1077.

[19] ———, *Optimization of nonlinear control systems with state space constraints*, Vestnik Leningrad. Univ., 21 (1966), no. 13, pp. 11–17.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

STABILITY CONDITIONS FOR SYSTEMS WITH MONOTONE AND SLOPE-RESTRICTED NONLINEARITIES*

G. ZAMES† AND P. L. FALB‡

1. Introduction. Consider the feedback system illustrated in Fig. 1, where **H** is a time invariant linear operator and **N** is a memoryless monotone (or odd monotone) nonlinearity. In recent years stability conditions for such a system have been derived in terms of the frequency response $H(j\omega)$ of **H**. Many of these conditions involve the use of multipliers and typically take the following form: If a multiplier $M(j\omega)$ can be found such that $\mathrm{Re}\{M(j\omega)H(j\omega)\} \geqq 0$ and certain auxilliary conditions depending on **N** are satisfied, then the system is stable. Here we derive more general results

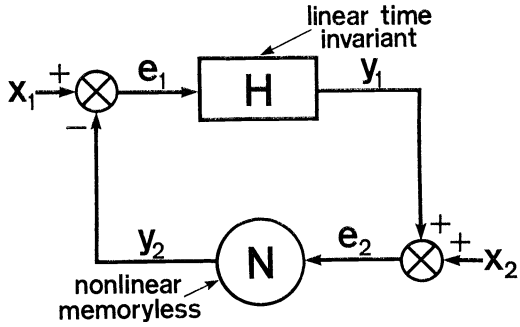


FIG. 1. *A feedback system*

than those previously available by broadening the class of allowable multipliers. Our derivation draws on the theory of positive operators on a Hilbert space and involves the factorization of a convolution operator into a product of two positive operators.

The problem of stability for feedback systems with a single nonlinearity was initially considered by Lur'e [1] and a well-known frequency response condition is Popov's theorem [2]. Lur'e and Popov assumed that the system was represented by a differential equation and that the nonlinearity was confined to the first and third quadrants. We pose our problem in the framework of operator equations and make quite different assumptions on **N**.

An operator approach to stability was used by Zames [3a], [3b], who

* Received by the editors March 7, 1967, and in revised form November 2, 1967.

† Guggenheim Fellow, NASA Electronics Research Center, Technology Square, Cambridge, Massachusetts 02939.

‡ Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912.

showed that a feedback system was stable if its "open-loop" operator could be factored into two positive definite parts. The class of "R.C." multipliers having the form

$$(1 - Z(s))^{-1} = \sum_{1}^{\infty} \frac{h_i}{s + a_i}, \qquad h_i > 0, \quad a_i > 0,$$

played a crucial role in [3a], [3b]. Closely related results were obtained, in the framework of Lyapunov theory, by Brockett and Willems [4], Yakubovicz [5], Narendra and Neuman [6], and Anderson [7] and, using operator methods, by Dewey and Jury [8] and Lee and Desoer [9]. A more general class of multipliers was introduced by O'Shea [10a] and used to prove stability by Zames and Falb [11]. (The original results of O'Shea were incorrectly proved.) In a subsequent report [10b], O'Shea used multipliers defined on $(-\infty, \infty)$ rather than on $[0, \infty)$; however, the results of O'Shea are valid for a very limited class of nonlinearities since his proof in effect involves the a priori assumption that the solutions are bounded. We require no such boundedness assumption here and our method of proof is novel. In point of fact, the main feature of this paper is a method for proving stability with multipliers defined on $(-\infty, \infty)$, which involves the factorization of a convolution operator on $L_2(-\infty, \infty)$ into a product of two operators, one having a kernel which vanishes on $(-\infty, 0]$ and the other having a kernel which vanishes on $[0, \infty)$.

In §2 we define the main problem and state our main result (Theorem 1). Then, in §3, we introduce our method for deriving stability conditions and prove a theorem which delineates the range of applicability of the method. As our approach involves the factorization of operators, we prove a relevant lemma in §4. Next, in §5, we show that the class of convolution operators considered forms a Banach algebra. In §6 we develop the requisite conditions ensuring the positivity of the operators involved in our proof of stability. We combine the results of §§3, 5 and 6 to give a proof of Theorem 1 in §7. Finally, we make some concluding remarks in §8.

2. The main problem and its solution. We consider the feedback equations (see Fig. 1):

$$e_1 = x_1 - y_2, \tag{1}$$

$$e_2 = x_2 + y_1, \tag{2}$$

$$y_1(t) = \mathbf{H}e_1(t) = \sum_{i=1}^{\infty} h_i e_1(t - \tau_i) + \int_0^{\infty} h(\tau) e_1(t - \tau)\, d\tau, \tag{3}$$

$$y_2(t) = \mathbf{N}e_2(t) = N(e_2(t)) \tag{4}$$

under the following assumptions.

A1. $N(\cdot)$ is a real-valued function on $(-\infty, \infty)$ with the following properties:

(i) $N(0) = 0$;

(ii) N is monotone nondecreasing, i.e., $(r - s)[N(r) - N(s)] \geqq 0$;

(iii) there is a constant $C > 0$ such that $|N(r)| \leqq C|r|$ for all r.

A2. $h(\cdot)$ is a (real-valued) element of $L_1[0, \infty)$, i.e., $\int_0^\infty |h(\tau)|\, d\tau < \infty$.

A3. $\{\tau_i\}$ is a sequence in $[0, \infty)$ and $\{h_i\}$ is a sequence in l_1, i.e., $\sum_{i=1}^\infty |h_i| < \infty$.

A4. x_1, x_2, y_1, y_2, e_1 and e_2 are real-valued functions on $(-\infty, \infty)$ and

(i) $\int_{-\infty}^t e_1(\tau)^2\, d\tau < \infty$, $\int_{-\infty}^t e_2(\tau)^2\, d\tau < \infty$ for all (finite) t;

(ii) $x_1(\cdot)$ and $x_2(\cdot)$ are elements of $L_2(-\infty, \infty)$, i.e.,

$$\int_{-\infty}^\infty x_1^2(\tau)\, d\tau < \infty \quad \text{and} \quad \int_{-\infty}^\infty x_2^2(\tau)\, d\tau < \infty.$$

Let $H(j\omega)$ denote the frequency function of $\mathbf{H}$, i.e.,

$$H(j\omega) = \sum_{i=1}^\infty h_i \exp(-j\omega\tau_i) + \int_0^\infty h(t) \exp(-j\omega t)\, dt.$$

We shall seek a solution to the following problem.

MAIN PROBLEM. Find conditions on $H(j\omega)$ which ensure that e_1 and e_2 are in $L_2(-\infty, \infty)$ and that $\lim_{t\to\infty} y_1(t) = 0$.

Our solution, which is given in Theorem 1, involves the following inequality:

$$\text{Re}(\{1 - Z(j\omega)\}H(j\omega)) \geqq \delta > 0, \qquad \omega \in (-\infty, \infty), \tag{5}$$

where $Z(j\omega)$ is a suitable frequency function. We refer to $1 - Z(j\omega)$ as the *stability multiplier*. More precisely, we have the following theorem.

THEOREM 1. *If there is an element $z(\cdot)$ in $L_1(-\infty, \infty)$, if there are sequences $\{\sigma_i\}$ and $\{z_i\}$ in $(-\infty, \infty)$ such that*

$$\int_{-\infty}^\infty |z(\tau)|\, d\tau + \sum_{i=1}^\infty |z_i| < 1, \tag{6}$$

and (5) is satisfied for $Z(j\omega)$ given by

$$Z(j\omega) = \sum_{i=1}^\infty z_i \exp(-j\omega\sigma_i) + \int_{-\infty}^\infty z(t) \exp(-j\omega t)\, dt, \tag{7}$$

and if either $z(\cdot) \geqq 0$ and $z_i \geqq 0$, $i = 1, 2, \cdots$, or $N(\cdot)$ is odd, then e_1 and e_2 are in $L_2(-\infty, \infty)$.

COROLLARY 1. *If, in addition to the hypotheses of the theorem, $h(\cdot)$ is in $L_2[0, \infty)$ and $h_i = 0$, $i = 1, 2, \cdots$, then $\lim_{t\to\infty} y_1(t) = 0$.*

COROLLARY 2 (Slope-restricted nonlinearities). *Suppose, in addition to the assumptions* A1–A4, *the following conditions are satisfied*:

(a) *there are constants* a, b, $\epsilon > 0$ *such that*

$$a \leqq \frac{N(x) - N(y)}{x - y} \leqq b - \epsilon$$

for all $x \neq y$;[1]

(b) $h_i = 0$ *for* $i = 1, 2, \cdots$;

(c) *there is an element* $z(\cdot)$ *in* $L_1(-\infty, \infty)$ *and there are sequences* $\{\sigma_i\}$ *and* $\{z_i\}$ *in* $(-\infty, \infty)$ *such that* (6) *holds and there is a constant* $\delta > 0$ *for which the inequality*

$$(5') \qquad \mathrm{Re}\{(1 - Z(j\omega))(H(j\omega) + b^{-1})(H^*(j\omega) + a^{-1})\} \geqq \delta > 0$$

is satisfied;

(d) *either* $z(\cdot) \geqq 0$ *and* $z_i \geqq 0$, $i = 1, 2, \cdots$, *or* $N(\cdot)$ *is odd*;

(e) *the Nyquist diagram*[2] *of* $H(j\omega)$ *does not encircle the point* $(-1/a, 0)$; *then* $e_1(\cdot)$ *and* $e_2(\cdot)$ *are in* $L_2(-\infty, \infty)$ *and* $\lim_{t\to\infty} y_1(t) = 0$.

We prove Theorem 1 and its corollaries in §7.

3. A method for generating stability conditions. In this section we develop the basic ideas underlying our method for generating stability conditions. We let Ω be a real Hilbert space and we view the space ${}^3L_2(R, \Omega)$ as a Hilbert space with inner product given by

$$(8) \qquad \langle x(\cdot), y(\cdot)\rangle = \int_{-\infty}^{\infty} \langle x(t), y(t)\rangle_\Omega \, dt.$$

We now have the following definition.

DEFINITION 1. Let $x(\cdot)$ be a mapping of R into Ω and let t be an element of R. Then the *t-truncation* of $x(\cdot)$, in symbols $x_t(\cdot)$, is the function defined by

$$(9) \qquad x_t(s) = \begin{cases} x(s) & \text{for } s \leqq t, \\ 0 & \text{for } s > t. \end{cases}$$

We let $L_{2e}(R, \Omega)$ denote the set of all mappings $x(\cdot)$ of R into Ω such that $x_t(\cdot) \in L_2(R, \Omega)$ for all t, i.e.,

$$(10) \qquad L_{2e}(R, \Omega) = \{x(\cdot) : x_t(\cdot) \in L_2(R, \Omega) \quad \text{for all } t \text{ in } R\}.$$

$L_{2e}(R, \Omega)$ shall be called the *extension* of $L_2(R, \Omega)$.

[1] Note that this implies that $| bx - N(x) | \geqq \epsilon | x |$ for all x.

[2] The Nyquist diagram of $H(\cdot)$ is the subset of the complex plane consisting of (i) the image of the $j\omega$-axis under $H(\cdot)$, and (ii) the origin. In our case, $H(j\omega)$ is a continuous curve with $\lim_{\omega\to\infty} H(j\omega) = 0$.

[3] R is the reals and $L_2(R, \Omega) = \left\{x(\cdot) : x(t) \in \Omega \text{ and } \int_{-\infty}^{\infty} \| x(t) \|^2 \, dt < \infty\right\}$.

For economy of notation, we write L_2 and L_{2e} in place of $L_2(R, \Omega)$ and $L_{2e}(R, \Omega)$, respectively.

If $x(\cdot)$ is an element of L_{2e}, then the *extended norm of* $x(\cdot)$, in symbols $\|x(\cdot)\|_e$, is given by

$$\|x(\cdot)\|_e = \begin{cases} \|x(\cdot)\| & \text{if } x(\cdot) \in L_2, \\ \infty & \text{if } x(\cdot) \notin L_2, \end{cases} \tag{11}$$

where $\|\cdot\|$ denotes the norm on L_2. Now let $\mathbf{T}$ be a mapping of L_2 into L_2 (or of L_{2e} into L_{2e}); then we have the next definition.

DEFINITION 2. $\mathbf{T}$ is *nonanticipative* if

$$(\mathbf{T}x)_t = (\mathbf{T}x_t)_t \tag{12}$$

for all x in L_2 (or L_{2e}) and all t in R.

Nonanticipative mappings play an important role in the sequel. We now have a third definition.

DEFINITION 3. Let $\mathbf{T}$ be a nonanticipative mapping of L_{2e} into L_{2e}. Then the *gain* of $\mathbf{T}$, $\gamma_e(\mathbf{T})$, is given by

$$\gamma_e(\mathbf{T}) = \sup_{\substack{x \in L_{2e} \\ t \in (-\infty, \infty) \\ x_t \neq 0}} \{\|(\mathbf{T}x(\cdot))_t\|_e / \|x_t(\cdot)\|_e\}. \tag{13}$$

$\mathbf{T}$ is said to be *positive* (e) if

$$\langle x_t(\cdot), (\mathbf{T}x(\cdot))_t \rangle \geqq 0 \tag{14}$$

for all $x(\cdot)$ in L_{2e} and all t in R, and to be *strongly positive* (e) if

$$\langle x_t(\cdot), (\mathbf{T}x(\cdot))_t \rangle \geqq \delta \|x_t(\cdot)\|_e^2 \tag{15}$$

for all $x(\cdot)$ in L_{2e}, all t in R, and some $\delta > 0$. A mapping $\mathbf{T}^*$ of L_{2e} into L_{2e} is said to be *a conjugate* (e) (or *adjoint* (e)) *of* $\mathbf{T}$ if

$$\langle x_t(\cdot), (\mathbf{T}y_t(\cdot)) \rangle = \langle (\mathbf{T}^* x_t(\cdot)), y_t(\cdot) \rangle \tag{16}$$

for all $x(\cdot)$, $y(\cdot)$ in L_{2e} and all t in R.

Definition 3 is a natural generalization of the corresponding properties for maps $\mathbf{T}$ of L_2 into L_2, and we shall say without further ado that $\mathbf{T}$ has a gain $\gamma(\mathbf{T})$, $\mathbf{T}$ is positive, $\mathbf{T}$ has a conjugate $\mathbf{T}^*$, etc.

Now consider the feedback system

$$f_1(\cdot) = v_1(\cdot) - w_2(\cdot), \tag{17}$$

$$f_2(\cdot) = v_2(\cdot) + w_1(\cdot), \tag{18}$$

$$w_1(\cdot) = \mathbf{G}_1 f_1(\cdot), \tag{19}$$

$$w_2(\cdot) = \mathbf{G}_2 f_2(\cdot), \tag{20}$$

where (a) $\mathbf{G}_1$ and $\mathbf{G}_2$ are nonanticipative maps of L_{2e} into L_{2e} ; (b) $f_1(\cdot)$, $f_2(\cdot)$, $v_1(\cdot)$, $v_2(\cdot)$, $w_1(\cdot)$ and $w_2(\cdot)$ are in L_{2e} ; and (c) $v_1(\cdot)$ and $v_2(\cdot)$ are in L_2 , i.e., $\| v_1(\cdot)\|_e$ and $\| v_2(\cdot)\|_e$ are finite. We wish to determine conditions which ensure that $f_1(\cdot)$ and $f_2(\cdot)$ are in L_2 . In particular, we want to prove a stability theorem for this system under the assumption that $\mathbf{G}_1$ and $\mathbf{G}_2$ are extensions of maps of L_2 into L_2 in the following sense.

DEFINITION 4. Let $\mathbf{T}$ be a nonanticipative map of L_2 into L_2 . A map $\mathbf{T}_e$ of L_{2e} is *an extension of* $\mathbf{T}$ if

$$(\mathbf{T}_e x(\cdot))_t = (\mathbf{T} x_t(\cdot))_t \tag{21}$$

for all $x(\cdot)$ in L_{2e} and all t in R.

Clearly $\mathbf{T}_e$ is nonanticipative and uniquely defined. Moreover, it is easy to see that $\mathbf{T}_e$ has the following properties:

(i) $\mathbf{T}_e x(\cdot) = \mathbf{T} x(\cdot)$ for all $x(\cdot)$ in L_2 ;
(ii) if $\mathbf{T}$ is linear (positive, strongly positive), then $\mathbf{T}_e$ is linear (positive (e), strongly positive (e));
(iii) $\gamma_e(\mathbf{T}_e) = \gamma(\mathbf{T})$;
(iv) if $\mathbf{T}^{-1}$ is defined and nonanticipative, then $\mathbf{T}_e^{-1}$ is defined and $\mathbf{T}_e^{-1} = (\mathbf{T}^{-1})_e$.

These properties will be used shortly.

We now prove a lemma which points the way toward our method for deriving stability conditions.

LEMMA 1. *If* $\mathbf{G}_1$ *and* $\mathbf{G}_2$ *are positive* (e) *and if either* $\mathbf{G}_1$ *is strongly positive* (e) *and* $\gamma_e(\mathbf{G}_1)$ *is finite or* $\mathbf{G}_2$ *is strongly positive* (e) *and* $\gamma_e(\mathbf{G}_2)$ *is finite, then* $f_1(\cdot)$ *and* $f_2(\cdot)$ *are in* L_2.

Proof. Let t be any element of R. Then

$$\begin{aligned}\langle v_{1t}(\cdot), w_{1t}(\cdot)\rangle + \langle v_{2t}(\cdot), w_{2t}(\cdot)\rangle \\ = \langle f_{1t}(\cdot), (\mathbf{G}_1 f_1(\cdot))_t\rangle + \langle f_{2t}(\cdot), (\mathbf{G}_2 f_2(\cdot))_t\rangle\end{aligned} \tag{22}$$

by virtue of (17)–(20). Suppose that $\mathbf{G}_1$ is strongly positive (e) and that $\gamma_e(\mathbf{G}_1)$ is finite. Then it follows from (22), the Schwarz inequality, and the positivity of $\mathbf{G}_2$ that

$$\delta \| f_{1t}(\cdot)\|_e^2 \leqq \| v_{1t}(\cdot)\|_e \| w_{1t}(\cdot)\|_e - \| v_{2t}(\cdot)\|_e \| w_{2t}(\cdot)\|_e \tag{23}$$

and hence, that

$$\begin{aligned}\delta \| f_{1t}(\cdot)\|_e^2 \leqq \| f_{1t}(\cdot)\|_e \{\gamma_e(\mathbf{G}_1)\| v_{1t}(\cdot)\|_e + \| v_{2t}(\cdot)\|_e\} \\ + \| v_{1t}(\cdot)\|_e \| v_{2t}(\cdot)\|_e .\end{aligned} \tag{24}$$

[4]

Since $\| v_{1t}(\cdot)\|_e \leqq \| v_1(\cdot)\|_e < \infty$ and $\| v_{2t}(\cdot)\|_e \leqq \| v_2(\cdot)\|_e < \infty$, there are

[4] If $f_{1t}(\cdot) \neq 0$ for some t, then $f_{1s}(\cdot) \neq 0$ for all $s > t$, and if $f_{1t}(\cdot) = 0$ for all t, then the assertions are trivially true.

positive constants α, β such that

$$\delta \| f_{1t}(\cdot) \|_e^2 \leqq \alpha \| f_{1t}(\cdot) \|_e + \beta \tag{25}$$

for all t. Since δ, α, and β are positive constants, it follows from (25) that there is an $M > 0$ such that $\| f_{1t}(\cdot) \|_e \leqq M$ for all t. Thus $f_1(\cdot)$ is in L_2. Furthermore, $\| f_2(\cdot) \|_e = \| v_2(\cdot) + \mathbf{G}_1 f_1(\cdot) \|_e \leqq \| v_2(\cdot) \|_e + \gamma_e(\mathbf{G}_1) \| f_1(\cdot) \|_e$ so that $f_2(\cdot)$ is also in L_2. The case where $\mathbf{G}_2$ is strongly positive (e) is treated in an entirely similar way and thus the lemma is established.

Now let us suppose that there are nonanticipative maps $\mathbf{H}'$ and $\mathbf{N}'$ of L_2 into L_2 such that $\mathbf{G}_1 = \mathbf{H}_e'$ and $\mathbf{G}_2 = \mathbf{N}_e'$. We then have the following theorem.

THEOREM 2. *Suppose that there is a mapping*[5] $\mathbf{M}$ (*the multiplier*) *of* L_2 *into* L_2 *such that*:

(i) *there are linear maps* $\mathbf{M}_+$ *and* $\mathbf{M}_-$ *of* L_2 *into* L_2 *with the following properties*:
  (a) $\mathbf{M} = \mathbf{M}_-\mathbf{M}_+$;
  (b) $\mathbf{M}_-$ *and* $\mathbf{M}_+$ *are invertible*;
  (c) $\mathbf{M}_+$, $\mathbf{M}_+^{-1}$, $\mathbf{M}^*$, *and* $\mathbf{M}_-^{*-1}$ *are nonanticipative and have finite gains* $\gamma(\cdot)$ (*i.e., are bounded*);

(ii) $\mathbf{M}\mathbf{H}'$ *and* $\mathbf{M}^*\mathbf{N}'$ *are positive*;

(iii) *either* $\mathbf{M}\mathbf{H}'$ *is strongly positive and* $\gamma(\mathbf{H}')$ *is finite or* $\mathbf{M}^*\mathbf{N}'$ *is strongly positive and* $\gamma(\mathbf{N}')$ *is finite.*

Then $f_1(\cdot)$ *and* $f_2(\cdot)$ *are in* L_2, *i.e.,* $\| f_1(\cdot) \|_e$ *and* $\| f_2(\cdot) \|_e$ *are finite.*

Proof. We shall transform the feedback equations (17)–(20) and apply Lemma 1. So let

$$\begin{aligned} v_1' &= (\mathbf{M}_-^*)_e v_1, & v_2' &= (\mathbf{M}_+)_e v_2, \\ f_1' &= (\mathbf{M}_-^*)_e f_1, & f_2' &= (\mathbf{M}_+)_e f_2, \\ w_2' &= (\mathbf{M}_-^*)_e w_2, & w_1' &= (\mathbf{M}_+)_e w_1, \end{aligned} \tag{26}$$

and

$$\begin{aligned} \mathbf{G}_1' &= (\mathbf{M}_+)_e \mathbf{H}_e' (\mathbf{M}_-^{*-1})_e, \\ \mathbf{G}_2' &= (\mathbf{M}_-^*)_e \mathbf{N}_e' (\mathbf{M}_+^{-1})_e. \end{aligned} \tag{27}$$

Then the feedback equations become

$$\begin{aligned} f_1' &= v_1' - w_2', & f_2' &= v_2' + w_1', \\ w_1' &= \mathbf{G}_1' f_1', & w_2' &= \mathbf{G}_2' f_2'. \end{aligned} \tag{28}$$

We shall show that (28) satisfies the hypotheses of Lemma 1.

Since $\mathbf{G}_1'$ and $\mathbf{G}_2'$ are compositions of nonanticipative maps, $\mathbf{G}_1'$ and $\mathbf{G}_2'$

[5] $\mathbf{M}$ may be anticipative.

are nonanticipative. Clearly f_1', v_1', w_1', f_2', v_2' and w_2' are in L_{2e}. Moreover, v_1' and v_2' are in L_2 since $\| v_1' \|_e = \| (\mathbf{M}_-^*)_e v_1 \|_e = \| (\mathbf{M}_-^*) v_1 \| \leqq \gamma(\mathbf{M}_-^*) \| v_1 \| < \infty$ as v_1 is in L_2 and $\| v_2' \|_e = \| (\mathbf{M}_+)_e v_2 \| = \| \mathbf{M}_+ v_2 \| \leqq \gamma(\mathbf{M}_+) \| v_2 \| < \infty$ as v_2 is in L_2.

By virtue of Lemma 2 which follows the theorem, $\mathbf{G}_1'$ and $\mathbf{G}_2'$ are positive (e). Moreover, if (say) $\mathbf{MH}'$ is strongly positive and $\gamma(\mathbf{H}')$ is finite, then it follows from Lemma 2 that $\mathbf{G}_1'$ is strongly positive (e). As $\gamma(\mathbf{H}') = \gamma_e(\mathbf{H}_e')$ we have $\gamma_e(\mathbf{G}') \leqq \gamma(\mathbf{M}_+)\gamma(\mathbf{H}')\gamma(\mathbf{M}_-^{*-1}) < \infty$. Thus, in this case, all the hypotheses of Lemma 1 are satisfied and so $\| f_1' \|_e < \infty$ and $\| f_2' \|_e < \infty$. But $\| f_1 \|_e = \| (\mathbf{M}_-^{*-1})_e f_1' \|_e \leqq \gamma(\mathbf{M}_-^{*-1}) \| f_1' \|_e < \infty$ and similarly, $\| f_2 \|_e < \infty$. The case where $\mathbf{M}^*\mathbf{N}'$ is strongly positive is treated in the same way and thus the theorem is established.

LEMMA 2. *Let* $\mathbf{P}$, $\mathbf{Q}$ *and* $\mathbf{R}$ *be nonanticipative maps of* L_2 *into* L_2. *If* $\gamma(\mathbf{Q}) < \infty$, *if* $\mathbf{Q}^{-1}$ *exists and is nonanticipative and if* $\mathbf{Q}^*\mathbf{PR}$ *is positive (strongly positive), then* $\mathbf{P}_e\mathbf{R}_e\mathbf{Q}_e^{-1}$ *is positive* (e) *(strongly positive* (e)*).*

Proof. We have

$$\begin{aligned} \langle x_t , (\mathbf{P}_e\mathbf{R}_e\mathbf{Q}_e^{-1}x)_t \rangle &= \langle x_t , (\mathbf{P}_e\mathbf{R}_e\mathbf{Q}_e^{-1}x_t)_t \rangle \\ &= \langle x_t , (\mathbf{PRQ}^{-1}x_t) \rangle \\ &= \langle y, \mathbf{Q}^*\mathbf{PR}y \rangle, \end{aligned}$$

where $y = \mathbf{Q}^{-1}x_t$, since $\mathbf{P}_e\mathbf{R}_e\mathbf{Q}_e^{-1}$ is nonanticipative. The lemma follows immediately as $\| x_t \| \leqq \gamma(\mathbf{Q}) \| y \|$.

4. Factorization of operators. In view of Theorem 2, we can see the importance of determining conditions which insure that an operator has a suitable factorization. If $\mathfrak{B}$ is some Banach algebra and P is an element of $\mathfrak{L}(\mathfrak{B}, \mathfrak{B})$,[6] then we shall call P a *projection* if (a) $P^2 = P$ and (b) xy is in the range of P if both x and y are in the range of P. The factorization that we shall ultimately use involves projections in a Banach algebra of convolution operators.

We now have the following lemma.

LEMMA 3. *Let* $\mathfrak{B}$ *be a commutative Banach algebra with an identity* E *and with norm* $\rho(\cdot)$. *Let* P_+ *be a projection on* $\mathfrak{B}$ *and let* $P_- = E_{\mathfrak{L}(\mathfrak{B},\mathfrak{B})} - P_+$. *Denote the ranges of* P_+ *and* P_- *by* $\mathfrak{B}_+$ *and* $\mathfrak{B}_-$ *respectively. Let* $\tilde{\mathfrak{B}}_+$ *be the subspace spanned by* $\mathfrak{B}_+$ *and* E, *and let* $\tilde{\mathfrak{B}}_-$ *be the subspace spanned by* $\mathfrak{B}_-$ *and* E. *If* Z *is any nonzero element of* $\mathfrak{B}$ *with* $\rho(Z) < 1$, *then there are elements* Z_+ *of* $\tilde{\mathfrak{B}}_+$ *and* Z_- *of* $\tilde{\mathfrak{B}}_-$ *such that*:

(i) $E + Z = Z_-Z_+$;

(ii) Z_+ *and* Z_- *are invertible and* Z_+^{-1}, Z_-^{-1} *are in* $\tilde{\mathfrak{B}}_+$ $\tilde{\mathfrak{B}}_-$ *respectively.*

[6] $\mathfrak{L}(\mathfrak{B}, \mathfrak{B})$ is the space of all continuous linear maps of $\mathfrak{B}$ into $\mathfrak{B}$.

Moreover,

$$\begin{aligned} Z_+ &= \exp\,[P_+(\log\,\{E + Z\})], \\ Z_- &= \exp\,[P_-(\log\,\{E + Z\})] \end{aligned} \tag{29}$$

and

$$\begin{aligned} Z_+^{-1} &= \exp\,[-P_+(\log\,\{E + Z\})], \\ Z_-^{-1} &= \exp\,[-P_-(\log\,\{E + Z\})]. \end{aligned} \tag{30}$$

Proof. Since $\rho(Z) < 1$, the series

$$\log\,(E + Z) = Z - \frac{Z^2}{2} + \frac{Z^3}{3} - \cdots \tag{31}$$

converges absolutely to an element of $\mathfrak{B}$. The series for

$$\exp\,[P_+(\log\,\{E + Z\})]$$

also converges to an element Z_+ of $\mathfrak{B}$. Z_+ is actually in $\mathfrak{B}_+$ since (a) $P_+(\log\,\{E + Z\})$ is in $\mathfrak{B}_+$, (b) $\mathfrak{B}_+$ is closed under addition and multiplication, and (c) P_+ is continuous. A similar argument applies to $Z_- = \exp\,[P_-(\log\,\{E + Z\})]$.

Moreover, since $\mathfrak{B}$ is commutative, we have

$$\begin{aligned} Z_-Z_+ &= \exp\,[P_-(\log\,\{E + Z\})]\exp\,[P_+(\log\,\{E + Z\})] \\ &= \exp\,[P_-(\log\,\{E + Z\}) + P_+(\log\,\{E + Z\})] \\ &= E + Z. \end{aligned} \tag{32}$$

Clearly Z_+^{-1} and Z_-^{-1} are defined by (30) and thus the lemma is established.

5. The Banach algebra of convolution operators. Suppose that $g_c(\cdot)$ is an element of $L_1(-\infty, \infty)$ and that $g_d(\cdot)$ is a real-valued function on $(-\infty, \infty)$ which has a countable support[7] $S(g_d)$ and which is absolutely summable on $S(g_d)$ (i.e., $\sum_{\tau_i \in S(g_d)} |g_d(\tau_i)| < \infty$). Then we shall consider convolutions of the form:

$$y(t) = \sum_{\tau_i \in S(g_d)} g_d(\tau_i)x(t - \tau_i) + \int_{-\infty}^{\infty} g_c(\tau)x(t - \tau)\,d\tau, \tag{33}$$

where $x(\cdot)$ is an element of $L_2(-\infty, \infty)$. We call the pair $\{g_d, g_c\}$ the *kernel of the convolution.* We shall show that any convolution of the form (33) is a bounded linear transformation of $L_2(-\infty, \infty)$ into itself and that the set of all such convolutions can be viewed as a commutative Banach algebra

[7] In other words, $g_d(\tau) = 0$ for all $\tau \notin S(g_d)$.

$\mathfrak{L}$ with an identity. Moreover, there is a natural projection P_+ on $\mathfrak{L}$. We begin with the following lemma.

LEMMA 4. *Let kl_1 be the set of all real-valued functions $\varphi(\cdot)$ on $(-\infty, \infty)$ which have countable support S_φ and which are absolutely summable on S_φ. Let $\|\varphi(\cdot)\|_k$ be given by*

$$\|\varphi(\cdot)\|_k = \sum_{\tau_i \in S_\varphi} |\varphi(\tau_i)| \tag{34}$$

for $\varphi(\cdot)$ in kl_1. Then kl_1 is a Banach space with $\|\cdot\|_k$ as norm.

Proof. Clearly kl_1 is a linear space with the usual definitions of sum and scalar multiplication. Moreover, $\|\cdot\|_k$ is obviously a norm on kl_1. Thus we need only show that kl_1 is complete. So let $\{\varphi_n\}_{n=1}^{\infty}$ be a Cauchy sequence in kl_1 and let $S = \bigcup_1^\infty S_{\varphi_n}$. Then S is countable. Letting σ_n be the element of l_1 given by $\sigma_n = \{\varphi_n(\tau_i)\}_{i=1}^{\infty}$, where $S = \{\tau_i : i = 1, 2, \cdots\}$, we can see that σ_n is a Cauchy sequence in l_1 and hence has a limit $\sigma = \{\alpha_i\}_{i=1}^{\infty}$ in l_1. If we let $\varphi(\cdot)$ be defined by $\varphi(\tau_i) = \alpha_i$ for τ_i in S and $\varphi(\tau) = 0$ for $\tau \notin S$, then $\varphi(\cdot)$ is in kl_1 and $\{\varphi_n\}_{i=1}^{\infty}$ converges to $\varphi(\cdot)$. Thus the lemma is established.

Let K be the direct sum of the Banach spaces kl_1 and $L_1(-\infty, \infty)$, i.e., $K = kl_1 \oplus L_1(-\infty, \infty)$. K shall be called the *kernel space* and we shall denote elements of K by $\{g_d, g_c\}$. Moreover, we also have

$$\|\{g_d, g_c\}\|_K = \|g_d\|_k + \|g_c\|_1$$

as the norm on K, or more explicitly,

$$\|\{g_d, g_c\}\|_K = \sum_{\tau_i \in S_{g_d}} |g_d(\tau_i)| + \int_{-\infty}^{\infty} |g_c(\tau)|\, d\tau. \tag{35}$$

We now have the next lemma.

LEMMA 5. *If $\{g_d, g_c\}$ is an element of K and $x(\cdot)$ is an element of $L_2(-\infty, \infty)$, then the convolution (33) converges for almost all t in $(-\infty, \infty)$. Furthermore, $y(\cdot)$ is in $L_2(-\infty, \infty)$ and $\|y(\cdot)\|_2 \leqq \|\{g_c, g_d\}\|_K \cdot \|x(\cdot)\|_2$. Hence the convolution (33) defines a bounded linear transformation of $L_2(-\infty, \infty)$ into itself.*

Proof. Formally, we have

$$\int_{-\infty}^{\infty} y^2(t)\, dt = I_{dd} + 2I_{dc} + I_{cc}, \tag{36}$$

where

$$\begin{aligned}
I_{dd} &= \int_{-\infty}^{\infty} \Big\{ \Big(\sum_{\tau_i \in S_{g_d}} g_d(\tau_i) x(t - \tau_i) \Big) \Big(\sum_{\sigma_i \in S_{g_d}} g_d(\sigma_i) x(t - \sigma_i) \Big) \Big\}\, dt, \\
I_{dc} &= \int_{-\infty}^{\infty} \Big\{ \Big(\sum_{\tau_i \in S_{g_d}} g_d(\tau_i) x(t - \tau_i) \Big) \Big(\int_{-\infty}^{\infty} g_c(\sigma) x(t - \sigma)\, d\sigma \Big) \Big\}\, dt, \qquad (37) \\
I_{cc} &= \int_{-\infty}^{\infty} \Big\{ \Big(\int_{-\infty}^{\infty} g_c(\tau) x(t - \tau)\, d\tau \Big) \Big(\int_{-\infty}^{\infty} g_c(\sigma) x(t - \sigma)\, d\sigma \Big) \Big\}\, dt.
\end{aligned}$$

We shall show that each of the terms I_{dd}, I_{dc}, I_{cc} is finite. Assuming this for the moment, we can see that the convolution (33) converges for almost all t and that $y(\cdot)$ is in $L_2(-\infty, \infty)$.

Consider, for example, I_{cc}. We have

$$I_{cc} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_c(\tau)g_c(\sigma)x(t-\tau)x(t-\sigma)\,dt\,d\tau\,d\sigma \tag{38}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_c(\tau)g_c(\sigma)\left\{\int_{-\infty}^{\infty} x(t-\tau)x(t-\sigma)\,dt\right\}d\tau\,d\sigma \tag{39}$$

$$\leqq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g_c(\tau)|\,|g_c(\sigma)|\left\{\int_{-\infty}^{\infty} x^2(t-\tau)\,dt\int_{-\infty}^{\infty} x^2(s-\sigma)\,ds\right\}^{1/2} d\tau\,d\sigma \tag{40}$$

$$\leqq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g_c(\tau)|\,|g_c(\sigma)|\cdot\|x(\cdot)\|_2^2\,d\tau\,d\sigma \tag{41}$$

$$\leqq \|g_c\|_1^2\cdot\|x(\cdot)\|_2^2. \tag{42}$$

The terms in (42) are finite by hypothesis; (39) is bounded by (40) as a consequence of the Schwarz inequality; (38) is obtained from (39) by inverting the orders of integration in accordance with the Fubini-Tonnelli theorems. Moreover, Tonnelli's theorem also implies that the bracketed integrand of I_{cc} in (37) converges for almost all t. The integrals I_{dd} and I_{dc} are treated in a similar way and thus are finite. Furthermore,

$$|I_{dd}| \leqq \|g_d\|_k^2\cdot\|x(\cdot)\|_2^2,$$

$$|I_{dc}| \leqq \|g_d\|_k\cdot\|g_c\|_1\cdot\|x(\cdot)\|_2^2,$$

and thus the lemma is established.

In view of Lemma 5, we let $\mathfrak{L}$ be the set of all operators $\mathbf{G}$ in $\mathfrak{L}(L_2(-\infty, \infty), L_2(-\infty, \infty))$ which are defined by convolutions with kernels in K. Addition, scalar multiplication and composition for operators in $\mathfrak{L}$ are defined in the usual way. Moreover, the identity operator $\mathbf{I}$ is in $\mathfrak{L}$. We shall show that $\mathfrak{L}$ can be viewed as a commutative Banach algebra with identity $\mathbf{I}$.

DEFINITION 5. Let $\{g_d, g_c\}$ and $\{h_d, h_c\}$ be elements of K. Then the *product of* $\{g_d, g_c\}$ *and* $\{h_d, h_c\}$, in symbols $\{g_d, g_c\} * \{h_d, h_c\}$, is the element of K given by $\{g_d \circ h_d, g_d \circ h_c + g_c \circ h_d + g_c \circ h_c\}$, where

$$\begin{aligned} g_d \circ h_d(t) &= \sum_{\tau_i \in S_{g_d}} g_d(\tau_i)h_d(t-\tau_i), \\ g_d \circ h_c(t) &= \sum_{\tau_i \in S_{g_d}} g_d(\tau_i)h_c(t-\tau_i), \\ g_c \circ h_d(t) &= \sum_{\mu_j \in S_{h_d}} h_d(\mu_j)g_c(t-\mu_j), \\ g_c \circ h_c(t) &= \int_{-\infty}^{\infty} g_c(\tau)h_c(t-\tau)\,d\tau. \end{aligned} \tag{43}$$

It is readily verified from well-known properties of ordinary convolutions of functions that $\{g_d, g_c\} * \{h_d, h_c\}$ is a well-defined element of K and that K is a commutative algebra with an identity (given by $\{e_d, 0\}$, where $e_d(0) = 1$, $e_d(t) = 0$ for $t \neq 0$) with respect to the product $*$. Moreover,

$$\|\{g_d, g_c\} * \{h_d, h_c\}\|_K \leqq \|\{g_d, g_c\}\|_K \|\{h_d, h_c\}\|_K \tag{44}$$

so that K is a commutative Banach algebra with an identity.

Let the mapping that takes kernels into operators be denoted by π, i.e., if $\{g_d, g_c\}$ is in K, then $\pi(\{g_d, g_c\})$ is the convolution operator $\mathbf{G}$ defined by (33). π is clearly a linear map of K onto $\mathfrak{L}$. We shall show that π is an algebraic isomorphism of K and $\mathfrak{L}$. In other words, we prove that

$$\pi(\{g_d, g_c\} * \{h_d, h_c\}) = \pi(\{g_d, g_c\})\pi(\{h_d, h_c\}) \tag{45}$$

and that π is an injection (i.e., $\pi(\{g_d, g_c\}) \neq \mathbf{0}$ if $\{g_d, g_c\} \neq 0$). It will then follow that $\mathfrak{L}$ can be viewed as a commutative Banach algebra with identity $\mathbf{I}$ since the function $\rho(\cdot)$ given by

$$\rho(\pi(\{g_d, g_c\})) = \|\{g_d, g_c\}\|_K \tag{46}$$

is a norm on $\mathfrak{L}$. So we now have the next lemma.

LEMMA 6. *If* $\mathbf{G} = \pi(\{g_d, g_c\})$ *and* $\mathbf{H} = \pi(\{h_d, h_c\})$, *then* $\mathbf{GH} = \pi(\{g_d, g_c\} * \{h_d, h_c\})$.

Proof. If $x(\cdot)$ is an element of $L_2(-\infty, \infty)$, then

$$\begin{aligned}\mathbf{G}(\mathbf{H}x)(t) &= \sum_{\tau_i \in S_{g_d}} g_d(\tau_i)\mathbf{H}x(t - \tau_i) + \int_{-\infty}^{\infty} g_c(\tau)\mathbf{H}x(t - \tau)\, dt \\ &= J_{dd} + J_{dc} + J_{cd} + J_{cc},\end{aligned} \tag{47}$$

where

$$J_{dd} = \sum_{\tau_i \in S_{g_d}} g_d(\tau_i) \sum_{\sigma_j \in S_{h_d}} h_d(\sigma_j)x(t - \tau_i - \sigma_j),$$

$$J_{dc} = \sum_{\tau_i \in S_{g_d}} g_d(\tau_i) \int_{-\infty}^{\infty} h_c(\sigma)x(t - \tau_i - \sigma)\, d\sigma,$$

$$J_{cd} = \int_{-\infty}^{\infty} g_c(\tau) \sum_{\sigma_j \in S_{h_d}} h_d(\sigma_j)x(t - \tau - \sigma_j)\, d\tau,$$

$$J_{cc} = \int_{-\infty}^{\infty} g_c(\tau) \int_{-\infty}^{\infty} h_c(\sigma)x(t - \tau - \sigma)\, d\sigma\, d\tau.$$

Since (47) converges for almost all t, all the terms J_{dd}, J_{dc}, J_{cd} and J_{cc} are well defined. Moreover, by inverting orders of integration or summation

and by translating variables, we easily see that

$$J_{dd} = \pi(\{g_d, 0\} * \{h_d, 0\}),$$
$$J_{dc} = \pi(\{g_d, 0\} * \{0, h_c\}),$$
$$J_{cd} = \pi(\{0, g_c\} * \{h_d, 0\}),$$
$$J_{cc} = \pi(\{0, g_c\} * \{0, h_c\}),$$

and so the lemma is established.

All that remains is to show that π is injective. We shall do this by means of the Fourier transform.

DEFINITION 6. If $x(\cdot)$ is in $L_2(-\infty, \infty)$, then *the l.i.m.*[8] *Fourier transform* $X(j\omega)$ *of* $x(\cdot)$ is given by

$$X(j\omega) = \text{l.i.m.}_{T\to\infty} \int_{-T}^{T} x(t) \exp(-j\omega t)\, dt. \tag{48}$$

If $\{g_d, g_c\}$ is in K, then the *Fourier transform* $G(j\omega)$ *of* $\{g_d, g_c\}$ is given by

$$G(j\omega) = \sum_{\tau_i \in S_{g_d}} g_d(\tau_i) \exp(-j\omega\tau_i) + \int_{-\infty}^{\infty} g_c(\tau) \exp(-j\omega\tau)\, d\tau, \tag{49}$$

which converges absolutely for $\omega \in (-\infty, \infty)$. If $\mathbf{G} = \pi(\{g_d, g_c\})$, then we often speak of $G(j\omega)$ as the Fourier transform of $\mathbf{G}$.

The basic properties of the Fourier transform that we need are:

(A) If $x(\cdot)$ is in $L_2(-\infty, \infty)$, if $\mathbf{G} = \pi(\{g_d, g_c\})$ and if $y = \mathbf{G}x$, then $Y(j\omega) = X(j\omega)G(j\omega)$ and

$$y(t) = \text{l.i.m.}_{w\to\infty} \frac{1}{2\pi} \int_{-w}^{w} X(j\omega)G(j\omega) \exp(+j\omega t)\, d\omega. \tag{50}$$

(B) If $x(\cdot)$ is a nonzero element of $L_2(-\infty, \infty)$ (or if $\{g_d, g_c\}$ is a nonzero element of K), then $X(j\omega) \not\equiv 0$ ($G(j\omega) \not\equiv 0$). In other words, the Fourier transformation is one-to-one.

(C) If $\{g_d, g_c\}$ is a nonzero element of K, then there is an $x(\cdot)$ in $L_2(-\infty, \infty)$ such that $\| \mathbf{G}x \|_2 \neq 0$, where $\mathbf{G} = \pi(\{g_d, g_c\})$. In other words π is injective.

Properties (A) and (B) are standard (see, for example, [14]) and property (C) is a direct consequence of (50) and Parseval's theorem. Thus we have shown that $\mathfrak{L}$ is a commutative Banach algebra with an identity.

We now exhibit the natural projection P_+ on $\mathfrak{L}$. If $\mathbf{G} = \pi(\{g_d, g_c\})$ is an element of $\mathfrak{L}$, then we let

$$P_+\mathbf{G} = \pi(\{g_d^+, g_c^+\}), \tag{51}$$

[8] Limit in the mean.

where

$$g_a^+(\tau) = \begin{cases} g_a(\tau) & \text{if } \tau \geqq 0, \\ 0 & \text{if } \tau < 0, \end{cases} \tag{52}$$

$$g_c^+(\tau) = \begin{cases} g_c(\tau) & \text{if } \tau \geqq 0, \\ 0 & \text{if } \tau < 0. \end{cases} \tag{53}$$

We observe that (a) P_+ is linear, (b) $\rho(P_+\mathbf{G}) \leqq \rho(\mathbf{G})$ so that P_+ is continuous, (c) $P_+^2 = P_+$, and (d) $(P_+\mathbf{G})(P_+\mathbf{H}) = \pi(\{g_a^+, g_c^+\} * \{h_a^+, h_c^+\})$. Thus P_+ is indeed a projection on $\mathfrak{L}$. Letting $P_- = \mathbf{I} - P_+$, we denote the ranges of P_+ and P_- by $\mathfrak{L}_+$ and $\mathfrak{L}_-$, respectively. We observe that now the factorization Lemma 3 applies to the Banach algebra $\mathfrak{L}$.

The operators in $\tilde{\mathfrak{L}}_+$ are nonanticipative. Furthermore, since the adjoint $\mathbf{G}^*$ of any operator $\mathbf{G} = \pi(\{g_a, g_c\})$ in $\mathfrak{L}$ is an operator in $\mathfrak{L}$ with kernel $\{g_a(-\tau_i), g_c(-\tau)\}$, we can see that the adjoint of any operator in $\tilde{\mathfrak{L}}_-$ lies in $\tilde{\mathfrak{L}}_+$ and is, therefore, nonanticipative. This fact is relevant to the proof of Theorem 1.

6. Positivity conditions. We now develop conditions for the positivity of compositions of operators. Our derivations are based on an area inequality which is closely related to Young's inequality.

Throughout this section we let $N(\cdot)$ be a real-valued function on $(-\infty, \infty)$ such that $N(0) = 0$.

LEMMA 7 (An area inequality). *If $N(\cdot)$ is monotone nondecreasing, then*

$$xN(x) - yN(x) \geqq P(x) - P(y) \tag{54}$$

for all x and y, where $P(x) = \int_0^x N(s)\, ds$.

Proof. Since $N(\cdot)$ is monotone nondecreasing,

$$[N(x + k\Delta x) - N(x)]k\Delta x \geqq 0 \tag{55}$$

for any integer k, and hence,

$$\sum_{k=1}^{m} [N(x + k\Delta x) - N(x)]\Delta x \geqq 0 \tag{56}$$

for any integer m. Setting $\Delta x = (y - x)/m$ and letting m approach infinity, we deduce that

$$\int_x^y N(s)\, ds - N(x) \int_x^y ds \geqq 0, \tag{57}$$

which is equivalent to (54).

LEMMA 8. *If $N(\cdot)$ is monotone nondecreasing and if there is a constant*

$C > 0$ such that $|N(s)| \leqq C|s|$, then

$$\int_{-\infty}^{\infty} x(t+\tau)N(x(t))\,dt \leqq \int_{-\infty}^{\infty} x(t)N(x(t))\,dt \tag{58}$$

for all τ and any $x(\cdot)$ in $L_2(-\infty, \infty)$. If, in addition, $N(\cdot)$ is odd, then

$$\left|\int_{-\infty}^{\infty} x(t+\tau)N(x(t))\,dt\right| \leqq \int_{-\infty}^{\infty} x(t)N(x(t))\,dt \tag{59}$$

for all τ and any $x(\cdot)$ in $L_2(-\infty, \infty)$.

Proof. Since $|N(s)| \leqq C|s|$ and $x(\cdot)$ is in $L_2(-\infty, \infty)$, $N(x(\cdot))$ is in $L_2(-\infty, \infty)$ and $P(x(\cdot))$ is in $L_1(-\infty, \infty)$. Thus,

$$\begin{aligned}\int_{-\infty}^{\infty} \{x(t)N(x(t)) - x(t+\tau)N(x(t))\}\,dt &\\ \geqq \int_{-\infty}^{\infty} P(x(t))\,dt - \int_{-\infty}^{\infty} P(x(t+\tau))\,dt &\\ \geqq 0,&\end{aligned}$$

so that (58) holds. If $N(\cdot)$ is odd, then $P(\cdot)$ is even, and so

$$\begin{aligned}\int_{-\infty}^{\infty} \{x(t)N(x(t)) + x(t+\tau)N(x(t))\}\,dt &\\ = \int_{-\infty}^{\infty} \{x(t)N(x(t)) - \{-(x(t+\tau)\}N(x(t))\}\,dt &\\ \geqq \int_{-\infty}^{\infty} \{P(x(t)) - P(-x(t+\tau))\}\,dt &\\ \geqq 0.&\end{aligned}$$

Thus, (59) holds.

Now suppose that $N(\cdot)$ satisfies the assumption A1 of §2 and let $\mathbf{N}$ be the mapping of $L_2(-\infty, \infty)$ into $L_2(-\infty, \infty)$ given by $\mathbf{N}x(t) = N(x(t))$. We then have the following proposition.

PROPOSITION 1. *If $\mathbf{Z} = \pi(\{z_d, z_c\})$ is an element of $\mathfrak{L}$ with $\rho(\mathbf{Z}) < 1$, and if either the kernel $\{z_d, z_c\}$ is nonnegative or $N(\cdot)$ is odd, then $(\mathbf{I} - \mathbf{Z})\mathbf{N}$ is positive.*

Proof. If $x(\cdot)$ is in $L_2(-\infty, \infty)$, then

$$\begin{aligned}\langle x(\cdot), \mathbf{ZN}x(\cdot)\rangle = \sum_{\tau_i \in S_{z_d}} z_d(\tau_i) \int_{-\infty}^{\infty} x(t)N(x(t-\tau_i))\,dt &\\ + \int_{-\infty}^{\infty} z_c(\tau) \int_{-\infty}^{\infty} x(t)N(x(t-\tau))\,dt\,d\tau.&\end{aligned} \tag{60}$$

If $\{z_d, z_c\}$ is nonnegative, then (58) and (60) yield

$$\langle x(\cdot), \mathbf{ZN}x(\cdot)\rangle \leqq \rho(\mathbf{Z})\langle x(\cdot), \mathbf{N}x(\cdot)\rangle, \tag{61}$$

which implies that

$$\langle x(\cdot), (\mathbf{I} - \mathbf{Z})\mathbf{N}x(\cdot)\rangle \geqq \langle x(\cdot), \mathbf{N}x(\cdot)\rangle\{1 - \rho(\mathbf{Z})\} \geqq 0. \tag{62}$$

If $N(\cdot)$ is odd, then (59) and (60) yield (61).

Proposition 1 is a positivity condition for the composition of a linear operator $\mathbf{I} - \mathbf{Z}$ and a nonlinear operator $\mathbf{N}$. In Proposition 2 which follows, we give a positivity condition for the composition of two convolution operators.

PROPOSITION 2. *If* $\mathbf{G}$ *and* $\mathbf{H}$ *are elements of* $\mathcal{L}$ *with Fourier transforms* $G(j\omega)$ *and* $H(j\omega)$ *respectively, and if*

$$\operatorname{Re}\{G(j\omega)H(j\omega)\} \geqq \delta, \tag{63}$$

then $\mathbf{GH}$ *is positive (strongly positive) if* $\delta = 0$ $(\delta > 0)$.

Proof. The Fourier transform of $\mathbf{GH}$ is simply $G(j\omega)H(j\omega)$. Thus, if $x(\cdot)$ is any element of $L_2(-\infty, \infty)$, then

$$\langle x(\cdot), \mathbf{GH}x(\cdot)\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(j\omega)H(j\omega) \mid X(j\omega) \mid^2 d\omega \tag{64}$$

by Parseval's theorem. It follows that

$$\begin{aligned}\langle x(\cdot), \mathbf{GH}x(\cdot)\rangle &= \frac{1}{\pi}\int_0^{\infty} \operatorname{Re}\{G(j\omega)H(j\omega)\} \mid X(j\omega) \mid^2 d\omega \\ &\geqq \frac{\delta}{\pi}\int_0^{\infty} \mid X(j\omega) \mid^2 d\omega = \delta \parallel x(\cdot) \parallel^2,\end{aligned} \tag{65}$$

which establishes the proposition.

7. Proof of the main result. Theorem 1 is now a simple consequence of Theorem 2. The hypotheses of Theorem 2 are satisfied by virtue of Propositions 1 and 2, the fact that $\mathcal{L}$ is a commutative Banach algebra with an identity and a projection P_+, and the fact that $\rho(\mathbf{Z}) < \infty$ implies $\gamma(\mathbf{Z}) < \infty$ for $\mathbf{Z}$ in $\mathcal{L}$.

As for Corollary 1, we observe that since $h_d(\cdot) = 0$ and $h_c(\cdot) = h(\cdot)$ is in $L_1(-\infty, \infty) \cap L_2(-\infty, \infty)$,

$$\begin{aligned}y_1(t) &= \int_0^{\infty} h(\tau)e_1(t - \tau)\, d\tau \\ &= \int_{t/2}^{t} h(t - \sigma)e_1(\sigma)\, d\sigma + \int_{t/2}^{\infty} h(\tau)e_1(t - \tau)\, d\tau,\end{aligned}$$

and hence that

$$|y_1(t)|^2 \leqq 2\left(\int_{t/2}^{\infty} h^2(t-\sigma)\,d\sigma\right)\left(\int_{t/2}^{\infty} e_1^2(\sigma)\,d\sigma\right) + 2\left(\int_{t/2}^{\infty} h^2(\tau)\,d\tau\right)\left(\int_{t/2}^{\infty} e_1^2(t-\tau)\,d\tau\right). \tag{66}$$

Since $e_1(\cdot)$ and $h(\cdot)$ are in $L_2(-\infty, \infty)$, it follows by taking square roots in (66) that $\lim_{t\to\infty} y_1(t) = 0$.

We now turn our attention to Corollary 2. We shall prove this corollary by transforming the feedback equations (1)–(4) into a form to which Theorem 1 applies. The transformed equations are of the form

$$\begin{aligned} x_1' &= x_1 - ax_2, & x_2' &= bx_2 - x_1, \\ e_1' &= e_1 + ay_1 = (\mathbf{I} + a\mathbf{H})e_1, & e_2' &= be_2 - y_2 = (b\mathbf{I} - \mathbf{N})e_2, \\ y_1' &= e_1 + by_1 = (\mathbf{I} + b\mathbf{H})e_1, & y_2' &= y_2 - ae_2 = (\mathbf{N} - a\mathbf{I})e_2. \end{aligned} \tag{67}$$

If the indicated inverses exist (as we shall soon show), then it is easy to see that

$$\begin{aligned} e_1' &= x_1' - y_2', & y_1' &= (\mathbf{I} + b\mathbf{H})(\mathbf{I} + a\mathbf{H})^{-1}e_1', \\ e_2' &= x_2' + y_1', & y_2' &= (\mathbf{N} - a\mathbf{I})(b\mathbf{I} - \mathbf{N})^{-1}e_2', \end{aligned} \tag{68}$$

and that the hypotheses of Theorem 1 are satisfied for (68). If we show that $(\mathbf{I} + a\mathbf{H})^{-1}$ is in $\mathfrak{L}$, and that $(b\mathbf{I} - \mathbf{N})^{-1}$ is a nonlinear operator with the same properties as the $\mathbf{N}$ of the theorem, then we can deduce that $e_1 = (\mathbf{I} + a\mathbf{H})^{-1}e_1'$ and $e_2 = (b\mathbf{I} - \mathbf{N})^{-1}e_2'$ are in $L_2(-\infty, \infty)$, which will establish Corollary 2.

Now the condition that $b \geqq \{N(x) - N(y)\}/(x - y)$ insures that the function $(b\mathbf{I} - \mathbf{N})(\cdot)$ is monotone nondecreasing which implies that $(b\mathbf{I} - \mathbf{N})^{-1}(\cdot)$ exists and is a monotone nondecreasing function. Furthermore, $(b\mathbf{I} - \mathbf{N})^{-1}(0) = 0$ and $|(b\mathbf{I} - \mathbf{N})^{-1}(x)| \leqq |x|/\epsilon$ (since $|(b\mathbf{I} - \mathbf{N})(x)| \geqq \epsilon|x|$). Thus $(b\mathbf{I} - \mathbf{N})^{-1}$ has all the requisite properties and, in view of the properties of compositions, so does $(\mathbf{N} - a\mathbf{I})(b\mathbf{I} - \mathbf{N})^{-1}$.

Let us examine $\mathbf{I} + a\mathbf{H}$. We first show that the inequality (5′) is equivalent to the following "circle condition":

$$\left|H(j\omega) + \frac{1}{2}\left\{(a^{-1} + b^{-1}) + j(a^{-1} - b^{-1})\frac{Z_i(j\omega)}{1 - Z_r(j\omega)}\right\}\right| \geqq \left|\frac{1}{2}(a^{-1} - b^{-1})\left\{1 - \frac{jZ_i(j\omega)}{1 - Z_r(j\omega)}\right\}\right| + \eta, \tag{69}$$

where $\eta > 0$ and $Z(j\omega) = Z_r(j\omega) + jZ_i(j\omega)$. To do this we note that

(5′) is equivalent to

$$\text{Re}\{[1 - Z_r(j\omega) - jZ_i(j\omega)][|H(j\omega) + \frac{1}{2}(a^{-1} + b^{-1})|^2 - \frac{1}{2}(a^{-1} - b^{-1})^2 + j(a^{-1} - b^{-1})H_i(j\omega)]\} \geqq \delta. \tag{70}$$

In view of (6), $1 - Z_r(j\omega) > 0$ and so (70) is equivalent to the inequality

$$\left|H(j\omega) + \frac{1}{2}(a^{-1} + b^{-1})\right|^2 - \frac{1}{2}(a^{-1} - b^{-1})^2 + \frac{(a^{-1} - b^{-1})Z_i(j\omega)H_i(j\omega)}{1 - Z_r(j\omega)} \geqq \frac{\delta}{1 - Z_r(j\omega)}. \tag{71}$$

Upon completing the square of $H_i(j\omega) + \frac{1}{2}(a^{-1} - b^{-1})Z_i(j\omega)/(1 - Z_r(j\omega))$, we see that (71) holds if and only if

$$\left|H(j\omega) + \frac{1}{2}\left\{(a^{-1} + b^{-1}) + j(a^{-1} - b^{-1})\frac{Z_i(j\omega)}{1 - Z_r(j\omega)}\right\}\right|^2 \geqq \frac{1}{4}(a^{-1} - b^{-1})^2\left|1 - j\frac{Z_i(j\omega)}{1 - Z_r(j\omega)}\right|^2 + \frac{\delta}{1 - Z_r(j\omega)}. \tag{72}$$

But (72) is clearly equivalent[9] to (69) (as $1 - Z_r(j\omega)$ is bounded away from 0 and $H(j\omega)$ is bounded) and so (5′) and (69) are equivalent. Since

$$H(j\omega) + a^{-1} = H(j\omega) + \frac{1}{2}(a^{-1} + b^{-1}) + \frac{1}{2}j\frac{(a^{-1} - b^{-1})Z_i(j\omega)}{1 - Z_r(j\omega)} - \left\{-\frac{1}{2}(a^{-1} - b^{-1}) + \frac{1}{2}j\frac{(a^{-1} - b^{-1})Z_i(j\omega)}{1 - Z_r(j\omega)}\right\},$$

we have $|H(j\omega) + a^{-1}| \geqq \eta > 0$ in view of (69). It follows from (69) and a lemma of Zames [3b, Part II] that $\mathbf{I} + a\mathbf{H}$ is injective (i.e., is one-to-one). In view of a theorem of Paley and Wiener [13, Theorem XVIII, p. 60] the nonanticipative operator $\mathbf{I} + a\mathbf{H}$ has a nonanticipative postinverse in $\mathfrak{L}$ and therefore, is surjective. Thus we have shown that $(\mathbf{I} + a\mathbf{H})^{-1}$ exists and is a nonanticipative operator in $\mathfrak{L}$.

Moreover, since the composition of two nonanticipative operators in $\mathfrak{L}$ is again a nonanticipative element of $\mathfrak{L}$, $(\mathbf{I} + b\mathbf{H})(\mathbf{I} + a\mathbf{H})^{-1}$ is a nonanticipative element of $\mathfrak{L}$. Thus the proof of Corollary 2 is complete.

8. Concluding remarks. We have derived stability conditions for a class of feedback systems in terms of the frequency response of the linear part $\mathbf{H}$ and a suitable multiplier $\mathbf{I} - \mathbf{Z}$. The key point was that $\mathbf{I} - \mathbf{Z}$ was

[9] Equation (72) is of the form $(\alpha - \beta)(\alpha + \beta) \geqq \gamma > 0$ with $\alpha, \beta > 0$.

defined on $(-\infty, \infty)$. The proof involved the factorization of operators in the Banach algebra $\mathfrak{L}$ of convolution operators and the development of positivity conditions for compositions of linear and nonlinear operators.

Although we have considered only the case of scalar-valued functions here, we can easily generalize Theorem 1 to the case of vector-valued e_i, x_i and y_i, $i = 1, 2, \cdots$. The operator $\mathbf{H}$ is defined just as in (3); however, the mapping $\mathbf{N}$ is no longer defined by a scalar function N but rather by a vector function N_v with the following properties:

(i) $N_v(\mathbf{0}) = \mathbf{0}$;

(ii) $\langle \mathbf{r} - \mathbf{s}, N_v(\mathbf{r}) - N_v(\mathbf{s}) \rangle \geqq 0$;

(iii) there is a constant $C > 0$ such that $\| N_v(\mathbf{r}) \| \leqq C \| \mathbf{r} \|$ for all $\mathbf{r}$.

A careful perusal of our proofs will clearly indicate the validity of this generalization.

Another avenue of generalization is via the extension of the kernel space to an $L^1(G)$, where G is a locally compact group; this generalization will be studied in a later paper.

REFERENCES

[1] A. I. Lur'e, *Some Nonlinear Problems in the Theory of Automatic Control*, Her Majesty's Stationery Office, London, 1957.

[2] V. M. Popov, *Absolute stability of nonlinear systems of automatic control*, Automat. Remote Control, 22 (1962), pp. 857–875.

[3a] G. Zames, *On the stability of nonlinear, time-varying feedback systems*, Proc. National Electronics Conference, 20 (1964), pp. 725–730.

[3b] ———, *On the input-output stability of time-varying nonlinear feedback systems, Parts I, II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, 465–476.

[4] R. W. Brockett and J. W. Willems, *Frequency domain stability criteria, Parts I, II*, Proc. Joint Automatic Control Conference, Troy, New York, 1965, pp. 735–747.

[5] V. A. Yakubovich, *Frequency conditions for the absolute stability and dissipativity of control systems with a single differentiable nonlinearity*, Soviet Math. Dokl., 6 (1965a), pp. 98–101.

[6] K. S. Narendra and C. P. Neuman, *Stability of a class of differential equations with a single monotone nonlinearity*, this Journal, 4 (1966), pp. 295–308.

[7] B. D. O. Anderson, *Stability of control systems with multiple nonlinearities*, J. Franklin Inst., 282 (1966), pp. 155–160.

[8] A. G. Dewey and E. I. Jury, *A stability inequality for a class of nonlinear feedback systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 54–62.

[9] C. T. Lee and C. A. Desoer, *Stability of single-loop nonlinear feedback systems*, Rep. ERL 66-13, Electronics Research Laboratory, University of California, Berkeley, 1966.

[10a] R. P. O'Shea, *A combined frequency-time domain stability criterion for autonomous continuous systems*, Proc. Joint Automatic Control Conference, Seattle, Washington, 1966, pp. 832–840.

[10b] ———, *An improved frequency-time domain stability criterion for continuous autonomous systems*, unpublished report.

[11] G. ZAMES AND P. L. FALB, *On the stability of systems with monotone and odd monotone nonlinearities*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 221–223.

[12] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, I*, Interscience, New York, 1966.

[13] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Providence, Rhode Island, 1934.

[14] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, 2nd ed., Oxford University Press, Oxford, 1964.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

A REMARK ON THE "BANG-BANG" PRINCIPLE FOR LINEAR CONTROL SYSTEMS IN INFINITE-DIMENSIONAL SPACE*

H. O. FATTORINI†

1. Introduction. The "bang-bang" principle for linear control systems

$$u'(t) = A(t)u(t) + B(t)f(t) \tag{1.1}$$

in finite-dimensional space E^n (LaSalle [7]) can be stated as follows: If the system (1.1) can be steered from a point $u \in E^n$ to another point $v \in E^n$ in a given time $t_1 > 0$ by a control f taking values, say, in the unit cube K of E^m, then the transfer of (1.1) from u to v can also be achieved in the same time by another control f_0 taking values in K_0, the set of extremal points of K. This result has been extended in various directions; let us only mention [9], where K is allowed to be any compact convex set in E^m (see also [4], [8] for other types of generalizations). The "bang-bang" principle does not hold in infinite-dimensional spaces; in fact, it is easy to construct control systems, even with A and B time-independent where the final state v at a given time t_2 depends *uniquely* on the control f (see [2], [3]). However, the principle subsists if we satisfy ourselves with approximating (and not actually attaining) the final state. Moreover, it turns out that we can also approximate the whole trajectory between u and v by means of K_0-valued controls (Theorem 1). This result is similar in form to the one in [5] for nonlinear control systems in finite-dimensional space (see also [11] where a very general treatment of this type of approximation problems is to be found).

2. The initial value problem. We shall denote by E, F two (real or complex) Banach spaces, $L(F; E)$ the Banach space of all linear bounded operators from F to E endowed, as usual, with the uniform operator topology.[1] For each t in $[t_0, t_1]$, $t_0 < t_1$, $A(t)$ will be a (possibly unbounded) linear operator with domain $D(A(t))$. We shall assume that the Cauchy

* Received by the editors May 22, 1967, and in revised form August 24, 1967.

† Department of Mathematics, University of California at Los Angeles, Los Angeles, California 90024. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant 693-67, and in part by the National Aeronautics and Space Administration under Grant NGR 40-002-015 at Brown University, Division of Applied Mathematics, Providence, Rhode Island.

[1] See [1] for definitions and results used here.

problem for

$$u'(t) = A(t)u(t) \tag{2.1}$$

is *well set*. This means there exists an *evolution operator* $U(t, s)$, i.e., a strongly continuous $L(E; E)$-valued function $U(t, s)$ defined in the triangle $t_0 \leqq s \leqq t \leqq t_1$ satisfying $U(t, t) = I$, $t_0 \leqq t \leqq t_1$, and such that for each $t \in [t_0, t_1)$, $u \in E$ the function

$$u(s) = U(s, t)u$$

is a (classical or generalized) solution of (2.1) in $[t, t_1]$. For any E-valued strongly measurable function $g(\cdot)$ defined and summable in $[t_0, t_1]$ and any $u \in E$ we shall define the expression[2]

$$u(t) = U(t, t_0)u + \int_{t_0}^{t} U(t, s)g(s)\, ds \tag{2.2}$$

to be a solution of the inhomogeneous equation

$$u'(t) = A(t)u(t) + g(t). \tag{2.3}$$

It is easy to see that the function $u(\cdot)$ defined by (2.2) exists and is continuous in $[t_0, t_1]$. Under additional conditions on $A(t)$, $U(t, s)$, $g(s)$ and u it is possible to show that (2.2) is a genuine solution of (2.3); we shall not dwell upon this point here (see, for instance, [12]).

Finally, we consider the linear control system

$$u'(t) = A(t)u(t) + B(t)g(t), \qquad t_0 \leqq t \leqq t_1. \tag{2.4}$$

For each t, $t_0 \leqq t \leqq t_1$, $B(t)$ is a bounded operator from F to E. We assume that $B(\cdot)$ is strongly measurable, i.e., that for any $u \in F$ the E-valued function $B(\cdot)u$ is strongly measurable; moreover, we suppose there exists a scalar-valued function $\eta(\cdot)$, summable in $[t_0, t_1]$, such that

$$|B(t)| \leqq \eta(t), \qquad t_0 \leqq t \leqq t_1. \tag{2.5}$$

The class of controls $\mathfrak{L}_K$ consists of all strongly measurable F-valued functions $f(\cdot)$ defined in $[t_0, t_1]$ with values in some fixed closed, bounded, convex set K. The *trajectories* of the system (2.4) are the solutions of (2.4) for some control $f \in \mathfrak{L}_K$, i.e., functions of the form

$$u(t) = U(t, t_0)u + \int_{t_0}^{t} U(t, s)B(s)f(s)\, ds \tag{2.6}$$

with $f \in \mathfrak{L}_K$. Since $B(\cdot)f(\cdot)$ is summable in E, each trajectory $u(\cdot)$ is continuous in $[t_0, t_1]$.

[2] All the integrals throughout this paper are Bochner integrals; see [6, Chap. 3] for an exposition of the theory of integration of vector-valued functions.

3. The "bang-bang" principle. In all that follows, K_0 will be a subset of K satisfying the following assumption.

ASSUMPTION 1. *Finite convex combinations of elements of* K_0 (*i.e., finite sums* $\sum \lambda_k u_k$, $\lambda_k \geqq 0$, $\sum \lambda_k = 1$, $u_k \in K_0$) *are dense in* K.

Let us call $\mathcal{L}^*_{K_0}$ the subset of $\mathcal{L}_K$ defined by the following two conditions:

(a) There exist disjoint intervals $I_1, \cdots, I_n$, $\bigcup I_k = [t_0, t_1]$, such that f is constant in each I_k.

(b) $f(t) \in K_0$ for all $t \in [t_0, t_1]$.

THEOREM 1. *Let* $u(\cdot)$ *be a trajectory of* (2.4) *corresponding to some* $f \in \mathcal{L}_K$ *and let* $\epsilon > 0$. *Then there exists an* $f_0 \in \mathcal{L}^*_{K_0}$ *such that the trajectory* $u_0(\cdot)$ *corresponding to* f_0 *satisfies*

$$| u(t) - u_0(t) |_E \leqq \epsilon, \qquad t_0 \leqq t \leqq t_1 .$$

The proof of Theorem 1 is a consequence of the following auxiliary result.[3]

LEMMA 1. *Let* X *be a Banach space,* $N(\cdot)$ *an* $L(F; X)$*-valued, strongly measurable function defined in* $[t_0, t_1]$ *such that* $| N(t) | \leqq \eta(t)$, $t_0 \leqq t \leqq t_1$, *for some summable function* $\eta(\cdot)$. *Let* $\mathcal{K}$ ($\mathcal{K}_0$) *be the set of all elements of* X *of the form*

$$\int_{t_0}^{t_1} N(s) f(s)\, ds, \tag{3.1}$$

$f \in \mathcal{L}_K$ ($f \in \mathcal{L}^*_{K_0}$). *Then* $\mathcal{K}_0$ *is dense in* $\mathcal{K}$.

In fact, assume Lemma 1 holds. Denote by $\mathcal{C}(E) = X$ the Banach space of all E-valued continuous functions $u(\cdot)$ defined in $[t_0, t_1]$ (norm $| u(\cdot) |_X = \sup_{t_0 \leqq t \leqq t_1} | u(t) |$). Let $\epsilon > 0$ and U_ϵ be the $L(E; E)$-valued function defined in the square $t_0 \leqq s, t \leqq t_1$ as being equal to $U(t, s)$ in the triangle $t_0 \leqq s \leqq t \leqq t_1$, null in the triangle $t_0 \leqq t \leqq s - \epsilon \leqq t_1 - \epsilon$ and defined elsewhere such as to be strongly continuous in the square and such that

$$\sup_{t_0 \leqq s, t \leqq t_1} | U_\epsilon(t, s) |_{L(E;E)} = C = \sup_{t_0 \leqq s \leqq t \leqq t_1} | U(t, s) |_{L(E;E)} . \tag{3.2}$$

(For instance, $U_\epsilon(t, s) = (1/\epsilon)(t + \epsilon - s) U(\frac{1}{2}(t + s), \frac{1}{2}(t + s))$ in $t \leqq s \leqq \min(t + \epsilon, t_1)$.)

It is not difficult to see that the $L(F; \mathcal{C}(E))$-valued function $N(s) = U_\epsilon(\cdot, s) B(s)$, $t_0 \leqq s \leqq t_1$, is strongly measurable, and (2.5) implies that it satisfies the rest of the assumptions of Lemma 1. Consequently, Lemma 1 tells us that the set of elements of $\mathcal{C}(E)$ of the form

$$\int_{t_0}^{t_1} U_\epsilon(t, s) B(s) f(s)\, ds \tag{3.3}$$

with $f \in \mathcal{L}^*_{K_0}$ is dense (in the $\mathcal{C}(E)$-topology) in the set of elements of the

[3] See [10], where a closely related result is proved.

form (3.3) with $f \in \mathfrak{L}_K$. We note now that

$$\left| \int_{t_0}^{t} U(t, s)B(s)f(s)\, ds - \int_{t_0}^{t_1} U_\epsilon(t, s)B(s)f(s)\, ds \right|_E \leqq CC_1 \int_t^{\min(t+\epsilon, t_1)} \eta(s)\, ds,$$

where C_1 is an upper bound for $\{ |f| ; f \in K\}$ and $\eta(\cdot)$ the function in (2.5). This completes the proof of Theorem 1.

Proof of Lemma 1. The proof is trivial if $N(\cdot)$ is *uniformly* measurable (i.e., measurable as an $L(F; X)$-valued function). For in this case, given $\epsilon > 0$, we can find disjoint intervals whose union differs from $[t_0, t_1]$ in a set of measure $\leqq \epsilon$ and operators $N_1, \cdots, N_n \in L(F; X)$ such that

$$|N(s) - N_k|_{L(F;X)} \leqq \epsilon, \quad s \in I_k.$$

This makes clear that we only need to prove Lemma 1 for the case N *constant*. Let $f \in \mathfrak{L}_K$. If $v = \int_{t_0}^{t_1} f(s)\, ds$, it follows from the fact that K is closed and convex that $(t_1 - t_0)^{-1}v \in K$. Then it can be approximated arbitrarily well by convex combinations $u = \sum \lambda_k u_k$, $u_k \in K_0$. But then Nv can be approximated by elements of the form $N(t_1 - t_0)u$, and $N(t_1 - t_0)u = \int_{t_0}^{t_1} Nf_0(s)\, ds$, where $f_0(s) = u_k$ for $s \in J_k$, J_k an arbitrary family of (disjoint) subintervals of $[t_0, t_1]$, length $(J_k) = (t_1 - t_0)\lambda_k$.

Observe next that if F is finite-dimensional, the concepts of strong and uniform measurability for $N(\cdot)$ coincide. We shall thus end the proof by reducing the general case to that in which $\dim F < \infty$. Let $f \in \mathfrak{L}_K$. Since f is strongly measurable, we can find a g of the form

$$g(s) = \sum_{\text{(finite)}} X_k(s)u_k, \tag{3.4}$$

$u_1, u_2, \cdots \in K$, $X_1, X_2, \cdots$ characteristic functions of disjoint measurable sets $e_1, e_2, \cdots$ in $[t_0, t_1]$, such that $|f(s) - g(s)| \leqq \epsilon$ in $[t_0, t_1]$ outside a set of measure $\leqq \epsilon$, thus we can assume f to be of the form (3.4). Now, since each u_k can be approximated by convex combinations $\sum_{j=1}^{m(k)} \lambda_{kj}u_{kj}$, $u_{kj} \in K_0$, we can assume that the values of f actually belong to the convex hull K' of the points u_{kj}, $k = 1, 2, \cdots$, $1 \leqq j \leqq m(k)$. But K' is contained in the finite-dimensional subspace F' of F generated by the u_{kj}, and those points satisfy Assumption 1 (with respect to K'); thus our result for finite-dimensional F applies, and the proof of Lemma 1 is completed.

Remark 1. Assume F is reflexive. Thus K is compact in F with respect to the weak topology. But then, by the Krein-Milman theorem [1, Chap. V, §8.4], K is the closed (in the weak topology) convex envelope of its set of

extremal points. However, the closed convex envelope of a set is the same in the strong as in the weak topology; thus K_e, the set of extremal points of K, satisfies Assumption 1. In some cases, K_0 can be chosen to be a proper subset of K_e. The most interesting case in application is that in which K_0 is substantially smaller than K; for instance, if K is a polyhedron in a finite-dimensional space F, we may take K_0 to be the set of its vertices. Thus, the steering of (2.1) can be achieved up to any degree of accuracy with controls assuming only a finite number of values.

Remark 2. Theorem 1 admits evident generalizations to infinite time intervals (t_0, ∞), higher order systems, etc.

REFERENCES

[1] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators I*, Interscience, New York, 1958.

[2] P. FALB, *Infinite dimensional control problems I: On the closure of the set of attainable states for linear systems*, J. Math. Anal. Appl., 9 (1964), pp. 12–22.

[3] H. O. FATTORINI, *Control in finite time of differential equations in Banach space*, Comm. Pure Appl. Math., 19 (1966), pp. 17–34.

[4] H. HALKIN, *A generalization of LaSalle's "bang-bang" control principle*, this Journal, 2 (1965), pp. 199–202.

[5] H. HERMES, *A note on the range of a vector measure; application to the theory of optimal control*, J. Math. Anal. Appl., 8 (1964), pp. 78–83.

[6] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, Colloquium Publications, vol. XXXI, American Mathematical Society, Providence, Rhode Island, 1957.

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

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

[9] L. M. SONNEBORN AND F. S. VAN VLECK, *The "bang-bang" principle for linear control systems*, this Journal, 2 (1965), pp. 151–159.

[10] C. CASTAING, *Sur une extension du théorème de Lyapunov*, C. R. Acad. Sci. Paris, 260 (1965), pp. 3834–3841.

[11] J. WARGA, *Functions of relaxed controls*, this Journal, 5 (1967), pp. 628–641.

[12] T. KATO AND H. TANABE, *On the abstract evolution equation*, Osaka J. Math, 14 (1962), pp. 107–133.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

DUALITY FOR CONTROL PROBLEMS*

BERTRAM MOND† AND MORGAN HANSON‡

Introduction. A number of duality theorems for the control problem have recently appeared in the literature (see, e.g., [2], [4], [8] and [9]). In general, these references give conditions under which an extremal solution of the control problem yields a solution of the corresponding dual.

The main result of this note is the converse duality theorem (Theorem 3) which gives conditions under which a solution of the dual problem yields a solution of the control problem. The relationship of our results to duality in mathematical programming is also discussed.

Notation. $f(t, x, u)$ will denote a scalar function with continuous derivatives up to and including the second order with respect to each of its arguments. $G(t, x, u)$ and $R(t, x, u)$ are, respectively, n- and r-dimensional vector functions with continuous derivatives up to and including the second order. x, u, λ and μ are, respectively, n-, m-, n- and r-dimensional functions of t.

A prime will denote derivative with respect to t. Superscripts denote vector components; subscripts denote partial derivatives. No notational distinction is made between row and column vectors.

f_x and f_u are the gradient vectors of f with respect to x and u. λG_x, λG_u, μR_x and μR_u are the gradient vectors with respect to x and u of λG and μR. Similarly, letting ν be an n-dimensional vector function of t, νf_{xx}, $\nu\lambda G_{xx}$ and $\nu\mu R_{xx}$ mean $(\nu f_x)_x$, $(\nu(\lambda G_x))_x$ and $(\nu(\mu R_x))_x$.

Broadly, the control problem is to choose, under given conditions, a control vector $u(t)$, such that the state vector $x(t)$ is brought from some specified initial state $x(t_0) = x_0$ to some specified final state $x(t_1) = x_1$ in such a way as to minimize a given functional. A more precise mathematical formulation is given in problem P below.

Duality. Consider the following two problems.

PRIMAL P. Minimize

$$\int_{t_0}^{t_1} f(t, x, u)\, dt$$

* Received by the editors May 5, 1967, and in revised form November 6, 1967.

† Aerospace Research Laboratories, Wright-Patterson Air Force Base, Dayton, Ohio 45433.

‡ Queen's University, Kingston, Ontario, and Aerospace Research Laboratories, Wright-Patterson Air Force Base, Dayton, Ohio.

subject to

$$x(t_0) = x_0, \quad x(t_1) = x_1, \tag{1}$$

$$G(t, x, u) = x', \tag{2}$$

$$R(t, x, u) \geqq 0. \tag{3}$$

DUAL D. Maximize

$$\int_{t_0}^{t_1} \{f(t, x, u) - \lambda(t)[G(t, x, u) - x'] - \mu(t)R(t, x, u)\}\, dt$$

subject to

$$x(t_0) = x_0, \quad x(t_1) = x_1, \tag{4}$$

$$f_x(t, x, u) - \lambda(t)G_x(t, x, u) - \mu(t)R_x(t, x, u) = \lambda'(t), \tag{5}$$

$$f_u(t, x, u) - \lambda(t)G_u(t, x, u) - \mu(t)R_u(t, x, u) = 0, \tag{6}$$

$$\mu(t) \geqq 0. \tag{7}$$

In the above two problems, $u(t)$ is required to have piecewise continuous first and second derivatives in the interval $t_0 \leqq t \leqq t_1$. $x(t)$ and $\lambda(t)$ are required to be continuous in $t_0 \leqq t \leqq t_1$; $x'(t)$, $\lambda'(t)$ and $\mu(t)$ are required to be continuous in $t_0 \leqq t \leqq t_1$ except for values of t corresponding to discontinuities of $u(t)$. The constraints must be fulfilled for all t, $t_0 \leqq t \leqq t_1$, except that for values of t corresponding to points of discontinuity of $u(t)$; (2) and (5) must be fulfilled for right- and left-hand limits.

THEOREM 1. *If f is convex in x and u, λG and R are concave in x and u, then the infimum of* P *is greater than or equal to the supremum of* D.

Proof. Let (x^*, u^*) satisfy (1), (2) and (3) and let (x, u, λ, μ) satisfy (4), (5), (6) and (7). Then

$$\int_{t_0}^{t_1} \{f(t, x^*, u^*) - f(t, x, u)\}\, dt$$

$$\geqq \int_{t_0}^{t_1} \{(x^* - x)f_x(t, x, u) + (u^* - u)f_u(t, x, u)\}\, dt$$

(by the convexity of f)

$$= \int_{t_0}^{t_1} (x^* - x)\lambda'\, dt + \int_{t_0}^{t_1} \{(x^* - x)[\lambda G_x(t, x, u) + \mu R_x(t, x, u)]$$

$$+ (u^* - u)[\lambda G_u(t, x, u) + \mu R_u(t, x, u)]\}\, dt \qquad \text{(by (5) and (6))}$$

$$= -\int_{t_0}^{t_1} (x^{*\prime} - x')\lambda\, dt + \int_{t_0}^{t_1} \{(x^* - x)\lambda G_x(t, x, u) + (u^* - u)\lambda G_u(t, x, u)$$

$$+ (x^* - x)\mu R_x(t, x, u) + (u^* - u)\mu R_u(t, x, u)\} \, dt$$

(by integration by parts, (1) and (4))

$$\geqq \int_{t_0}^{t_1} [-\lambda x^{*\prime} + \lambda x'] \, dt + \int_{t_0}^{t_1} \{\lambda[G(t, x^*, u^*) - G(t, x, u)]$$

$$+ \mu[R(t, x^*, u^*) - R(t, x, u)]\} \, dt \qquad \text{(by the concavity of } \lambda G \text{ and } R\text{)}$$

$$\geqq \int_{t_0}^{t} \{-\lambda[G(t, x, u) - x'] - \mu R(t, x, u)\} \, dt \qquad \text{(by (2), (3) and (7))}.$$

Hence

$$\int_{t_0}^{t_1} f(t, x^*, u^*) \, dt$$

$$\geqq \int_{t_0}^{t_1} \{f(t, x, u) - \lambda(t)[G(t, x, u) - x'] - \mu(t)R(t, x, u)\} \, dt,$$

and, therefore, the infimum of P is greater than or equal to the supremum of D.

Note that if G is linear in x and u, then λG is always concave in x and u. This is the case that is considered in [2].

The *convexity* of f and the *concavity* of λG and R with respect to x and u will henceforth be assumed.

We assume also, for the next theorem, that the functions R^i, $i = 1, \cdots, r$, satisfy the constraint conditions:

(i) if $r > m$, then at each (t, x, u) at most m components of R can vanish;

(ii) at each (t, x, u) the matrix $(\partial R^i/\partial u^j)$, where i ranges over those indices where $R^i(t, x, u) = 0$ and $j = 1, \cdots, m$, has maximum rank.

Necessary conditions for the existence of an extremal solution for a variational problem subject to both equality and inequality constraints were given by Valentine [10]. Using Valentine's results, Berkovitz [1] obtained corresponding necessary conditions for the control problem P. These may be stated in the following way. If (x^*, u^*) is an optimal solution for P, then there exists a function of the form

$$F \equiv \lambda_0 f - \lambda(t)[G - x'] - \mu(t)R \tag{8}$$

such that

$$F_x = \frac{d}{dt} F_{x'}, \tag{9}$$

$$F_u = 0, \tag{10}$$

$$\mu^i R^i = 0, \qquad i = 1, \cdots, r, \tag{11}$$

$$\mu \geqq 0 \tag{12}$$

hold throughout $t_0 \leqq t \leqq t_1$ (except that for values of t corresponding to points of discontinuity of $u(t)$, (9) holds for right- and left-hand limits). Here λ_0 is a nonnegative constant, $\lambda(t)$ is continuous in $t_0 \leqq t \leqq t_1$, and λ_0, $\lambda(t)$, $\mu(t)$ cannot vanish simultaneously for any t, $t_0 \leqq t \leqq t_1$.

It will be assumed that the minimizing arc determined by (x^*, u^*) is *normal*, i.e., that λ_0 can be taken equal to 1.

THEOREM 2. *If (x^*, u^*) is an optimal solution of* P, *then there exist functions $\lambda(t)$ and $\mu(t)$ such that $(x^*(t), u^*(t), \lambda(t), \mu(t))$ is an extremal solution of* D *and the extreme values of* P *and* D *are equal.*

Proof. It follows from Berkovitz's results [1] that there exist $\lambda(t)$ and $\mu(t)$ such that (5), (6) and (7) are satisfied. Thus (x^*, u^*, λ, μ) satisfies the constraints of D. In addition, we have, from (11),

$$\mu(t)R(t, x^*, u^*) = 0. \tag{13}$$

Equations (2) and (13) and Theorem 1 imply that (x^*, u^*, λ, u) maximizes D.

Converse duality. We now consider the converse dual problem, that is, of finding conditions under which the existence of an extremal solution of problem D implies the existence of an extremal solution to the control problem P.

We assume now that f, G and R have continuous derivatives up to and including the third order with respect to each of their arguments.

Let us write (6) as $F(t, x, u, \lambda, \mu) = 0$ and let $z \equiv (x, u)$. We assume that the functions F^i, $i = 1, \cdots, m$, satisfy the constraint condition:

(iii) at each (t, x, u, λ, μ) the matrix $(\partial F^i/\partial z^j)$, $i = 1, \cdots, m$, and $j = 1, \cdots, n + m$, has rank m.

The results of Valentine [10] applied to D state that if $(x^*, u^*, \lambda^*, \mu^*)$ is an extremal solution of D, then there exists a function

$$H \equiv \nu_0[f - \lambda^*(G - x^{*\prime}) - \mu^* R] - \nu(t)[f_x - \lambda^* G_x - \mu^* R_x - \lambda^{*\prime}] - \tau(t)[f_u - \lambda^* G_u - \mu^* R_u] - \gamma(t)\mu^*$$

such that

$$H_x = \frac{d}{dt} H_{x'}, \tag{14}$$

$$H_u = 0, \tag{15}$$

$$H_\lambda = \frac{d}{dt} H_{\lambda'}, \tag{16}$$

$$H_\mu = 0, \tag{17}$$

$$\gamma(t) \leqq 0, \tag{18}$$

$$\gamma^i \mu^{*i} = 0, \qquad i = 1, \cdots, m, \tag{19}$$

hold throughout $t_0 \leqq t \leqq t_1$ (except that for values of t corresponding to points of discontinuity of $u(t)$, (14) and (16) hold for right- and left-hand limits). Here ν_0 is a nonnegative constant, $\nu(t)$, $\tau(t)$ and $\gamma(t)$ are n-, m- and r-dimensional functions of t, continuous except possibly for values of t corresponding to points of discontinuity of $u(t)$. ν_0, $\nu(t)$, $\tau(t)$ and $\gamma(t)$ cannot vanish simultaneously for any t, $t_0 \leqq t \leqq t_1$.

We shall assume that the arc corresponding to the extremal solution $(x^*, u^*, \lambda^*, \mu^*)$ is *normal*, i.e., that ν_0 may be taken equal to 1.

THEOREM 3. *If* $(x^*, u^*, \lambda^*, \mu^*)$ *is an extremal solution of* D *such that the matrix*

$$\begin{pmatrix} f_{xx} - \lambda G_{xx} - \mu R_{xx} & f_{ux} - \lambda G_{ux} - \mu R_{ux} \\ f_{xu} - \lambda G_{xu} - \mu R_{xu} & f_{uu} - \lambda G_{uu} - \mu R_{uu} \end{pmatrix} \tag{20}$$

is nonsingular for all t, $t_0 \leqq t \leqq t_1$, *then* (x^*, u^*) *is an optimal solution of* P, *and the extreme values of* P *and* D *are equal.*

Proof. It follows from Valentine's results [10] applied to D that there exist ν_0, $\nu(t)$, $\tau(t)$ and $\gamma(t)$ such that

$$\begin{aligned} f_x - \lambda^* G_x - \mu^* R_x - \nu(t)[f_{xx} - \lambda^* G_{xx} - \mu^* R_{xx}] \\ - \tau(t)[f_{ux} - \lambda^* G_{ux} - \mu^* R_{ux}] = \lambda^{*\prime}, \end{aligned} \tag{21}$$

$$\begin{aligned} f_u - \lambda^* G_u - \mu^* R_u - \nu(t)[f_{xu} - \lambda^* G_{xu} - \mu^* R_{xu}] \\ - \tau(t)[f_{uu} - \lambda^* G_{uu} - \mu^* R_{uu}] = 0, \end{aligned} \tag{22}$$

$$-R + \nu(t)R_x + \tau(t)R_u - \gamma(t) = 0, \tag{23}$$

$$-G + x' + \nu(t)G_x + \tau(t)G_u = \nu'(t), \tag{24}$$

$$\gamma(t) \leqq 0, \tag{25}$$

$$\gamma^i \mu^i = 0, \qquad i = 1, \cdots, m. \tag{26}$$

Equations (5), (6), (21) and (22) imply

$$-\nu(t)[f_{xx} - \lambda^* G_{xx} - \mu^* R_{xx}] - \tau(t)[f_{ux} - \lambda^* G_{ux} - \mu^* R_{ux}] = 0, \tag{27}$$

$$-\nu(t)[f_{xu} - \lambda^* G_{xu} - \mu^* R_{xu}] - \tau(t)[f_{uu} - \lambda^* G_{uu} - \mu^* R_{uu}] = 0. \tag{28}$$

By the hypothesis of the theorem, $\nu(t) = 0, \tau(t) = 0, t_0 \leqq t \leqq t_1$, is the only solution of (27) and (28).

Equations (23) and (24) now become

$$-R - \gamma(t) = 0, \tag{29}$$

$$-G + x' = 0. \tag{30}$$

Equations (29) and (25) yield

$$R \geqq 0. \tag{31}$$

Hence (x^*, u^*) satisfies the constraints of P.

From (26) and (29) it follows that

$$\mu^* R = 0. \tag{32}$$

The theorem now follows from (30), (32) and Theorem 1.

Remark 1. The matrix (20) will be positive semidefinite because of the convexity of f and the concavity of $\lambda^* G$ and R. The condition that the matrix be nonsingular is thus equivalent to the condition that the matrix be positive definite.

Remark 2. If P and D are both independent of t and x, they reduce essentially to the static cases of mathematical programming. Putting $t_1 - t_0 = 1$, P and D become the following problems.

PROBLEM 1*. Minimize

$$f(u)$$

subject to

$$G(u) = 0,$$

$$R(u) \geqq 0.$$

PROBLEM 2*. Maximize

$$f(u) - \lambda G(u) - \mu R(u)$$

subject to

$$f_u(u) - \lambda G_u(u) - \mu R_u(u) = 0,$$

$$\mu \geqq 0,$$

where f is convex, λG and R are concave in u. Theorems 1–3 then reduce to corresponding duality theorems for mathematical programming. If only inequality constraints are given, then Problems 1* and 2* become the dual mathematical programs of [3], [5] and [11].

Remark 3. It is possible to formulate P and D with different types of endpoint conditions. See [4] and [8] as well as [6] and [7] for further discussions pertinent to this point.

REFERENCES

[1] L. D. BERKOVITZ, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3(1961), pp. 145–169.

[2] M. A. HANSON, *Bounds for functionally convex optimal control problems*, Ibid., 8(1964), pp. 84–89.

[3] P. HUARD, *Dual programs*, IBM J. Res. Develop., 6(1962), pp. 137–139.

[4] E. KREINDLER, *Reciprocal optimal control problems*, J. Math. Anal. Appl., 14 (1966), pp. 141–152.

[5] O. MANGASARIAN, *Duality in nonlinear programming*, Quart. Appl. Math., 20 (1962), pp. 300–302.

[6] ———, *Sufficient conditions for the optimal control of nonlinear systems*, this Journal, 4(1966), pp. 139–152.

[7] B. MOND AND M. A. HANSON, *Duality for variational problems*, J. Math. Anal. Appl., 18(1967), pp. 355–364.

[8] J. D. PEARSON, *Reciprocity and duality in control programming problems*, Ibid., 10(1965), pp. 383–408.

[9] R. J. RINGLEE, *Bounds for convex variational programming problems arising in power system scheduling and control*, IEEE Trans. Automatic Control, AC-10(1965), pp. 28–35.

[10] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, Contributions to the Calculus of Variations 1933–37, University of Chicago Press, Chicago, 1937, pp. 407–488.

[11] P. WOLFE, *A duality theorem for non-linear programming*, Quart. Appl. Math., 19(1961), pp. 239–244.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

CONTROLLABILITY, OBSERVABILITY AND STABILITY OF LINEAR SYSTEMS*

L. M. SILVERMAN† AND B. D. O. ANDERSON‡

1. Introduction. Of the many types of stability which may be defined for dynamical systems, at least two are of special importance when the systems are linear. These are bounded-input bounded-output (BIBO) stability [1] and exponential stability [2]. The aim of this paper is to establish an equivalence between these two types of stability for a large class of linear time-variable systems. The basic system description we shall consider is an *impulse response matrix* H which maps the system inputs u into the system outputs y via the formula

$$y(t) = \int_{t_0}^{t} H(t, \tau)u(\tau)\, d\tau, \tag{1}$$

when the system is in the zero state at time t_0. An alternate description is provided by a set of state equations of the form

$$\dot{x} = Ax + Bu, \tag{2a}$$

$$y = Cx, \tag{2b}$$

where A, B and C are time-variable matrices, and x is the state vector associated with the coordinate basis used in setting up (2). The dimensions of the vectors x, u and y will be taken to be n, r and m, respectively. The well-known [3] relationship between the two representations is that $H(t, \tau) = C(t)\Phi(t, \tau)B(\tau)$ for $t \geqq \tau$, where Φ is the transition matrix of the homogeneous part of (2a).

Recall (see [1], [4]–[6]) that a system of the above type is (zero-state) BIBO stable if and only if there exists a positive constant c_1 such that

$$\int_{-\infty}^{t} \| H(t, \tau) \|\, d\tau \leqq c_1 \quad \text{for all} \quad t, \tag{3}$$

where $\| \cdot \|$ denotes the Euclidean norm. It should be noted that this type of stability is independent of the particular *realization* (2) of H. In con-

* Received by the editors May 18, 1967, and in revised form August 21, 1967. This research was supported in part by the Joint Services Electronics Program (United States Army, United States Navy and United States Air Force) under Grant AF-AFOSR-139-66, and by the National Science Foundation under Grant GK-716.

† Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720.

‡ Department of Electrical Engineering, University of Newcastle, Newcastle, New South Wales, Australia.

trast, exponential stability is a characteristic of the internal structure of the system. As is well known [1], the realization (2) is exponentially stable if and only if there exist positive constants c_2 and c_3 such that

$$\| \Phi(t, \tau) \| \leqq c_2 e^{-c_3(t-\tau)} \quad \text{for all} \quad \tau \quad \text{and for all} \quad t \geqq \tau. \tag{4}$$

In the time-invariant case (A, B and C constant matrices), relations between the two types of stability are well known. Exponential stability implies BIBO stability, while BIBO stability, together with complete controllability [3] and complete observability [3], implies exponential stability.[1] Hence, in synthesizing a time-invariant impulse response matrix one is assured that all minimal realizations will have appropriate stability properties. Unfortunately, no such simple and analogous statements can be made in the time-variable case. Indeed, as was observed by Kalman [7], who first investigated this problem, it is impossible to conclude the existence of any sort of relation between the two types of stability without further constraints on the realizations (2). The reason for this is that one can construct a realization for H with an essentially arbitrary A matrix. If (2) is to represent a practical physical system (e.g., an analogue computer), then a natural restriction is that the elements of the coefficient matrices be bounded functions. Consequently, we shall assume that a constant K exists such that for all t,

$$\| A(t) \| \leqq K, \quad \| B(t) \| \leqq K, \quad \| C(t) \| \leqq K. \tag{5}$$

A system representation satisfying (5) will be termed a *bounded realization*.

Even with the restriction to bounded realizations, complete controllability and observability do not suffice to insure the equivalence of BIBO and exponential stability. It is shown below, however, that a somewhat more stringent, but physically reasonable, set of constraints does provide a connection between the two types of stability. Several important classes of systems which satisfy these constraints are also derived.

2. Uniform controllability and observability. As defined by Kalman [8], the system representation (2) is *uniformly completely controllable* if, for some $\delta_c > 0$, any two of the following three conditions hold for all s (any two imply the third):[2]

$$0 < \alpha_1(\delta_c) I \leqq M(s - \delta_c, s) \leqq \alpha_2(\delta_c) I, \tag{6}$$

$$0 < \alpha_3(\delta_c) I \leqq \Phi(s - \delta_c, s) M(s - \delta_c, s) \Phi'(s - \delta_c, s) \leqq \alpha_4(\delta_c) I, \tag{7}$$

[1] As pointed out by a reviewer, a proof of this widely used result does not seem to exist in the literature. Such a proof is provided here by Theorem 3 and Theorem 4 specialized to time-invariant systems.

[2] If A and B are symmetric matrices, $A > B$ ($A \geqq B$) means $A - B$ is positive (nonnegative) definite.

$$\| \Phi(t, \tau) \| \leqq \alpha_5(\mid t - \tau \mid) \quad \text{for all} \quad t, \tau, \tag{8}$$

where

$$M(s - \delta_c, s) = \int_{s-\delta_c}^{s} \Phi(s, t)B(t)B'(t)\Phi'(s, t)\, dt. \tag{9}$$

The above criteria greatly simplify for bounded realizations. Condition (8) is immediately implied by the bound on A, and it is a routine matter to show that the upper bound in (6) is always satisfied. Hence, we have the following lemma.

LEMMA 1. *A bounded system of the form* (2) *is uniformly completely controllable if and only if there exists* $\delta_c > 0$ *such that for all* s,

$$M(s - \delta_c, s) \geqq \alpha_1(\delta_c)I > 0. \tag{10}$$

Uniform complete observability is defined [8] in a dual [3] manner to the above in terms of the matrix

$$W(s - \delta_0, s) = \int_{s-\delta_0}^{s} \Phi'(t, s - \delta_0)C'(t)C(t)\Phi(t, s - \delta_0)\, dt, \tag{11}$$

so that we need not state the definition explicitly here.

3. Equivalence of BIBO and exponential stability. If the system (2a) with x considered as the output is BIBO stable, it will be said that the system is bounded-input bounded-state (BIBS) stable. We shall first establish an equivalence between BIBS and exponential stability. As a preliminary, we prove the following lemma which gives a useful alternate characterization of uniform complete controllability.

LEMMA 2. *A bounded realization* (2a) *is uniformly completely controllable if and only if there exists* $\delta_c > 0$ *such that for every state* $\xi \in R^n$ *and for any time* s, *there exists an input* u *defined on* $(s - \delta_c, s)$ *such that if* $x(s - \delta_c) = 0$, *then* $x(s) = \xi$ *and* $\| u(t) \| \leqq \gamma(\delta_c, \| \xi \|)$ *for all* $t \in (s - \delta_c, s)$.

Proof. If the system (2) is uniformly completely controllable, then the input $u(t) = B'(t)\Phi'(s, t)M^{-1}(s - \delta_c, s)\xi$ will transfer the system from the zero state at time $s - \delta_c$ to the state ξ at time s. From (5), (6) and (8) it is clear that a constant γ independent of s and t exists such that $\| u(t) \| < \gamma$ for all $t \in (s - \delta_c, s)$.

The converse will be established by contradiction. If the system is not uniformly completely controllable, then Lemma 1 implies that, for each $\delta > 0$ and for any $\alpha > 0$, there is a vector $\lambda \in R^n$, with $\| \lambda \| = 1$, such that for some s, $\lambda' M(s - \delta, s)\lambda < \alpha$, or equivalently, for some s,

$$\int_{s-\delta}^{s} \| \lambda'\Phi(s, \tau)B(\tau) \|^2\, d\tau < \alpha. \tag{12}$$

Suppose that a bounded control u exists which transfers the zero state at time $s - \delta$ to the state λ at time s. Then,

$$\lambda = \int_{s-\delta}^{s} \Phi(s, \tau)B(\tau)u(\tau)\, d\tau,$$

which together with the Schwarz inequality implies

$$\| \lambda \|^2 \leqq \left[\int_{s-\delta}^{s} \| \lambda' \Phi(s, \tau)B(\tau) \|^2 \, d\tau \right]^{1/2} \left[\int_{s-\delta}^{s} \| u(\tau) \|^2 \, d\tau \right]^{1/2}. \tag{13}$$

If $\| u(t) \| < \gamma\ (\delta, 1)$ for all $t \in (s - \delta, s)$ and for all s, then (12) and (13) imply that for some s, $\gamma \sqrt{\alpha\delta} \geqq 1$, a contradiction since α can be made arbitrarily small. This completes the proof.

THEOREM 1. *If* (2a) *is bounded and uniformly completely controllable, then it is BIBS stable if and only if it is exponentially stable.*

Proof. It is well known and straightforward to show [1] that if B is bounded, then exponential stability implies BIBS stability.

To prove the converse, let λ be any unit norm vector in R^n. It follows from Lemma 2 that if (2a) is uniformly completely controllable and bounded, then there exists a $\delta_c > 0$ such that, for all s, an input u exists which satisfies

$$\lambda = \int_{s-\delta_c}^{s} \Phi(s, \tau)B(\tau)u(\tau)\, d\tau, \tag{14}$$

and $\| u(\tau) \| \leqq \gamma_1(\delta_c)$ for $t \in (s - \delta_c\ ,\ s)$. Multiplying both sides of (14) by $\Phi(t, s)$ and integrating the norm of the result yields the inequality

$$\int_{t_0}^{t} \| \Phi(t, s)\lambda \| \, ds \leqq \gamma_1 \int_{t_0}^{t} \left\{ \int_{s-\delta_c}^{s} \| \Phi(t, \tau)B(\tau) \| \, d\tau \right\} ds. \tag{15}$$

Letting $r = \tau - s + \delta_c$, and interchanging the order of integration on the right-hand side of (15), it is then seen that

$$\int_{t_0}^{t} \| \Phi(t, s)\lambda \| \, ds \leqq \gamma_1 \int_{0}^{\delta_c} \left\{ \int_{t_0+r-\delta_c}^{t+r-\delta_c} \| \Phi(t, \tau)B(\tau) \| \, d\tau \right\} dr, \tag{16}$$

and for $0 \leqq r \leqq \delta_c$ it is clear that

$$\int_{t_0+r-\delta_c}^{t+r-\delta_c} \| \Phi(t, \tau)B(\tau) \| \, d\tau \leqq \int_{-\infty}^{t} \| \Phi(t, \tau)B(\tau) \| \, d\tau. \tag{17}$$

Since (2a) is assumed BIBS, it follows from (3) that the right-hand side of (17) is bounded by a constant γ_2 so that (15)–(17) imply

$$\int_{-\infty}^{t} \| \Phi(t, s)\lambda \| \, ds \leqq \gamma_1 \gamma_2 \delta_c \quad \text{for all} \quad t. \tag{18}$$

Hence, if the supremum of (18) over all $\| \lambda \| = 1$ is taken, the bound

$$\int_{-\infty}^{t} \| \Phi(t, s) \| \, ds \leqq \gamma_1 \gamma_2 \delta_c \quad \text{for all} \quad t \tag{19}$$

is obtained. But (19) together with the bound (5) on A suffices to imply exponential stability [4], [9]. This completes the proof.

To complement the above theorem, we now relate BIBO and BIBS stability.

THEOREM 2. *If* (2) *is bounded and uniformly completely observable, then it is BIBO stable if and only if it is BIBS stable.*

Proof. Suppose that BIBO stability does not imply BIBS stability, i.e., there exists a bounded input u which produces both a bounded output and an unbounded state. Then, corresponding to an arbitrary positive number N, there is a value of time $s - \delta_0$ for which $\| x(s - \delta_0) \| > N$. Set u equal to zero over the interval $(s - \delta_0, s)$. Then the output y over this interval is given by $y(t) = C(t)\Phi(t, s - \delta_0)x(s - \delta_0)$. Consequently, using the dual of (10),

$$\int_{s-\delta_0}^{s} y'(t)y(t) \, dt = x'(s - \delta_0)W(s - \delta_0, s)x(s - \delta_0) \geqq \beta_1(\delta_0)N^2.$$

Hence, at some point t in $(s - \delta_0, s)$, $\| y(t) \| > N\sqrt{\beta_1/\delta_0}$. But since N is arbitrary, while u is bounded this contradicts the assumption of BIBO stability. This completes the proof of the theorem, since it is obvious that BIBS implies BIBO stability if C is bounded.

Following immediately from Theorems 1 and 2 is the main result, as given in the following theorem.

THEOREM 3. *If* (2) *is bounded, uniformly completely controllable and uniformly completely observable, then it is BIBO stable if and only if it is exponentially stable.*

A valid question at this point is whether the boundedness constraint of Theorem 3 is essential to the conclusion. It is clear that the constraint on the matrix A can be relaxed since (8) holds under somewhat weaker conditions [8] than (5). However, as shown by the following example, the constraints on B and C are essential.

Example. Consider the system realization $x = -x + u$, $y = gx$, where $g(t) = k$ for $t \in (k, k + (1/k))$, $k = 1, 2, \cdots$, and is zero elsewhere. It is easily verified that this system is uniformly completely controllable and observable, yet it is simultaneously exponentially stable and BIBO unstable.

4. Periodic systems. Periodic systems (A, B and C periodic with the same period) are an important subclass of linear systems. It is shown below that minimal (completely controllable and observable) periodic sys-

tems are uniformly completely controllable and observable. This together with Theorem 3 establishes the apparently known [10] but previously unproven fact that BIBO and exponential stability are equivalent in periodic systems.

THEOREM 4. *If* (2) *is periodic, then it is uniformly completely controllable* (*observable*) *if and only if it is completely controllable* (*observable*).

Proof. If (2) is completely controllable, there must exist a finite $\sigma > 0$ such that $M(0, \sigma) \geqq \epsilon I > 0$. Let k be a positive integer such that $kT > \sigma$, where T is the period of the matrices A, B and C. Clearly, for $s \in (kT, 2kT)$, $M(s - 2kT) \geqq \epsilon I$. It is easily verified, however, that $M(s - 2kT, s)$ is periodic in s with period T. Hence, $M(s - 2kT, s) \geqq \epsilon I$ for all s. By Theorem 3, therefore, (2) is uniformly completely controllable. Since the converse is obviously true, this completes the proof.

5. Classes of uniformly completely controllable systems. In order to apply the results of the previous sections in stability analysis or system synthesis, it is useful to have criteria for uniform complete controllability which do not require calculation of the transition matrix.[3] Such criteria are derived below, and it is shown that several broad classes of systems have the uniform complete controllability property. A basic tool in this development is the following lemma establishing the invariance of uniform complete controllability under bounded state-variable feedback of the form $u = Gx + Fv$, where v is the input to the closed loop system.

LEMMA 3.[4] *A bounded realization* (2) *is uniformly completely controllable if and only if the system* $(A + BG, BF, C)$ *is uniformly completely controllable, where* G *is any* $r \times n$ *bounded matrix and* F *is any* $r \times r$ *bounded matrix whose inverse is also bounded.*

Proof. Let (2) be uniformly completely controllable. Then by Lemma 2 there are a $\delta > 0$ and an input u_1 which takes $x(s - \delta) = 0$ to $x(s) = \xi$, such that $\| u_1(t) \| \leqq \gamma(\delta, \|\xi\|)$ for all $t \in (s - \delta, s)$ and for all s. It is readily verified that if $v_1(t) = F^{-1}u_1(t) - Gx_1(t)$ is the input to $(A + BG, BF, C)$, where x_1 is the trajectory in (A, B, C) due to u_1, then $z_1(s - \delta) = 0$ and $z_1(s) = \xi$, where z_1 is the trajectory of $(A + BG, BF, C)$ due to v_1 (in fact, $z_1(t) = x_1(t)$ for all $t \in (s - \delta, s)$). Since for all $t \in (s - \delta, s)$,

$$\| v_1(t) \| \leqq \| F^{-1}(t) \| \| u_1(t) \| + \| G(t) \| \int_{s-\delta}^{s} \| \Phi(t, \tau)B(\tau)u_1(\tau) \| \, d\tau,$$

it is easily shown that $\| v_1(t) \| \leqq \gamma_1(\delta, \|\xi\|)$. Hence, by Lemma 2, $(A + BG, BF, C)$ is uniformly completely controllable.

[3] Such criteria are also applicable in other problems [8] which involve uniform complete controllability.

[4] The proof of this lemma is based on an argument of Brockett [11] used in proving the invariance of complete controllability in time-invariant systems under time-invariant state-variable feedback.

The converse follows by a similar argument.

Remark. Lemma 3 can be applied directly to a class of problems studied extensively in recent years—stability analysis of a constant linear system with bounded time-variable feedback from output to input. If G is a bounded $r \times m$ matrix, then such a system has the form $(A + BGC, B, C)$, where (A, B, C) is a time-invariant completely controllable and observable system. It follows immediately from Lemma 3 and its dual that the closed loop system is uniformly completely controllable and observable so that BIBO and exponential stability are equivalent in this class of systems. Consequently, only one of the two types of stability need be examined and several existing results can be strengthened. For example, a recent criterion for Lyapunov instability given by Brockett and Lee [12, Theorem 1] extends to a criterion for BIBO instability.

THEOREM 5. *If A and B are bounded and B contains an $n \times n$ submatrix $\tilde{B}$ whose inverse is also bounded, then* (2a) *is uniformly completely controllable.*

Proof. Without loss of generality, we may take $\tilde{B} = B$. Letting $G = -AB^{-1}$ and $F = B^{-1}$, we obtain the time-invariant closed loop system (O, I, C). The result then follows from Lemma 3.

A corollary to the above is the well-known result of Perron [4], [9] that BIBS and exponential stability are equivalent in systems satisfying the hypothesis of the theorem.

The constraint on B in Theorem 5 is quite restrictive. A much weaker condition under which the result holds will now be presented for single-input systems ($B = b$ in (2)). Let $Q_c = [p_0\ p_1 \cdots p_{n-1}]$, where $p_0 = b$ and $p_{k+1} = -Ap_k + \dot{p}_k$, $k = 1, 2, \cdots$. In terms of this controllability matrix [13] we have the following theorem.

THEOREM 6. *If* (2a) *is a bounded, single-input realization and Q_c is a Lyapunov transformation*[5] [5], *then the system is uniformly completely controllable.*

Proof. Let λ be an arbitrary constant vector, and let $g(s, \tau) = \lambda'\Phi(s, \tau)b(\tau)$. Also, let

$$\bar{M}(s - \delta, s) = \int_{s-\delta}^{s} \Phi(s, \tau)Q_c(\tau)Q_c'(\tau)\Phi'(s, \tau)\, d\tau.$$

It is easily shown that

$$\frac{\partial^i}{\partial \tau^i} g(s, \tau) = \lambda'\Phi(s, \tau)p_i(\tau),$$

so that

$$\lambda'\bar{M}(s - \delta, s)\lambda = \sum_{i=0}^{n-1} \int_{s-\delta}^{s} \left[\frac{\partial^i}{\partial \tau^i} g(s, \tau)\right]^2 d\tau. \tag{20}$$

[5] For time-invariant systems, this condition on Q_c is equivalent to complete controllability.

It can also be shown [14] that for all s, each element of $\Phi(s, t)b(t)$, and hence $g(s, t)$, is a solution of the differential equation

$$z^{(n)}(t) + \sum_{i=0}^{n-1} a_i(t) z^{(i)}(t) = 0, \tag{21}$$

where $[a_0\, a_1 \cdots a_{n-1}]' = -Q_c^{-1} p_n$. By virtue of the assumptions on A and Q_c, the coefficients $a_i(t)$ are bounded for all t, so that the following inequality holds [14]:

$$\int_{s-\delta}^{s} \left[\frac{\partial^i}{\partial \tau^i} g(s, \tau)\right]^2 d\tau \leqq K_1 \int_{s-\delta}^{s} g^2(s, \tau)\, d\tau \quad \text{for} \quad 1 \leqq i \leqq n, \tag{22}$$

where K_1 is a constant which depends only on δ. From (20) and (22), therefore, it follows that

$$\lambda' M(s - \delta, s)\lambda \geqq \frac{1}{nK_1} \lambda' \bar{M}(s - \delta, s)\lambda. \tag{23}$$

Since the system (A, Q_c, C) satisfies the hypothesis of Theorem 5, (A, B, C) must be uniformly completely controllable.

A second class of uniformly completely controllable systems is delineated by the following theorem, the proof of which is an immediate consequence of Lemma 3.

THEOREM 7. *The (phase-variable) canonical form*

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{24}$$

where the coefficients $a_i(t)$ are bounded for all t, is uniformly completely controllable.

Theorem 7 implies that BIBS and exponential stability are equivalent in systems represented in phase-variable canonical form (this result was established previously in [15]). It should be noted that any representation which can be transformed to this form via a Lyapunov transformation also has this property. A general method for calculating a transformation to phase-variable form was given in [16], and it is clear from the form of this transformation that with some additional constraints on the derivatives of the matrices A and b, the classes of systems considered in Theorems 6 and 7 are equivalent. Without such constraints, however, they are distinct.

An interesting corollary to Theorems 6 and 7 is the following for systems

represented in the form

$$y^{(n)} + \sum_{i=0}^{n-1} a_i y^{(i)} = u, \tag{25}$$

where the $a_i(t)$ are bounded for all t, $0 \leqq i \leqq n-1$.

COROLLARY. *If the system represented by* (25) *is BIBO stable, then there exist positive constants* c_1 *and* c_2 *such that for any solution* y *of the homogeneous part of* (25),

$$\| \bar{y}(t) \| \leqq c_1 \| \bar{y}(t_0) \| e^{-c_2(t-t_0)} \tag{26}$$

for all $t \geqq t_0$, *where* $\bar{y} = [y\ y^{(1)} \cdots y^{(n-1)}]$.

Proof. If we let $x_i = y^{(i)}$, $0 \leqq i \leqq n-1$, then (25) has the state representation (24), with $y = [1\ 0 \cdots 0]x$. From Theorem 7, this representation is uniformly completely controllable, and from the dual version of Theorem 6 it is uniformly completely observable. Hence, by Theorem 3, the result (26) follows.

A weaker version of the above corollary was established by Kaplan [17, Chap. 8, Theorem 25]. He showed that (26) holds under the more restrictive condition that the impulse response matrix of (25) is exponentially bounded.

In conclusion, we note that Theorems 4–7 are applicable to the synthesis of impulse response matrices. Under appropriate conditions [18], H can be realized as a member of one of the classes discussed above. Thus the internal stability of the corresponding physical realizations is guaranteed, if H represents a BIBO stable system.

REFERENCES

[1] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

[2] N. P. BHATIA, *On exponential stability of linear differential systems*, this Journal, 2 (1965), pp. 181–191.

[3] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[4] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design via the 'second method' of Lyapunov. I. Continuous-time systems*, Trans. ASME Ser. D. J. Basic Engrg., 82 (1960), pp. 371–393.

[5] D. C. YOULA, *On the stability of linear systems*, IEEE Trans. Circuit Theory, CT-10 (1963), pp. 276–279.

[6] C. A. DESOER AND A. J. THOMASIAN, *A note on zero-state stability of linear systems*, Proc. 1st Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1963, pp. 50–52.

[7] R. E. KALMAN, *On the stability of time-varying linear systems*, IRE Trans. Circuit Theory, CT-9 (1962), pp. 420–422.

[8] ———, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102–119.

[9] O. PERRON, *Die Stabilitätsfrage der Differentialgleichungen*, Math. Z., 32 (1930), pp. 703–728.

[10] R. W. BROCKETT, *The status of stability theory for deterministic systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 596–606.

[11] ———, *Poles, zeros and feedback; state space interpretation*, Ibid., AC-10 (1965), pp. 129–134.

[12] R. W. BROCKETT AND H. B. LEE, *Frequency-domain instability criteria for time-varying and nonlinear systems*, Proc. IEEE, 55 (1967), pp. 604–618.

[13] L. M. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–73.

[14] L. H. HAINES AND L. M. SILVERMAN, *Internal and external stability of linear systems*, J. Math. Anal. Appl., to appear.

[15] B. D. O. ANDERSON, *Stability properties of linear systems in phase-variable form*, in preparation.

[16] L. M. SILVERMAN, *Transformation of time-variable systems to canonical (phase-variable) form*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 300–303.

[17] W. KAPLAN *Operational Methods for Linear Systems*, Addison-Wesley, Reading, Massachusetts, 1962.

[18] L. M. SILVERMAN, *Stable realization of impulse response matrices*, IEEE International Convention Record, 15 (1967), pp. 32–36.

SIAM J. CONTROL
Vol. 6, No. 1, 1968
Printed in U.S.A.

A COUNTEREXAMPLE IN STOCHASTIC OPTIMUM CONTROL*

H. S. WITSENHAUSEN†

Abstract. It is sometimes conjectured that nothing is to be gained by using nonlinear controllers when the objective is to minimize the expectation of a quadratic criterion for a linear system subject to Gaussian noise and with unconstrained control variables.

In fact, this statement has only been established for the case where all control variables are generated by a single station which has perfect memory.

Without this qualification the conjecture is false.

1. Introduction. In a stochastic control problem control actions have to be taken at various instants in time as functions of the data then available. One seeks the functions for which the expected value of the cost, under given noise distributions, is minimized. It is usually assumed that all actions to be taken at a given time are based on the same data and that any data available at time t will still be available at any later time $t' > t$. This situation is the *classical information pattern.*

Considering in particular unconstrained control of linear systems with Gaussian noise and quadratic criteria, it is well known that the search for an optimum can safely be confined to the class of affine (linear plus constant) functions [1]. This is the case for both discrete and continuous time systems, with classical information pattern.

In this paper it is shown that the class of affine functions is not always adequate (complete, in decision theory parlance) when the information pattern is not classical.

A counterexample is presented for which it is established that an optimal design exists and that no affine design is optimal. There does not appear to exist any counterexample involving fewer variables than the one presented here.

The practical importance of nonclassical information patterns is discussed.

2. Problem description.

Original Statement. Let x_0 and v be independent random variables with finite second moments. Consider the following 2-stage stochastic control problem. (All variables are real scalars.)

State equations.

$$x_1 = x_0 + u_1,$$

$$x_2 = x_1 - u_2.$$

* Received by the editors August 7, 1967.

† Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07971.

Output equations.
$$y_0 = x_0,$$
$$y_1 = x_1 + v.$$

Cost function. $k^2(u_1)^2 + (x_2)^2, \quad k^2 > 0.$

Admissible controllers.
$$u_1 = \gamma_1(y_0),$$
$$u_2 = \gamma_2(y_1),$$

where (γ_1, γ_2) is any pair of Borel functions. The set of such pairs is designated by Γ.

Objective. For any choice of (γ_1, γ_2) the variables u_1 and x_2 become random variables, and since the cost function is nonnegative it has an expectation that is possibly infinite. The problem is to minimize over Γ the expression $E\{k^2(u_1)^2 + (x_2)^2\}$. The information pattern is nonclassical because the value of y_0 is known at the first control stage but not at the second.

It will be shown that for x_0 and v Gaussian and suitable parameter values the best affine controller is not optimal over Γ.

Restatement. Denoting x_0 by x, γ_2 by g and letting f be defined by $f(x) = x + \gamma_1(x)$, the problem amounts to minimizing, over the set Γ of all pairs of Borel functions (f, g), the expression

$$(1) \qquad J(f, g) = E\{k^2(x - f(x))^2 + (f(x) - g(f(x) + v))^2\},$$

where $k^2 > 0$. Without loss of generality one may assume

$$(2) \qquad E\{x\} = E\{v\} = 0, \quad E\{v^2\} = 1.$$

This reduction amounts to ordinate shifts of f and g, abscissa shift of g and rescaling. The case $E\{v^2\} = 0$ is trivial. Problem $\pi(k^2, F)$ is the problem of minimizing (1) with v Gaussian subject to (2) and x having the distribution function F subject to (2) and $0 < E\{x^2\} \equiv \sigma^2 < \infty$. Problem $\pi(k^2, \sigma^2)$ is the special case of problem $\pi(k^2, F)$ with F the Gaussian distribution with zero mean and variance σ^2.

3. Existence of an optimum for problem $\pi(k^2, F)$.

LEMMA 1. (a) $J^* = \inf \{J(f, g) \mid (f, g) \in \Gamma\}$ *satisfies* $0 \leqq J^* \leqq \min(1, k^2\sigma^2)$.

(b) *If* $(f, g) \in \Gamma$, *then there exists* $(f_1, g_1) \in \Gamma$ *such that* $E\{f_1(x)\} = 0$, $E\{(x - f_1(x))^2\} \leqq \sigma^2$, $J(f_1, g_1) \leqq J(f, g)$ *and* $E\{f_1^2(x)\} \leqq 4\sigma^2$.

Proof. (a) For $f \equiv 0$, $g \equiv 0$ one has $J(f, g) = k^2E\{x^2\} = k^2\sigma^2$, while for $f(x) \equiv x$, $g(y) \equiv y$ one has $J(f, g) = E\{v^2\} = 1$.

(b) If $E\{(x - f(x))^2\} > \sigma^2$ so that $J(f, g) \geqq k^2E\{(x - f(x))^2\} > k^2\sigma^2$, then $f_1 \equiv g_1 \equiv 0$ satisfies all requirements. If $E\{(x - f(x))^2\} \leqq \sigma^2$, then

$E\{f^2(x)\} \leqq 4\sigma^2$, so that $m = E\{f(x)\}$ exists. Let $f_1(x) \equiv f(x) - m$, $g_1(y) \equiv g(y + m) - m$. Then $E\{f_1(x)\} = 0$, $E\{(x - f_1(x))^2\} = E\{(x - f(x))^2\} - m^2 \leqq \sigma^2$, hence $E\{f_1^2(x)\} \leqq 4\sigma^2$ and $J(f_1, g_1) = J(f, g) - k^2m^2 \leqq J(f, g)$.

Hence one need only consider pairs (f, g) for which $f(x)$ has zero mean and variance not exceeding $4\sigma^2$. For such f we now select $g = g_f{}^*$ to minimize $J(f, g)$ for fixed f.

With $\varphi(x) \equiv (2\pi e^{x^2})^{-1/2}$ define

$$D_f(y) = \int \varphi(y - f(x))\, dF(x),$$

$$N_f(y) = \int f(x)\varphi(y - f(x))\, dF(x),$$

$$g_f{}^*(y) = N_f(y)/D_f(y),$$

$$J_2{}^*(f) = J(f, g_f{}^*).$$

First we recall a well-known fact.

LEMMA 2. *Let μ be a measure and h a measurable function. Consider the integral*

$$H(s) = \int_{-\infty}^{+\infty} \varphi(s - t)h(t)\, d\mu(t).$$

Then the set of real values of s for which the integral is finite is convex and H is analytic on the interior of this set.

Proof. Since $\varphi(s - t) = \sqrt{2\pi}\varphi(s)e^{st}\varphi(t)$, one can interpret H as

$$H(s) = \sqrt{2\pi}\varphi(s) \int_{-\infty}^{+\infty} e^{st}\varphi(t)h(t)\, d\mu(t).$$

The claim then follows from the properties of convergence strips of two-sided Laplace transforms.

LEMMA 3. *Assume $E\{f^2(x)\} < \infty$. Then*

(a) *N_f, D_f, $g_f{}^*$ are analytic with $D_f > 0$;*

(b) *D_f is a density of the random variable $y \equiv f(x) + v$;*

(c) *$g_f{}^*(y) = E\{f(x) \mid y\}$ a.s.;*

(d) *$J_2{}^*(f) = \min\{J(f, g) \mid g \text{ Borel}\}$;*

(e) *$dg_f{}^*(y)/dy = \text{var}\{f(x) \mid y\} \geqq 0$;*

(f) *$J_2{}^*(f) - k^2E\{(x - f(x))^2\} = E\{\text{var}\{f(x) \mid y\}\}$*
$= E\{f^2(x)\} - E\{g_f{}^{*2}(y)\}$
$= 1 - I(D_f)$,

where

$$I(D_f) = \int \left(\frac{d}{dy} D_f(y)\right)^2 \frac{dy}{D_f(y)}$$

$$= \int \frac{dD_f(y)}{dy}\, d\log D_f(y)$$

$$= 4 \int \left(\frac{d}{dy}\sqrt{D_f(y)}\right)^2 dy$$

is the Fisher information of the random variable y;

(g) $\max(0, 1 - E\{f^2(x)\}) \leqq I(D_f) \leqq 1$, *and for* $E\{(x - f(x))^2\} \leqq \sigma^2$ *one has* $J_2^*(f) \leqq k^2\sigma^2 + \min(1, 4\sigma^2)$.

Proof. (a) For each y the integrands $\varphi(y - z)$ and $z\varphi(y - z)$, with $z = f(x)$, are bounded, hence the integrals defining N_f and D_f exist for all y. By Lemma 2, N_f and D_f are analytic. Since φ is strictly positive, so is D_f, hence g_f^* is analytic.

(b) The joint distribution of y and x is defined by

$$\varphi(y - f(x))\, dy\, dF(x)$$

because the measurable transformation $\begin{pmatrix} x \\ v \end{pmatrix} \leftrightarrow \begin{pmatrix} x \\ y \end{pmatrix}$ with $y = f(x) + v$ is measure preserving by Cavalieri's principle (though a Jacobian does not exist for general f). Hence the marginal distribution of y has density D_f.

(c) Since $f(x)$ has finite second moment, its conditional expectation exists. With the joint distribution of x and v as in (b), (c) is immediate.

(d) This states the quadratic minimization property of expectations.

(e) and (f). These follow by simple manipulations. Note that

$$N_f(y) = yD_f(y) + \frac{d}{dy} D_f(y);$$

hence,

$$g_f^*(y) = y + \frac{d}{dy} \log D_f(y).$$

(g) This follows as in Lemma 1.

The problem is thus reduced to the minimization of

$$J_2^*(f) = k^2 E\{(x - f(x))^2\} - I(D_f) + 1$$

over all Borel functions (or only those of zero mean and variance $\leqq 4\sigma^2$).

The designer is trying to find a compromise between (i) keeping the cost of the first stage correction small, and (ii) making the Fisher information of the observation available at the second stage large.

The difficulty is that J_2^* is not a convex functional.

Now for f_0 and $g_{f_0}^*$ as in Lemma 3 we attempt to minimize $J(f, g_{f_0}^*)$ over f for fixed $g_{f_0}^*$.

LEMMA 4. *Let* P *be the distribution of a real random variable. Let* α_1 *be the set of all points* x *for which both* $(-\infty, x]$ *and* $[x, +\infty)$ *have positive*

probability. Let α_2 be the set obtained by removing from the convex hull of the support of P those boundary points which are not atoms. Let α_3 be the intersection of all convex sets of probability one. Then $\alpha_1 = \alpha_2 = \alpha_3 \equiv \alpha(P)$, the smallest convex set of probability one.

Proof. (i) $\alpha_1 \supset \alpha_2$: If x belongs to the interior of α_2, both $[x, \infty)$ and $(-\infty, x]$ have positive probability. If x is a boundary point of α_2, then it is an atom, hence belongs to α_1.

(ii) $\alpha_2 \supset \alpha_3$: By construction α_2 is a convex set of probability one.

(iii) $\alpha_3 \supset \alpha_1$: If E is a convex set and x a point in α_1 but not in E, then E is disjoint from one of the sets $(-\infty, x]$, $[x, +\infty)$, and thus E has probability less than one. Hence all convex sets of probability one contain α_1.

Note that in two (or more) dimensions the intersection of all convex sets of probability one may have probability zero, because the boundary of a nontrivial convex set is uncountable.

LEMMA 5. *For f and $g_f{}^*$ as in Lemma 3 let P be the distribution of the random variable $f(x)$. Then the range of $g_f{}^*$ is contained in $\alpha(P)$.*

Proof. By contradiction suppose that for some y the set $[g(y), \infty)$ (or $(-\infty, g(y)]$) has probability zero under P. Then

$$\begin{aligned} g(y)D_f(y) = N_f(y) &= \int dF(x)f(x)\varphi(y - f(x)) \\ &= \int dP(z)z\varphi(y - z) \\ &= \int_{(-\infty, g(y))} dP(z)z\varphi(y - z) \\ &< g(y) \int_{(-\infty, g(y))} dP(z)\varphi(y - z) \\ &= g(y)D_f(y), \end{aligned}$$

which is a contradiction.

LEMMA 6. *For f_0 and $g^*_{f_0}$ as in Lemma 3, fixed, one has*

$$J(f, g^*_{f_0}) = \int dF(x)[k^2(x - f(x))^2 + K(f(x))],$$

where K is a nonnegative analytic function.

Proof. To shorten notation let $g = g^*_{f_0}$ and let P be the distribution of the random variable $f_0(x)$. One has

$$\begin{aligned} J(f, g) &= \int dF(x) \left[k^2(x - f(x))^2 + \int dv\varphi(v)(f(x) - g(f(x) + v))^2 \right] \\ &= \int dF(x)[k^2(x - f(x))^2 + K(f(x))] \end{aligned}$$

with

$$K(z) = \int dv\varphi(v)(z - g(z + v))^2$$

$$= \int dy\varphi(y - z)(z - g(y))^2.$$

Since the integrands are nonnegative, the above formulas are valid whether the integrals are finite or not. Let β be the set $\{z \mid K(z) < \infty\}$. Because of the inequalities

$$g^2(y) \leqq 2z^2 + 2(z - g(y))^2,$$

$$(z - g(y))^2 \leqq 2z^2 + 2g^2(y),$$

the set β coincides with the set of all z for which

$$\int dy\varphi(y - z)g^2(y) < \infty.$$

By Lemma 2 the set β is thus convex. By construction of g,

$$J_2{}^*(f_0) = J(f_0, g) < \infty,$$

and therefore,

$$\int dF(x)K(f_0(x)) = \int dP(z)K(z) < \infty.$$

Hence the set β has probability one under P. Since it is convex, β contains $\alpha(P)$ defined in Lemma 4 and, by Lemma 5, $\alpha(P)$ contains the range of g. Also by Lemma 3(e) g is monotone nondecreasing.

This author claims that $\beta = (-\infty, +\infty)$; indeed otherwise by convexity of β at least one of the inequalities $-\infty < \inf \beta$, $\sup \beta < \infty$ holds. If both hold, g is bounded which implies $\beta = (-\infty, +\infty)$. If $\inf \beta = -\infty$, $\sup \beta < \infty$, then

$$\int_{-\infty}^{0} dy\varphi(y - z)g^2(y) = \sqrt{2\pi}\varphi(z) \int_{-\infty}^{0} dye^{yz}\varphi(y)g^2(y)$$

converges for $z < \sup \beta$ by the assumption on β and a fortiori for $z \geqq \sup \beta$. But, for all z,

$$\int_{0}^{\infty} dy\varphi(y - z)g^2(y) < \infty$$

because for $y > 0$, g is bounded, according to Lemmas 3(e) and 5, by

$$g(0) \leqq g(y) \leqq \sup \beta.$$

Hence $\beta = (-\infty, +\infty)$ and a symmetric argument applies for $-\infty < \inf \beta$, $\sup \beta = \infty$.

In conclusion, $\int dy\varphi(y - z)g^2(y)$ is finite for all z, and a fortiori $\int dy\varphi(y - z)g(y)$ is finite for all z. By Lemma 2, both these integrals are analytic in z. Therefore,

$$K(z) = z^2 - 2z \int dy\varphi(y - z)g(y) + \int dy\varphi(y - z)g^2(y)$$

is analytic.

LEMMA 7. *For* $E\{(f_0(x) - x)^2\} \leqq \sigma^2$ *and* $g = g^*_{f_0}$ *as in Lemma* 3, *there exists a function* f^*, *monotone nondecreasing on* $\alpha(F)$, *such that*

(a) $J(f^*, g) = \min \{J(f, g) \mid f \text{ Borel}\}$,

(b) $|f^*(x)| < c(x)$ *for* x *in* $\alpha(F)$, *where the real-valued function* c *depends only on* F *and* k^2, *not on* f_0.

Proof. For each x the function

$$V_x(z) = k^2(x - z)^2 + K(z)$$

is nonnegative, continuous (by Lemma 6) and radially unbounded (because $K \geqq 0$). Hence it attains its minimum on a nonempty compact set. For each x define $f^*(x)$ as one of the minimizing values of z (e.g., the largest). Then for any x and x',

$$V_x(f^*(x)) \leqq V_x(f^*(x'))$$

and

$$V_{x'}(f^*(x')) \leqq V_{x'}(f^*(x)).$$

Adding these inequalities gives

$$(x - x')(f^*(x) - f^*(x')) \geqq 0.$$

Hence the function f^* is monotone nondecreasing and a fortiori Borel.

$$V_x(f^*(x)) \leqq V_x(f(x))$$

for all x in $\alpha(F)$, which implies

$$\int dF(x)V_x(f^*(x)) \leqq \int dF(x)V_x(f(x))$$

or

$$J(f^*, g) \leqq J(f, g),$$

so that f^* is optimal for fixed g. In particular, $J(f^*, g) \leqq J(f_0, g) = J_2^*(f_0) \leqq k^2\sigma^2 + \min(1, 4\sigma^2)$, a constant independent of f_0.

Hence $E\{f^{*2}(x)\} \leqq a$, where a is a constant. Then for $x \in \alpha(F)$,

$$-\left(\frac{a}{F((-\infty, x])}\right)^{1/2} \leqq f^*(x) \leqq \left(\frac{a}{F([x, +\infty))}\right)^{1/2}.$$

Indeed, if $f^*(x) > (a/F([x, +\infty)))^{1/2}$, then

$$\int dF(\xi) f^{*2}(\xi) \geqq \int_{[x,+\infty)} dF(\xi) f^{*2}(\xi)$$

$$> \frac{a}{F([x, +\infty))} \int_{[x,+\infty)} dF(\xi) = a,$$

and similarly for the lower bound.

One also needs a special form of Helly's selection theorem.

LEMMA 8. *Let S be a convex set of reals and f_n a sequence of monotone nondecreasing functions on S. Assume that, for all n and all x in S, $|f_n(x)| \leqq c(x) < \infty$. Then there exists a subsequence which converges pointwise on S to a monotone nondecreasing function f.*

Proof. Because at each x in S the numerical sequence $f_n(x)$ is bounded, there exists a subsequence converging at that value of x. Given a countable subset of S a subsequence converging on it can be formed by the diagonal process. Let S_0 be the set of rational points in S. It is countable, hence we may assume that f_n is a subsequence converging on S_0, reindexed. Then $\limsup f_n(x)$ is a monotone nondecreasing function to which, by monotony, the sequence f_n converges at all points of continuity interior to S. Since the points of discontinuity of a monotone function are countable and the number of boundary points of S belonging to S is at most two, a second application of the diagonal process yields a subsequence converging on S.

THEOREM 1. *For any $k^2 > 0$ and any distribution F the problem $\pi(k^2, F)$ has an optimal solution.*

Proof. Let $(f_n^{(0)}, g_n^{(0)})$ be a minimizing sequence in Γ, that is,

$$\lim_{n\to\infty} J(f_n^{(0)}, g_n^{(0)}) = J^* = \inf \{J(f, g) \mid f, g \text{ Borel}\}.$$

Observe that $J(f, g)$ depends only upon f through its restriction to $\alpha(F)$. Henceforth we shall only consider functions f so restricted. Observe also that when the construction of Lemma 1(b) is applied to a pair (f, g), where f is monotone on $\alpha(F)$, the resulting function f_1 is monotone on $\alpha(F)$.

For each value of n replace $(f_n^{(0)}, g_n^{(0)})$ by $(f_n^{(1)}, g_n^{(1)})$ according to Lemma 1(b). Then replace by $(f_n^{(1)}, g_n^{(2)})$ with $g_n^{(2)} = g^*_{f_n^{(1)}}$ according to Lemma 3. Then replace by $(f_n^{(2)}, g_n^{(2)})$, where $f_n^{(2)}$ is optimal versus $g_n^{(2)}$ and monotone as by Lemma 7. Then replace by $(f_n, g_n^{(3)})$ according to Lemma 1(b) noting that f_n is still monotone. Then replace by (f_n, g_n), where $g_n = g^*_{f_n}$ according to Lemma 3.

Then

$$J(f_n, g_n) = J_2^*(f_n) \leqq J(f_n, g_n^{(3)}) \leqq J(f_n^{(2)}, g_n^{(2)}) \leqq J(f_n^{(1)}, g_n^{(2)}) \leqq J(f_n^{(1)}, g_n^{(1)}) \leqq J(f_n^{(0)}, g_n^{(0)}),$$

and therefore the sequence (f_n, g_n) is a fortiori a minimizing sequence, that is,

$$\lim J_2^*(f_n) = J^*.$$

By Lemmas 7 and 8 there exists a subsequence f_{n_k} converging to a limit f pointwise on $\alpha(F)$. Relabel (f_{n_k}, g_{n_k}) as (f_n, g_n). By Fatou's lemma,

$$E\{f^2(x)\} \leqq \liminf E\{f_n^2(x)\} \leqq 4\sigma^2.$$

Let $g = g_f^* = N_f/D_f$. For each y the functions $\varphi(y - z)$ and $z\varphi(y - z)$ are bounded functions of z, and since φ is continuous,

$$\varphi(y - f_n(x)) \to \varphi(y - f(x)),$$

$$f_n(x)\varphi(y - f_n(x)) \to f(x)\varphi(y - f(x)),$$

pointwise in x, for all y.

By the bounded convergence theorem,

$$D_{f_n}(y) = \int_{\alpha(F)} dF(x)\varphi(y - f_n(x)) \to D_f(y) > 0$$

for each y, and similarly,

$$N_{f_n}(y) \to N_f(y).$$

Hence $g_n(y) \to g(y)$ pointwise.

For all x in $\alpha(F)$ and all y the nonnegative expression

$$A_n(x, y) = [k^2(x - f_n(x))^2 + (f_n(x) - g_n(y))^2]\varphi(y - f_n(x))$$

converges to

$$A(x, y) = [k^2(x - f(x))^2 + (f(x) - g(y))^2]\varphi(y - f(x)).$$

By Fatou's lemma,

$$\int_{\alpha(F)} dF(x) \int dy A(x, y) \leqq \liminf_{n\to\infty} \int_{\alpha(F)} dF(x) \int dy A_n(x, y)$$

or

$$J(f, g) \leqq \liminf_{n\to\infty} J(f_n, g_n) = J^*.$$

But, by definition, $J^* \leqq J(f, g)$, hence $J^* = J(f, g)$ and the pair (f, g) is optimal. (Define f as zero outside $\alpha(F)$.)

Note that when $\alpha(F)$ has a (say upper) boundary point b not belonging to $\alpha(F)$ (because b is not an atom), then the function $c(x)$ of Lemma 7 approaches ∞ as $x \to b$ and, in consequence, the function f of Theorem 1 may approach ∞ as $x \to b$. Then a *monotone* real-valued extension of f to $(-\infty, \infty)$ does not exist.

Taking the first variation of ${J_2}^*$ gives, formally,

$$\delta {J_2}^*(f) = \int dF(x) G_f(x) \delta f(x),$$

where

$$G_f(x) = 2k^2(f(x) - x) + \int dy \varphi(y - f(x)) \frac{D_f'(y)}{D_f(y)} \left[2(y - f(x))^2 + \frac{D_f'(y)}{D_f(y)} (y - f(x)) - 2 \right]$$

with

$$D_f'(y) = \frac{d}{dy} D_f(y) = N_f(y) - y D_f(y).$$

Hence one has the following necessary condition.

LEMMA 9. *If f is optimal, then $E\{f(x)\} = 0, E\{f^2(x)\} \leqq 4\sigma^2$, and $G_f(x) = 0$ F-almost surely, provided the formal differentiation holds at least in the sense of Gâteaux for δf in $L_\infty[(-\infty, \infty), F]$.*

This condition is of little use because there are in general many local minima of ${J_2}^*(f)$. Steepest descent in function space can be used to improve a suboptimal solution but not, safely, to find an absolute optimum.

An alternative existence proof can be based on a generalization of Theorem 378 of Hardy, Littlewood and Pólya [2]. All functions f which give the same distribution to $f(x)$ also give the same optimal cost $1 - I(D_f)$ for the second stage. According to the theorem in question, among all these "equimeasurable" functions, the monotone nondecreasing rearrangement maximizes $E\{xf(x)\}$, hence minimizes $E\{(f(x) - x)^2\}$. This establishes the existence of a minimizing sequence $(f_n, g_{f_n}^*)$ with monotone f_n.

4. Optimization of $\pi(k^2, \sigma^2)$ over the affine class. For problem $\pi(k^2, \sigma^2)$ let

$${J_a}^* = \inf \{J(f, g) \mid f, g \text{ affine}\}.$$

Observe that the transformation of (f, g) into (f_1, g_1) in Lemma 1(b) maps the class of affine pairs into itself. Hence one need only consider $E\{f(x)\} = 0$ or

$$f(x) = \lambda x.$$

For such f,

$$g_f^*(y) = \mu y$$

with

$$\mu = \frac{\sigma^2\lambda^2}{1 + \sigma^2\lambda^2}$$

and

$$J_2^*(f) = J_{2a}^*(\lambda) = k^2\sigma^2(1 - \lambda)^2 + \frac{\lambda^2\sigma^2}{1 + \lambda^2\sigma^2}.$$

This expression being nonnegative, analytic and radially unbounded, optimal values of λ exist and are stationary points of J_{2a}^*.

LEMMA 11. *Optimal affine solutions exist and are of the form* $f(x) = \lambda x$, $g(y) = \mu y$, *where*

$$\mu = \frac{\sigma^2\lambda^2}{1 + \sigma^2\lambda^2},$$

and $t = \sigma\lambda$ *is a real root of the equation*

$$(t - \sigma)(1 + t^2)^2 + \frac{1}{k^2}t = 0.$$

Proof. Set $dJ_{2a}^*(\lambda)/d\lambda = 0$.

A great deal of insight is gained by interpreting graphically the condition of Lemma 11. It may be written

$$\frac{t}{(1 + t^2)^2} = k^2(\sigma - t).$$

Hence the stationary points are the abscissas of the points of intersection of the curve

$$s = \frac{t}{(1 + t^2)^2}$$

with the line

$$s = k^2(\sigma - t).$$

The curve is odd and positive for $t > 0$. Since σ and k^2 are positive, all solutions are positive. The curve has a maximum at $t = \sqrt{3}/3$ with value $3\sqrt{3}/16$ and then decays asymptotically to zero with an inflection at $t = 1$, where the value is $\frac{1}{4}$ and the slope $-\frac{1}{4}$.

Hence for $k^2 \geqq \frac{1}{4}$ and any σ there is exactly one root which defines a unique optimum.

For $k^2 < \frac{1}{4}$ and σ sufficiently small there is a unique solution with t small. For σ sufficiently large there is a unique solution with t large. For intermediate values of σ there are 3 solutions corresponding to two local minima of J_{2a}^* separated by a local maximum. There is a value σ_c of σ for which the two local minima are equal, hence both optimal. For $\sigma < \sigma_c$ the lowest root is optimal; for $\sigma > \sigma_c$ the largest root is optimal. Hence for fixed $k^2 < \frac{1}{4}$ the plot of the optimal λ versus σ has a jump at σ_c, though J_a^* is continuous in σ. At, and only at, the jump there are two optimal solutions.

LEMMA 12. *For $k^2 < \frac{1}{4}$ the critical value of σ is $\sigma_c = k^{-1}$. At this value the two optimal solutions are $\lambda = \mu = \frac{1}{2}(1 \pm \sqrt{1 - 4k^2})$, both of which yield $J_a^* = 1 - k^2$.*

Proof. Let $k^2\sigma^2 = 1$, $k^2 < \frac{1}{4}$. Then the stationarity condition is $(t - \sigma)(1 + t^2)^2 + \sigma^2 t = 0$ and can be factored into

$$(t^2 - \sigma t + 1)(t^3 + t - \sigma) = 0.$$

Since the two roots $t = \frac{1}{2}(\sigma \pm \sqrt{\sigma^2 - 4})$ give the same value $1 - k^2$ to J_{2a}^*, they are the two local minima, and the real root of the cubic is the intermediate local maximum. Hence $k^2\sigma^2 = 1$ is the critical condition.

Note that for $k^2\sigma^2 = 1$, $k^2 = \frac{1}{4}$, there is a triple root at the inflection point, and for $k^2\sigma^2 = 1$, $k^2 > \frac{1}{4}$, the only real root is that of the cubic and this is then the optimum.

LEMMA 13. *If a design is optimal in the affine class, it is optimal in the class of all pairs of Borel functions (f, g) of which at least one is affine.*

Proof. If either f or g is affine and fixed, the determination of an optimal choice for the other function is a Gaussian-linear-quadratic single-stage problem with classical information pattern, hence it is an affine function.

Clearly this lemma holds in far more general problems with "at least one" replaced by "all but at most one."

LEMMA 14. *If $f(x) = \lambda x$ and $g(y) = \sigma^2\lambda^2 y/(1 + \sigma^2\lambda^2)$ is stationary (in particular, optimal) over the affine class, then it satisfies the formal conditions of Lemma 9.*

Proof. With $f(x) = \lambda x$, substitution yields

$$G_f(x) = 2\left(k^2(\lambda - 1) + \frac{\lambda}{(1 + \lambda^2\sigma^2)^2}\right)x,$$

which vanishes by the stationarity condition of Lemma 11.

Despite the facts stated in Lemmas 13 and 14, we shall find that $J^* < J_a^*$ is possible.

5. Two-point symmetric distributions. Consider problem $\pi(k^2, F)$ for F the two-point symmetric distribution assigning probability $\frac{1}{2}$ to $x = \sigma > 0$ and $x = -\sigma$.

Let $a = f(\sigma)$. For optimization we may assume by Lemma 1(b) that $f(-\sigma) = -a$ and by Lemma 7 that $a \geqq 0$.

The first stage cost is thus $k^2(a - \sigma)^2$. At the second stage,

$$\begin{aligned} D_f(y) &= \tfrac{1}{2}(\varphi(y - a) + \varphi(y + a)) \\ &= \sqrt{2\pi}\,\varphi(a)\varphi(y) \cosh ay. \end{aligned}$$

Hence,

$$\frac{D_f'(y)}{D_f(y)} = -y + a \tanh ay$$

and

$$\begin{aligned} g_f^*(y) &= a \tanh ay, \\ g_f^{*2}(y) &= a^2 - a^2 \operatorname{sech}^2 ay, \\ E\{g_f^{*2}(y)\} &= a^2 - h(a), \end{aligned}$$

where

$$h(a) = \sqrt{2\pi}\, a^2 \varphi(a) \int \frac{\varphi(y)}{\cosh ay}\, dy.$$

Thus,

$$\begin{aligned} J_2^*(f) &= k^2 E\{(x - f(x))^2\} + E\{f^2(x)\} - E\{g_f^{*2}(y)\} \\ &= k^2(a - \sigma)^2 + h(a). \end{aligned}$$

This is a radially unbounded analytic function of a, and therefore attains a minimum $J^* = V_k(\sigma)$ at one or more optimal values of a. Any optimal value must satisfy the transcendental equation

$$k^2(\sigma - a) = -\tfrac{1}{2}h'(a).$$

A plot of $-\frac{1}{2}h'(\sigma)$ is similar in shape to the plot of $t/(1 + t^2)^2$ which occurred in the optimization of the Gaussian case over the affine class. Hence the qualitative discussion of that case applies also in the present instance. The possible appearance of two local minima has now a simple interpretation. For small k^2 and appropriate σ one policy is to bring a close to zero by means of f so that the second stage will have little work to do; another policy is to make a larger than σ, creating a vast gap between a and $-a$, so that the second stage can almost infallibly separate these two values.

In summary one has the following lemma.

LEMMA 15. *When F is the two-point symmetric distribution with variance σ^2, then the design $f(x) = (a/\sigma)x$, $g(y) = a \tanh ay$ is optimal for an appropriate constant a which gives the minimum in the formula*

$$J^* = V_k(\sigma) \equiv \min_a [k^2(a - \sigma)^2 + h(a)].$$

Note that the functions $h(a)$, $h'(a)$ and $V_k(\sigma)$ can be obtained by computer programs with relative ease. Note also that, for the general problem $\pi(k^2, F)$, whenever $f(x)$ has a two-point symmetric distribution with the values $\pm a$, then the minimum over g of $E\{(f(x) - g(f(x) + v))^2\}$ is $h(a)$.

When $a \gg 1$ (the variance of the noise v), the second stage cost should be close to zero. More precisely one has the following lemma.

LEMMA 16. *The function $h(a)$ is bounded by $\sqrt{2\pi}\, a^2\varphi(a) = a^2 e^{-a^2/2}$.*

Proof.

$$\int \frac{\varphi(y)}{\cosh ay}\, dy \leqq \int \varphi(y)\, dy = 1.$$

6. Nonlinear design for the Gaussian case.

THEOREM 2. *There exist values of the parameters k and σ for problem $\pi(k^2, \sigma^2)$ such that J^*, the optimal cost, is less than J_a^*, the optimal cost achievable in the class of affine designs.*

Proof. Consider the design

$$f(x) = \sigma \operatorname{sgn} x, \quad g(y) = \sigma \tanh \sigma y.$$

For this choice $f(x)$ has a two-point symmetric distribution and $g = g_f^*$. Then

$$J(f, g) = k^2 E\{(x - \sigma \operatorname{sgn} x)^2\} + h(\sigma),$$

where h is the function defined in §5.

The first term is readily evaluated to be

$$2k^2\sigma^2\left(1 - E\left\{\left|\frac{x}{\sigma}\right|\right\}\right) = 2k^2\sigma^2\left(1 - \sqrt{\frac{2}{\pi}}\right).$$

For $k^2\sigma^2 = 1$, by Lemma 16,

$$J(f, g) \leqq 2\left(1 - \sqrt{\frac{2}{\pi}}\right) + \sqrt{2\pi}\,\frac{1}{k^2}\,\varphi\left(\frac{1}{k}\right).$$

As $k \to 0$, the right-hand side approaches $2(1 - \sqrt{2/\pi}) = 0.404230878$, while by Lemma 12, J_a^* approaches 1.

Hence, for small k^2, $J^* \leqq J(f, g) < J_a^*$.

The design of Theorem 2 is far from optimal. Lower values of $J(f, g_f^*)$ for $k^2\sigma^2 = 1$, k^2 small, are obtained by starting with f a $(2n + 1)$-level quantization and then improving this choice by the gradient method in function space.

The optimum, which exists by Theorem 1, is not known.

Computer experimentation suggests that the functional J_2^* has a large (possibly infinite) number of stationary points.

7. A lower bound for the Gaussian case. Since only suboptimal designs for the Gaussian case were found in §6 and these give only upper bounds on J^*, it may be useful to have a loose but positive lower bound on J^*.

Let ξ, u, v be independent random variables: ξ, v Gaussian of zero mean and variances σ^2, 1; u taking the values $+1$ and -1 with probability $\frac{1}{2}$.

Let J_3^* be the infimum, over all pairs (f, g) of Borel functions of two variables, of the expression

$$J_3(f, g) = E\{k^2(u\xi - f(u\xi, \xi))^2 + (f(u\xi, \xi) - g(f(u\xi, \xi) + v, \xi))^2\}.$$

Let $x = u\xi$ and $y = f(u\xi, \xi) + v$; then x is a Gaussian random variable independent of v and distributed like ξ.

Hence for any pair (f_1, g_1) of Borel functions of one variable, the choice

$$f(x, \xi) = f_1(x), \quad g(y, \xi) = g_1(y)$$

is a possible design, for which

$$J_3(f, g) = J(f_1, g_1),$$

where J is the cost functional of problem $\pi(k^2, \sigma^2)$. Hence $J_3^* \leqq J^*$. But

$$J_3(f, g) = E\{E\{\text{expression} \mid \xi\}\},$$

and for fixed ξ the minimization of the conditional expectation is the problem of §5 with the variable σ of that section having the value ξ. Hence for all pairs (f, g) the conditional expectation is, almost surely in ξ, bounded from below by the function $V_k(\xi)$ defined in Lemma 15. This establishes the next theorem.

THEOREM 3. *For problem* $\pi(k^2, \sigma^2)$ *one has*

$$J^* \geqq \frac{1}{\sigma} \int d\xi \varphi(\xi/\sigma) V_k(\xi).$$

Since V_k can be obtained by computer, this bound can be evaluated for any k and σ.

Theorem 3 may be considered a special case of the following observation. Suppose the expected cost, in a stochastic optimization problem with nonnegative cost function, is considered as a function of the design γ and of the distribution F of some of the noise variables. Suppose that the conditional distribution of the other noise variables, given those described by F, is fixed, for instance, because they are independent. Let $K(\gamma, F)$ be this function, with values in $[0, +\infty]$. Then for each γ, K is a linear function of F on the set on which it is finite and is $+\infty$ elsewhere. Therefore $K^*(F) = \inf_\gamma K(\gamma, F)$ satisfies, for all distributions F_1, F_2 and $0 < \theta < 1$, the extended-real number inequality

$$K^*(\theta F_1 + (1 - \theta)F_2) \geqq \theta K^*(F_1) + (1 - \theta)K^*(F_2).$$

In other words, K^* is concave in the extended-real sense. If F is construed as a mixture of distributions F_α under some distribution of α, then by the concavity of K^*, the expectation of $K^*(F_\alpha)$ under α is a lower bound on $K^*(F)$.

In Theorem 3, α is the Gaussian random variable ξ and F_α is the two-point symmetric distribution supported on $\pm\xi$.

8. Physical situations leading to nonclassical information patterns. (a) Nonclassical patterns arise when the controller memory is limited. In particular, one may want to determine an optimal zero-memory controller, that is, one for which each control action depends only upon the most recent output [3].

(b) Whenever the physical system to be controlled is of large size or comprises mobile subsystems, nonclassical patterns appear. Indeed control is then effected from several stations widely separated and in relative motion. Hence the actions applied at a given time-stage by the stations are not based all on the same data, even when each station has perfect memory. Communication links between stations are subject to delay, noise and operating costs. These links should be considered as part of the controlled system and the communication policy as part of the control policy. The nonclassical effects are most likely to be of practical import in such cases, as for control of space missions, air traffic or high-speed ground transportation.

(c) When communications problems are considered as control problems (which they are), the information pattern is never classical since at least two stations, not having access to the same data, are always involved.

If one considers the transmission of Gaussian signals over Gaussian channels with quadratic (power and distortion) criteria, then there is a possibility, in complex cases such as with noisy feedback channels, etc., that the optimum "controller" (i.e., modulator or coder) not be affine.

9. Conclusions. (i) Further study of linear, Gaussian, quadratic control problems with general information patterns appears to be required.

(ii) The existence of an optimum and the question of completeness of the class of affine designs must be examined as a function of the information pattern.

(iii) It would be interesting if a relation could be found between the appearance of several local minima over the affine class and lack of completeness of this class.

(iv) Algorithms for approaching an optimal solution need to be developed. Because of the occurrence of local minima, this appears to be a most difficult task.

REFERENCES

[1] C. STRIEBEL, *Sufficient statistics in the optimum control of stochastic systems*, J. Math. Anal. Appl., 12 (1965), pp. 576–592.

[2] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA, *Inequalities*, 2nd ed., Cambridge University Press, London, 1952.

[3] H. CHERNOFF, *Backward induction in dynamic programming*, Unpublished memorandum, 1963.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

ON A NEW COMPUTING TECHNIQUE IN OPTIMAL CONTROL*

A. V. BALAKRISHNAN†

1. Introduction. Most known methods for computing optimal control laws for dynamic systems involve the solution of the dynamic equations as an essential step. In dealing with distributed parameter systems in particular, the complexity attendant on solving boundary value problems for partial differential equations in addition to the optimal control can often be prohibitively large. Since the primary aim is to compute the optimal control, there would be intrinsic advantage if a method could be devised that would avoid having to solve the dynamic equations explicitly. The possibility of such a method was first brought to the author's attention by J. L. Lions. It turns out, however, that quite apart from computational attractiveness the method can provide a constructive approach to characterizing the optimal solutions and, in particular, to the maximum principle. First results in this direction were announced in [1]. In this paper we give detailed proofs as well as extensions, including infinite-dimensional problems and time-optimal problems.

The particular computing methods that we discuss can be stated quite simply. In reference to a fixed endpoint problem, suppose the dynamics are characterized (in the usual notation) by

$$\dot{x}(t) = f(t; x(t); u(t)), \qquad x(0) = x_0,$$

and it is required to minimize

$$\int_0^T g(t; x(t); u(t))dt,$$

where $x(t)$ is the state and $u(t)$ the control, the latter being subject to constraint conditions C. Instead of solving the dynamic equation, we formulate first a nondynamic problem for each $\epsilon > 0$. Let us call this the *epsilon problem*. We minimize

$$\frac{1}{2\epsilon}\int_0^T \| \dot{x}(t) - f(t; x(t); u(t))\|^2 \, dt + \int_0^T g(t; x(t); u(t)) \, dt$$

over, say, the class of absolutely continuous functions $x(t)$ subject to $x(0) = x_0$, with derivative square integrable over $[0, T]$, and over the class of control functions subject to C. It can be shown (see §3) that the solutions

* Received by the editors January 4, 1968, and in revised form January 29, 1968.

† Department of Engineering, University of California, Los Angeles, California 90024. This research was supported in part by the United States Air Force Office of Scientific Research, Applied Mathematics Division, under Grant 68-1408.

of this problem approximate as closely as desired the infimum in the original control problem for sufficiently small ϵ. The epsilon problem can be solved (computationally) in many ways, but the gradient methods have a theoretical significance as well. In fact, we can obtain an *epsilon maximum principle* for the problem which in the limit, as ϵ goes to zero, yields the familiar Pontryagin maximum principle [2]. Results bearing on computational aspects will be reported elsewhere.[1]

We begin in §2 with the simplest type of control problem, namely, with linear dynamics and quadratic criteria where the solution to the epsilon problem can be made explicit, and the behavior, as ϵ goes to zero, of the solution of the epsilon problem can be made explicit as well. We consider both the general fixed endpoint "pursuit" problem as well as the final value problem, and extend the results to the infinite-dimensional state-space case. Equations (2.11), (2.12) characterizing the solutions of the epsilon problem (and the infinite-dimensional versions (2.22), (2.23)) would appear to be new and of independent interest. Similarly, the results obtained for the epsilon problem in the case of boundary control for partial differential equations appear to be new, and the limiting case as ϵ goes to zero of course provides a new method of establishing the character of the optimal control law.

The main results are presented in §3. Here we consider the general nonlinear problem with fixed endpoint. We characterize the solutions of the epsilon problem, obtain an epsilon maximum principle, and show that the corresponding optimal controls provide a minimizing sequence for the original optimal control problem, the approach being monotonic in the cost function. If the optimal control for the epsilon problem converges pointwise to an admissible control, the latter is actually an optimal control, and the epsilon maximum principle yields the familiar maximum principle in the limit. We also show that an optimal solution for the epsilon problem will exist under some mild conditions. It is probable that the conditions imposed in existence and limiting theorems can be weakened.

In §4 we indicate the modifications necessary to handle the time-optimal problem. For a restricted class of dynamics we show how the epsilon maximum problem yields the maximum principle in the limit.

2. Linear systems—quadratic criteria. We begin with the simplest and most common kind of control system: linear finite-dimensional plant with quadratic cost functions. Thus we may take the plant dynamics as

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(0) = x_0, \tag{2.1}$$

[1] In the forthcoming doctoral thesis of S. DeJulio, Department of Engineering, University of California at Los Angeles.

where all functions are rectangular matrices, $x(\cdot)$ being the state and $u(\cdot)$ the control, and $A(t)$ and $B(t)$ are bounded by square summable functions in each finite interval so that (2.1) has a unique solution. The optimization problem we shall consider is to minimize

$$\int_0^T g(x(t),u(t))\, dt,$$

where T is fixed and

$$g(x(t), u(t)) = \| x(t) - x_d(t)\|^2 + \lambda \| u(t)\|^2, \tag{2.2}$$

$x_d(t)$ being a given (square summable) desired function and λ a fixed positive parameter. Here there is no constraint on the control function other than that imposed through the usual "penalty function" in (2.2).

We begin by considering the intermediate nondynamic problem for fixed $\epsilon > 0$, which is to minimize

$$h(\epsilon; x(\cdot), u(\cdot)) = \frac{1}{2\epsilon}\int_0^T \| \dot{x}(t) - A(t)x(t) - B(t)u(t)\|^2\, dt + \int_0^T g(x(t), u(t))\, dt \tag{2.3}$$

over the class of absolutely continuous functions $x(t)$, with $\dot{x}(t)$, $u(t)$ in $L_2(0, T)$ (a generic notation for the appropriate dimensional functions square integrable over $(0, T)$). We shall first show that (2.3) has a unique minimizing solution. Let $u_n(\cdot)$, $x_n(\cdot)$ be a sequence of admissible functions such that

$$\lim_{n\to\infty} h(\epsilon; x_n(\cdot), u_n(\cdot)) = \inf h(\epsilon; x(\cdot), u(\cdot)) = h(\epsilon).$$

Let

$$z_n(t) = \dot{x}_n(t) - A(t)x_n(t) - B(t)u_n(t).$$

Then we must have

$$x_n(t) = \Phi(t)x_0 + \int_0^t \Phi(t)\Phi(s)^{-1}(B(s)u_n(s) + z_n(s))\, ds,$$

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

$$\dot{\Phi}(t) = A(t)\Phi(t); \qquad \Phi(0) = I,$$

where I is the identity matrix. Now

$$\sup \int_0^T \| z_n(t)\|^2\, dt < \infty, \qquad \sup \int_0^T \| u_n(t)\|^2\, dt < \infty.$$

Hence we can find subsequences that converge weakly to $z_0(\cdot)$ and $u_0(\cdot)$. Also the mapping

$$\int_0^t \Phi(t)\Phi(s)^{-1}B(s)u(s)\,ds, \qquad 0 < t < T,$$

on $u(\cdot)$, as well as the mapping

$$\int_0^t \Phi(t)\Phi(s)^{-1}z(s)\,ds, \qquad 0 < t < T,$$

into $L_2(0\,,\,T)$ are both compact. Hence it follows that the corresponding subsequence $x_n(\cdot)$ converges strongly in $L_2(0\,,\,T)$ to a function $x_0(\cdot)$. In fact,

$$x_0(t) = \Phi(t)x_0 + \int_0^t \Phi(t)\Phi(s)^{-1}(B(s)u_0(s) + z_0(s))\,ds \tag{2.4}$$

establishes in particular that $x_0(\cdot)$ is absolutely continuous with derivative in $L_2(0,\,T)$, and that

$$\dot{x}_0(t) = A(t)x_0(t) + B(t)u_0(t) + z_0(t) \quad \text{a.e.} \tag{2.5}$$

But weak convergence of $z_n(\cdot)$, $u_n(\cdot)$ implies that

$$\begin{aligned} h(\epsilon;\, x_0(\cdot),\, u_0(\cdot)) &= \frac{1}{2\epsilon}\int_0^T \| z_0(t) \|^2\,dt + \int_0^T \| x_0(t) - x_d(t) \|^2\,dt \\ &\qquad + \lambda \int_0^T \| u(t) \|^2\,dt \\ &\leqq \liminf h(\epsilon;\, x_n(\cdot);\, u_n(\cdot)) \\ &= h(\epsilon), \end{aligned}$$

or, in other words, $x_0(\cdot)$, $u_0(\cdot)$ is a minimizing solution. To see that the solution is unique we have only to note that $h(\epsilon;\, x(\cdot);\, u(\cdot))$ is (strictly) convex; indeed let $x_1(\cdot)$, $u_1(\cdot)$ be another minimizing solution. Then if

$$\begin{aligned} z(t) &= (1 - \theta)z_0(t) + \theta z_1(t), \\ x(t) &= (1 - \theta)x_0(t) + \theta x_1(t), \\ u(t) &= (1 - \theta)u_0(t) + \theta u_1(t), \end{aligned}$$

we have

$$\begin{aligned} \frac{d^2}{d\theta^2} h(\epsilon:\, x(\cdot);\, u(\cdot)) &= \frac{1}{\epsilon}\int_0^T \| z_0(t) - z_1(t) \|^2\,dt \\ &\qquad + \int_0^T \| x_0(t) - x_1(t) \|^2\,dt + \lambda \int_0^T \| u_0(t) - u_1 \|^2\,dt \\ &= 0 \end{aligned}$$

if and only if $u_0(\cdot) = u_1(\cdot)$, $x_0(\cdot) = x_1(\cdot)$.

Having established the existence and uniqueness of the minimizing solution for the problem (2.3), we now proceed to characterize it further by examining the first variation which must of course vanish at the optimal solution. Thus let us denote the optimal solutions by $u_0(t;\epsilon)$, $x_0(t;\epsilon)$ to indicate the dependence on ϵ. Let

$$z_0(t;\epsilon) = \dot{x}_0(t;\epsilon) - A(t)x_0(t;\epsilon) - B(t)u_0(t;\epsilon). \tag{2.6}$$

Let $h(t)$ be an element in the Schwartz space of infinitely smooth functions vanishing outside compact subsets of the open interval $(0, T)$. Let $v(t)$ be any function in the same $L_2(0, T)$ space as $u_0(t;\epsilon)$. Then setting the first variation to zero we obtain

$$\int_0^T [z_0(t;\epsilon), \dot{h}(t) - A(t)h(t)]\,dt + \epsilon\int_0^T [x_0(t;\epsilon) - x_d(t), h(t)]\,dt = 0, \tag{2.7}$$

$$\int_0^T [z_0(t;\epsilon), B(t)v(t)]\,dt = \lambda\epsilon \int_0^T [u_0(t;\epsilon), v(t)]\,dt, \tag{2.8}$$

where $[\,,]$ indicates the appropriate inner product. Now (2.7) implies that

$$\int_0^T [z_0(t;\epsilon), \dot{h}(t)]\,dt = \int_0^T [A^*(t)z_0(t;\epsilon) - \epsilon(x_0(t;\epsilon) - x_d(t)), h(t)]\,dt$$

for every function in the Schwartz space, or the distributional derivative of $z_0(t;\epsilon)$ coincides with the function

$$-A^*(t)z_0(t;\epsilon) + \epsilon(x_0(t;\epsilon) - x_d(t)) \tag{2.9}$$

which is an ordinary (square integrable) function. Hence $z_0(t;\epsilon)$ is absolutely continuous with the derivative given a.e. by (2.9). Next let $h(t)$ be any absolutely continuous function wth $h(0) = 0$ and $\dot{h}(t)$ square integrable on $[0, T]$. Then

$$\int_0^T [z_0(t;\epsilon), \dot{h}(t)]\,dt = \int_0^T [A^*(t)z_0(t;\epsilon) - \epsilon(x_0(t;\epsilon) - x_d(t)), h(t)] + [z_0(T;\epsilon), h(T)]$$

and since $h(T)$ is arbitrary, it follows that $z_0(T;\epsilon)$ must be zero. Hence it follows that $z_0(t;\epsilon)$ is the unique solution of

$$\dot{z}_0(t;\epsilon) + A^*(t)z_0(t;\epsilon) - \epsilon(x_0(t;\epsilon) - x_d(t)) = 0, \quad z_0(T;\epsilon) = 0, \tag{2.10}$$

and of course from (2.8),

$$B^*(t)z_0(t;\epsilon) = \lambda\epsilon u_0(t;\epsilon).$$

Let

$$\psi(t) = \Phi(t)^{*-1}$$

Then (2.10) has the solution

$$z_0(t;\epsilon) = \epsilon\int_T^t \psi(t)\psi(s)^{-1}(x_0(s;\epsilon) - x_d(s))\,ds. \tag{2.10a}$$

Hence finally we note that $x_0(t;\epsilon)$, $u_0(t;\epsilon)$ are characterized as the unique solution of

$$\begin{aligned}&\dot{x}_0(t;\epsilon) - A(t)x_0(t;\epsilon) - B(t)u_0(t;\epsilon)\\ &\qquad = \epsilon\int_T^t \psi(t)\psi(s)^{-1}(x_0(s;\epsilon) - x_d(s))\,ds,\quad x_0(0;\epsilon) = x_0,\end{aligned} \tag{2.11}$$

$$\lambda u_0(t;\epsilon) = B^*(t)\int_T^t \psi(t)\psi(s)^{-1}\,(x_0(s;\epsilon) - x_d(s))\,ds. \tag{2.12}$$

The function $x_0(t;\epsilon)$ is given by itself as the solution of

$$\begin{aligned}&\dot{x}_0(t;\epsilon) - A(t)x_0(t;\epsilon)\\ &\qquad = \left[\frac{B(t)B^*(t)}{\lambda} + \epsilon I\right]\int_T^t \psi(t)\psi(s)^{-1}(x_0(s;\epsilon) - x_d(s))\,ds,\\ &x_0(0;\epsilon) = x_0.\end{aligned} \tag{2.13}$$

One can also rewrite (2.12) as

$$\begin{aligned}\lambda u_0(t;\epsilon) &= B^*(t)\psi(t)\frac{z_0(0;\epsilon)}{\epsilon}\\ &\quad + B^*(t)\int_0^t \psi(t)\psi(s)^{-1}(x_0(s;\epsilon) - x_d(s))\,ds.\end{aligned} \tag{2.14}$$

Finally, if we set

$$\begin{aligned}&H(\epsilon; z_0(t;\epsilon); x_0(t;\epsilon); t; u)\\ &\qquad = [z_0(t;\epsilon), A(t)x_0(t;\epsilon) + B(t)u] - \epsilon g(x_0(t;\epsilon); u),\end{aligned} \tag{2.14a}$$

we have the epsilon maximum principle that the left side of (2.14a) is a maximum for each t for $u = u_0(t;\epsilon)$.

We shall next study the behavior of the optimal solution $x_0(t;\epsilon)$, $u_0(t;\epsilon)$ as a function of ϵ in a neighborhood of the origin. Now $x_0(t;\epsilon)$ is the unique

solution of the linear equation (2. 13). Putting $\epsilon = 0$ we obtain

$$\dot{x}_0(t; 0) - A(t)x_0(t; 0) = \frac{B(t)B^*(t)}{\lambda} \cdot \int_T^t \psi(t)\psi(s)^{-1}(x_0(s; 0) - x_d(s))\, ds. \tag{2.15}$$

Also

$$\lambda u_0(t; 0) = B^*(t) \int_T^t \psi(t)\psi(s)^{-1}(x_0(s; 0) - x_d(s))\, ds. \tag{2.16}$$

But $x_0(t; 0)$, $u_0(t; 0)$ are readily recognized as the well-known solution to the optimal control problem. In particular, since $z_0(t; 0)$ is identically zero, $x_0(t; 0)$ of course satisfies the plant equation with control given by (2.16). Moreover, it follows from standard perturbation theory for linear equations that $x_0(t; \epsilon)$ is actually analytic in ϵ in a neighborhood of the origin. Indeed, we can proceed to obtain a power series expansion for $x_0(t; \epsilon)$ by the usual technique of taking partial derivatives of (2.13) with respect to the parameter. We shall omit the standard calculations here and note that we can write

$$x_0(t; \epsilon) = x_0(t; 0) + \sum_1^\infty \epsilon^k y_k(t),$$

where

$$y_k = L^k(x_0(\cdot\,; 0) - x_d(\cdot)),$$

where L is the linear transformation defined by taking the solution of the linear equation

$$\dot{y}(t) - A(t)y(t) - \frac{B(t)B(t)^*}{\lambda} \int_T^t \psi(t)\psi(s)^{-1}y(s)\, ds = \frac{B(t)B(t)^*}{\lambda} \int_t^T \psi(t)\psi(s)^{-1}x(s)\, ds, \tag{2.17}$$

$$y(0) = 0$$

and

$$y = Lx.$$

Also

$$u_0(t; \epsilon) = u_0(t; 0) - \frac{B(t)^*}{\lambda} \int_T^t \psi(t)\psi(s)^{-1} \sum_1^\infty \epsilon^k y_k(s)\, ds. \tag{2.18}$$

The expansions are clearly valid in a sufficiently small neighborhood of the

origin. Finally we note that from (2.10a),

$$z_0(t;\epsilon) = \epsilon \int_T^t \psi(t)\psi(s)^{-1} \sum_1^\infty \epsilon^k y_k(s)\, ds + \epsilon \int_T^t \psi(t)\psi(s)^{-1}(x_0(s;0) - x_d(s))\, ds.$$

It follows in particular that

$$\lim_{\epsilon\to 0} \frac{z_0(t;\epsilon)}{\epsilon} = \lim_{\epsilon\to 0} \frac{\partial}{\partial\epsilon} z_0(t;\epsilon) = \int_T^t \psi(t)\psi(s)^{-1}(x_0(s;0) - x_d(s))\, ds,$$

which is clearly absolutely continuous with derivative square integrable over $[0, T]$. Also, for any $h(t)$ that is in the same $L_2(0, T)$ space as $x_0(t; 0)$,

$$\epsilon \int_0^T \left[\frac{\partial}{\partial\epsilon}(x_0(t;\epsilon) - x_d(t)), h(t)\right] dt = \int_0^T \left[\epsilon \frac{\partial}{\partial\epsilon} \sum_1^\infty \epsilon^k y_k(t), h(t)\right] dt \to 0 \quad \text{as} \quad \epsilon \to 0.$$

Again for any $v(t)$ as before, by (2.18),

$$\int_0^T \left[\epsilon \frac{\partial}{\partial\epsilon} u_0(t;\epsilon), v(t)\right] dt = -\int_0^T \left[\epsilon B(t)^* \int_T^t \psi(t)\psi(s)^{-1} \frac{\partial}{\partial\epsilon} \sum_1^\infty \epsilon^k y_k(s)\, ds, v(t)\right] dt \to 0 \quad \text{as} \quad \epsilon \to 0.$$

Also, of course letting

$$\phi(t) = \lim_{\epsilon\to 0} \frac{\partial}{\partial\epsilon} z_0(t;\epsilon) = \int_T^t \psi(t)\psi(s)^{-1}(x_0(s;0) - x_d(s))\, ds,$$

we have

$$\dot\phi(t) + A(t)^*\phi(t) = x_0(t;0) - x_d(t), \qquad \phi(T) = 0,$$
$$u_0(t;0) = B(t)^*\phi(t),$$

and finally,

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon} H(\epsilon;\, z_0(t;\epsilon);\, x_0(t;\epsilon);\, t;\, u) = \lim_{\epsilon\to 0} \frac{\partial}{\partial\epsilon} H(\epsilon; z_0(t;\epsilon);\, x_0(t;\epsilon);\, t;\, u) = [\phi(t), A(t)x_0(t) + B(t)u] - g(x_0(t); u) \tag{2.19}$$

and the right side is a maximum for $u = u_0(t)$, which is the maximum principle for the problem.

2.1. Final value problems. We shall now indicate how similar results can be obtained for the final value problem. Thus, with the dynamical equations

as in (2.1), let it be required to minimize

$$\| x(T) - x_d \|^2 + \lambda \int_0^T \| u(t) \|^2 \, dt, \tag{2.20}$$

where x_d is the desired "final" value at time T. We take the corresponding epsilon problem as

$$\frac{1}{2\epsilon} \int_0^T \| \dot{x}(t) - A(t)x(t) - B(t)u(t) \|^2 \, dt + \| x(T) - x_d \|^2 + \lambda \int_0^T \| u(t) \|^2 \, dt.$$

The existence of a unique minimizing solution is trivial, as before. Using the same notation $u_0(t;\epsilon)$, $x_0(t;\epsilon)$, $z_0(t;\epsilon)$ as before to denote the optimal solution, we readily obtain the following necessary conditions by the same perturbation methods (first order variation):

$$\begin{aligned} \dot{z}_0(t;\epsilon) + A(t)^* z_0(t;\epsilon) &= 0, \\ \lambda\epsilon u_0(t;\epsilon) &= B(t)^* z_0(t;\epsilon), \\ z_0(T;\epsilon) &= \epsilon(x_d - x_0(T;\epsilon)), \end{aligned}$$

the last equation being the only one that is different. We have thus:

$$\begin{aligned} x_0(t;\epsilon) &= \Phi(t)x_0 \\ &+ \int_0^t \Phi(t)\Phi(s)^{-1}\left(\frac{B(s)B(s)^*}{\lambda} + \epsilon I\right)\Phi(s)^{*^{-1}}\Phi(T)^*(x_d - x_0(T;\epsilon)) \, ds. \end{aligned} \tag{2.20a}$$

In particular, we can obtain $x_0(T;\epsilon)$ from

$$x_0(T;\epsilon) = \left(I + \frac{R}{\lambda} + \epsilon R\right)^{-1}\left(\Phi(T)x_0 + \frac{R}{\lambda} x_d + \epsilon R x_d\right),$$

where R is the operator defined by

$$Rx = \int_0^T \Phi(T)\Phi(s)^{-1}B(s)B(s)^*\Phi(s)^{*^{-1}} \Phi(T)^* x \, ds.$$

The limiting case as ϵ goes to zero is thus readily obtained as

$$\begin{aligned} \lim_{\epsilon\to 0} \frac{z_0(t;\epsilon)}{\epsilon} &= \Phi(t)^{*^{-1}}\Phi(T)^*(x_d - x_0(T)), \\ x_0(T) &= \left(I + \frac{R}{\lambda}\right)^{-1}\left(\Phi(T)x_0 + \frac{R}{\lambda} x_d\right), \\ \lim_{\epsilon\to 0} u_0(t;\epsilon) &= \frac{B(t)^*}{\lambda}\Phi(t)^{*^{-1}}\Phi(T)^*(x_d - x_0(T)). \end{aligned}$$

The epsilon maximum principle and the maximum principle for this problem can also be obviously obtained by appropriate modification of (2.14a) and (2.19).

2.2. Infinite-dimensional problems. The infinite-dimensional versions (that is, where the state space and control space are infinite-dimensional) of the above problems are readily deduced. Thus let us (for simplicity) consider the time-invariant problem where the dynamics are now given by

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = x_0 \in \text{domain of } A, \tag{2.21}$$

where $x(t)$ has its range in a Hilbert space $\mathcal{H}_1$, and A is the infinitesimal generator of a strongly continuous semigroup $T(s)$ over $\mathcal{H}_1$; the control $u(t)$ has its range in a possibly different Hilbert space $\mathcal{H}_2$, and B is a linear bounded transformation mapping $\mathcal{H}_2$ into the domain of A. Then for the problem (2.2), we again take the epsilon problem as (2.3). The existence and uniquenes of the solution for the epsilon problem is immediate; using the same notation $u_0(t;\epsilon)$, $x_0(t;\epsilon)$ for the optimal control and state functions and $z_0(t;\epsilon)$ as before, the necessary conditions may be deduced as before, and (2.13) becomes

$$\dot{x}_0(t;\epsilon) - Ax_0(t;\epsilon) = \left(\frac{BB^*}{\lambda} + \epsilon I\right)\int_T^t T(s-t)^*(x_0(s;\epsilon) - x_d(s))\,ds,$$

$$x_0(0;\epsilon) = x_0, \tag{2.22}$$

which is readily shown to have a unique solution. Also, corresponding to (2.12), we have

$$\lambda u_0(t;\lambda) = B^* \int_T^t T(s-t)^*(x_0(s;\epsilon) - x_d(s))\,ds. \tag{2.23}$$

The limiting optimal control is also deduced as before leading to well-known solutions (see [3], [4]).

There is, however, another and more important class of problems, namely, the boundary control problems, where the method actually leads to new results. Not to complicate matters unduly, we shall again consider a time-invariant problem for a concrete partial differential equation. Thus let the partial differential equation be

$$\frac{\partial}{\partial t} f(t,x) = \nabla^2 f(t,x), \qquad 0 < t, \quad x \in \Omega \subset E_n, \tag{2.24}$$

where Ω is a bounded open domain in E_n with smooth boundary S, and ∇^2 stands for the Laplacian. The control is now on the boundary

$$u(t,\sigma) = f(t,\sigma), \qquad \sigma \in S. \tag{2.25}$$

We shall consider a fixed endpoint problem, where it is desired to minimize

$$\int_0^T \int_\Omega |f(t,x) - g(t,x)|^2 \, d|x| \, dt + \lambda \int_0^T \int_S |u(t,\sigma)|^2 \, dS \, dt \tag{2.26}$$

and (for simplicity) we take the initial "value" $f(0, x)$ to be zero almost everywhere (in x). (For a standard treatment of such a problem, see [5].) We begin by formulating the epsilon problem as that of minimizing

$$\begin{aligned} &\frac{1}{2\epsilon} \int_0^T \int_\Omega \left| \frac{\partial f}{\partial t} - \nabla^2 f \right|^2 d|x| \, dt \\ &\quad + \int_0^T \int_\Omega |f(t,x) - g(t,x)|^2 \, d|x| \, dt \\ &\quad + \lambda \int_0^T \int_S |u(t,\sigma)|^2 \, dS \, dt, \end{aligned} \tag{2.27}$$

where $u(t, \sigma)$ is given by (2.25), over the class of functions $f(t, x)$ defined on $\Lambda = [0, T] \times \bar{\Omega}$ such that $f(t, x)$ is continuous thereon, absolutely continuous in t for each x, $(\partial f/\partial t)(t, \cdot)$ is an element of $L_2(\Lambda)$ of class C^2 in x on Ω such that $\nabla^2 f(\cdot, x)$ is also in $L_2(\Lambda)$. Because this is again a linear problem with a quadratic cost function, the existence and uniqueness of an optimal solution may be demonstrated essentially as in the finite-dimensional case. Let us now examine the necessary conditions that the optimal solution must satisfy. Let $f_0(t, x, \epsilon)$ denote the optimal solution with

$$u_0(t, \sigma, \epsilon) = f_0(t, \sigma, \epsilon), \qquad \sigma \in S.$$

Let $h(t, x)$ be any infinitely differentiable function vanishing outside compact subsets of $(0, T) \times \Omega$. Setting

$$f(t, x) = f_0(t, x, \epsilon) + \theta h(t, x)$$

and taking the first derivative of the functional (2.27) with respect to θ at $\theta = 0$, we have

$$\begin{aligned} &\int_0^T \int_\Omega \left(\frac{\partial h}{\partial t} - \nabla^2 h \right) z_0(t, x, \epsilon) \, d|x| \, dt \\ &\quad + \epsilon \int_0^T \int_\Omega (f_0(t, x, \epsilon) - g(t, x)) h(t, x) \, d|x| \, dt \\ &\quad + \lambda\epsilon \int_0^T \int_\Omega u_0(t, \sigma, \epsilon) h(t, \sigma) \, dS \, dt = 0, \end{aligned}$$

where

$$z_0(t, x, \epsilon) = \frac{\partial f}{\partial t} f_0(t, x, \epsilon) = \nabla^2 f_0(t, x, \epsilon), \qquad x \in \Omega. \tag{2.28}$$

As in the finite-dimensional case we readily obtain that

$$\frac{\partial}{\partial t} z_0(t, x, \epsilon) + \nabla^2 z_0(t, x, \epsilon) = \epsilon(g(t, x) - f_0(t, x, \epsilon)), \qquad x \in \Omega. \tag{2.29}$$

By using Green's integral formula,

$$\int_\Omega U\nabla^2 V = \int_\Omega V\nabla^2 U = \int_S V \frac{\partial U}{\partial \nu}\, dS - \int_S U \frac{\partial U}{\partial \nu}\, dS,$$

where ν is the inner normal to S, and with the appropriate choice of a perturbing function $h(t, x)$, it follows that we must have

$$z_0(t, \sigma, \epsilon) = 0, \qquad \sigma \in S, \tag{2.30}$$

$$\lambda \in u_0(t, \sigma, \epsilon) = \frac{\partial}{\partial \nu} z_0(t, \sigma, \epsilon). \tag{2.31}$$

We have thus obtained the necessary conditions for optimality. It may be shown (essentially as in the finite-dimensional case) that (2.29) has a unique solution subject to (2.30), and so does (2.28) subject to (2.31).

Passing on to the limiting case as ϵ goes to zero, we obtain

$$\phi(t, x) = \lim_{\epsilon \to 0} \frac{z_0(t, x, \epsilon)}{\epsilon},$$

$$u_0(t, \sigma) = \lim_{\epsilon \to 0} u_0(t, \sigma, \epsilon),$$

$$f_0(t, x) = \lim_{\epsilon \to 0} f_0(t, x, \epsilon),$$

and we observe that the optimal control $u_0(t, \sigma)$ satisfies (see [5])

$$\lambda u_0(t, \sigma) = \frac{\partial}{\partial \nu} \phi(t, \sigma),$$

and the function $\phi(t, x)$ satisfies

$$\phi(t, x) + \nabla^2 \phi(t, x) = g(t, x) - f_0(t, x), \qquad x \in \Omega,$$

$$\phi(t, \sigma) = 0, \qquad \sigma \in S.$$

3. Fixed endpoint problems with finite-dimensional state space. We shall now consider the general fixed endpoint problem but with the state spaces still finite-dimensional. We shall discuss both necessary and sufficient conditions for optimality.

In what follows we assume that the dynamics are described by the equations

$$\dot{x}(t) = f(t; x(t); u(t)), \qquad x(0) = x_0, \tag{3.1}$$

where (as usual, see [2], [6]) it is assumed that $f(t; x; u)$ is continuous in all of the variables and is continuously differentiable with respect to x. As in §2, we shall take the state $x(t)$ as $n \times 1$, the control $u(t)$ as $p \times 1$. The control problem is to minimize the functional

$$\int_0^T g(t; x(t); u(t))\, dt \tag{3.2}$$

with the control function $u(t)$ subject to certain constraint conditions C which will be described below. Here $g(t, x, u)$ is assumed continuous in all of the variables and continuously differentiable in x.

The nondynamic epsilon problem is then to minimize, for fixed $\epsilon > 0$,

$$\begin{aligned} h(\epsilon; x(\cdot); u(\cdot)) &= \frac{1}{2\epsilon} \int_0^T \| \dot{x}(t) - f(t; x(t); u(t)) \|^2 \, dt \\ &\quad + \int_0^T g(t; x(t); u(t))\, dt \end{aligned} \tag{3.3}$$

over the class of functions $x(t)$, $u(t)$ such that $x(t)$ is absolutely continuous with $x(0) = x_0$, the derivative $\dot{x}(t)$ square integrable over $[0, T]$, and $u(t)$ measurable and subject to constraint C. A measurable control function $u(t)$ satisfying the constraint C will be referred to as an *admissible* control. It is assumed that (3.1) has a unique solution for each admissible control. It will be further assumed that C is such that the admissible controls are uniformly bounded in $[0, T]$. In addition we shall assume throughout that $f(t; x(t); u(t))$ is such that for admissible $u(t)$,

$$[x(t), f(t; x(t); u(t))] \leqq m[1 + \| x(t) \|^2]. \tag{3.1a}$$

This implies that $x(t)$ in (3.1) is uniformly bounded for admissible controls (see [6]).

The first result relating the epsilon problem to the control problem of minimizing (3.2) is the following theorem.

THEOREM 3.1. *Let $h(\epsilon)$ denote the infimum (over the stipulated class) of (3.3). Suppose the infimum is attained so that*

$$h(\epsilon) = h(\epsilon; x_0(\cdot\,; \epsilon); u_0(\cdot\,; \epsilon))$$

for each $\epsilon > 0$. Let $\hat{x}(\cdot, \epsilon)$ be the solution of (3.1) with $u(t)$ given by $u_0(t, \epsilon)$. Then

$$\begin{aligned} \lim_{\epsilon \to 0} h(\epsilon) = \inf_{\epsilon \to 0} h(\epsilon) &= \inf \int_0^T g(t; x(t); u(t))\, dt = g_0 \\ &= \lim_{\epsilon \to 0} \int_0^T g(t; \hat{x}(t; \epsilon); u_0(t; \epsilon))\, dt. \end{aligned} \tag{3.4}$$

In other words, the epsilon problem approximates the actual control

problem and provides an approximating sequence of controls that approximate the optimum, if any.

Proof. Let $0 < \epsilon_2 < \epsilon_1$ and let

$$z_0(t;\epsilon) = \dot{x}_0(t;\epsilon) - f(t; x_0(t;\epsilon); u_0(t;\epsilon)).$$

Let us observe that

$$\frac{1}{2\epsilon_2}\int_0^T \| z_0(t;\epsilon_2)\|^2\, dt + \int_0^T g(t; x_0(t;\epsilon_2); u(t;\epsilon_2))\, dt$$

$$\leqq \frac{1}{2\epsilon_2}\int_0^T \| z_0(t;\epsilon_1)\|^2\, dt + \int_0^T g(t; x_0(t;\epsilon_1); u_0(t;\epsilon_1))\, dt.$$

Similarly,

$$\frac{1}{2\epsilon_1}\int_0^T \| z_0(t;\epsilon_1)\|^2\, dt + \int_0^T g(t; x_0(t;\epsilon_1); u_0(t;\epsilon_1)\, dt$$

$$\leqq \frac{1}{2\epsilon_1}\int_0^T \| z_0(t;\epsilon_2)\|^2\, dt + \int_0^T g(t; x_0(t;\epsilon_2)u_0(t;\epsilon_2))\, dt.$$

Since $\epsilon_2 < \epsilon_1$ it readily follows that

$$\int_0^T \| z_0(t;\epsilon_2)\|^2\, dt \leqq \int_0^T \| z_0(t;\epsilon_1)\|^2\, dt,$$

$$\int_0^T g(t; x_0(t;\epsilon_2); u_0(t;\epsilon_2))\, dt \geqq \int_0^T g(t; x_0(t;\epsilon_1); u_0\,(t;\epsilon))\, dt.$$

Hence we have that

$$\int_0^T g(t; x_0(t;\epsilon); u_0(t;\epsilon))\, dt \tag{3.5}$$

is monotone increasing as ϵ goes to zero, while

$$\int_0^T \| z_0(t;\epsilon)\|^2\, dt \tag{3.6}$$

decreases monotonically as ϵ goes to zero. On the other hand, we know that

$$h(\epsilon) \leqq \inf \int_0^T g(t; x(t); u(t))\, dt = g_0\,, \tag{3.7}$$

where the infimum is over admissible controls with state satisfying (3.1). In particular, then, (3.6) converges monotonically to zero as ϵ goes to zero, and

$$\lim_{\epsilon\to 0} \int_0^T g(t; x_0(t;\epsilon); u_0(t;\epsilon)) \leqq g_0\,. \tag{3.8}$$

Let $\hat{x}(t;\epsilon)$ denote the solution of (3.1) with $u(t) = u_0(t;\epsilon)$. Since $g(t;x;u)$ is continuously differentiable in x, and the admissible controls are uniformly bounded, it follows that we can find a constant M such that

$$\begin{aligned}(3.9)\qquad &|g(t;\hat{x}(t;\epsilon);u_0(t;\epsilon)) - g(t;x_0(t;\epsilon);u_0(t;\epsilon))| \\ &\qquad \leqq M\,\|x_0(t;\epsilon) - x(t;\epsilon)\|, \quad 0 \leqq t \leqq T.\end{aligned}$$

Next let

$$y(t) = \hat{x}(t;\epsilon) - x_0(t;\epsilon).$$

Then because of the existence of the continuous derivative in x of $f(t;x;u)$ it follows that

$$\begin{aligned}[\dot{y}(t), y(t)] &= [f(t;\hat{x}(t;\epsilon);u_0(t;\epsilon)) - f(t;x_0(t;\epsilon);u_0(t;\epsilon)), y(t)] \\ &\quad - [z_0(t;\epsilon), y(t)] \\ &= [M(t)y(t), y(t)] - [z_0(t;\epsilon), y(t)],\end{aligned}$$

where $M(t)$ is uniformly bounded independently of ϵ, and so is $y(t)$, by (3.1a) and the fact that $\int_0^T \|z_0(t;\epsilon)\|^2\,dt$ is going to zero. Hence it follows that if

$$m(t) = \|y(t)\|^2,$$

then

$$(3.10)\qquad \dot{m}(t) \leqq M_1 m(t) + M_2\,\|z_0(t;\epsilon)\|, \qquad m(0) = 0.$$

Hence it follows that $m(t)$ goes to zero in such a way that from (3.9) we have

$$\lim_{\epsilon\to 0}\int_0^T |g(t;x(t;\epsilon);u_0(t;\epsilon)) - g(t;x_0(t;\epsilon);u_0(t;\epsilon))|\,dt = 0.$$

Hence,

$$\begin{aligned}(3.11)\qquad g_0 &\leqq \lim_{\epsilon\to 0}\int_0^T g(t;\hat{x}(t;\epsilon);u_0(t;\epsilon))\,dt \\ &= \lim_{\epsilon\to 0}\int_0^T g(t;x_0(t;\epsilon);u_0(t;\epsilon))\,dt.\end{aligned}$$

From (3.8) therefore it follows that equality holds in (3.11); thus we have proved (3.4). In particular,

$$(3.12)\qquad \frac{1}{2\epsilon}\int_0^T \|z_0(t;\epsilon)\|^2\,dt \to 0.$$

3.1. Necessary conditions for optimality. We shall next turn to necessary conditions for optimality. We shall first obtain necessary conditions for optimality for the nondynamic epsilon problem (3.3) and then show how under suitable limiting conditions they lead to the Pontryagin maximum principle for optimal control.

With $x_0(t;\epsilon)$, $u_0(t;\epsilon)$, $z_0(t;\epsilon)$ an optimal solution for the problem (3.3), let $h(t)$ be any $n \times 1$ function in the Schwartz space of infinitely smooth functions vanishing outside compact subsets of $(0, T)$. Let

$$x(t) = x_0(t;\epsilon) + \theta h(t),$$

where θ is a real variable. Then we must have

$$\frac{d}{d\theta} h(\epsilon; x(\cdot); u_0(\cdot;\epsilon))\,|_{\theta=0} = 0.$$

But because of our assumptions on $f(\cdot)$ and $g(\cdot)$ it follows that this derivative equals

$$\frac{1}{\epsilon}\int_0^T [z_0(t;\epsilon), \dot{h}(t) - f_1(t; x_0(t;\epsilon); u_0(t;\epsilon))h(t)]\,dt + \int_0^T [g_1(t; x_0(t;\epsilon); u_0(t;\epsilon))]\,dt, \tag{3.13}$$

where

$$f_1(t;x;u) = \nabla_x f(t;x;u), \qquad g_1(t;x;u) = \nabla_x g(t;x;u),$$

$f_1(t;x;u)$ being $n \times n$ and $g_1(t;x;u)$, $n \times 1$, both continuous. As in §2 this is enough to imply that $z_0(t;\epsilon)$ is absolutely continuous with

$$\dot{z}_0(t;\epsilon) = -f_1(t; x_0(t;\epsilon); u_0(t;\epsilon))^* z_0(t;\epsilon) + \epsilon g_1(t; x_0(t;\epsilon); u_0(t;\epsilon)). \tag{3.14}$$

Next, specializing $h(\cdot)$ to any smooth function with

$$h(0) = 0, \quad h(T) \quad \text{arbitrary and nonzero,}$$

it follows as before that

$$z_0(T;\epsilon) = 0. \tag{3.15}$$

We observe that (3.14) is a linear equation for $z_0(t;\epsilon)$ and has a unique solution subject to (3.15). Next let us form the Hamiltonian

$$H(\epsilon; z; x; u; t) = [z, f(t; x; u)] - \epsilon g(t; x; u). \tag{3.16}$$

Let us note that for any admissible control $u(t)$,

$$\frac{1}{2}\int_0^T \| \dot{x}_0(t;\epsilon) - f(t; x_0(t;\epsilon); u(t)) \|^2\,dt + \epsilon\int_0^T g(t; x_0(t;\epsilon); u(t))\,dt \tag{3.17}$$

can be expressed as

$$\frac{1}{2}\int_0^T \| \dot{x}_0(t;\epsilon) \|^2\, dt - \frac{1}{2}\int_0^T \| f(t; x_0(t;\epsilon); u_0(t;\epsilon)) \|^2\, dt$$

$$+ \frac{1}{2}\int_0^T \| f(t; x_0(t;\epsilon); u_0(t;\epsilon)) - f(t; x_0(t;\epsilon); u(t)) \|^2\, dt \tag{3.18}$$

$$- \int_0^T H(\epsilon; z_0(t;\epsilon); x_0(t;\epsilon); u(t); t)\, dt.$$

It follows that

$$\int_0^T H(\epsilon; z_0(t;\epsilon); x_0(t;\epsilon); u(t); t)\, dt \tag{3.19}$$

attains its maximum over admissible control functions when $u(t) = u_0(t;\epsilon)$.

Under certain additional conditions (usually satisfied in control problems in practice) we can also obtain a pointwise version of this maximum property when the constraint C is also expressed as a pointwise condition. In the simplest of these conditions we take C to be such that for each t the control $u(t)$ is required to be in a closed bounded convex set Q and measurable in t, and that both $f(t; x; u)$ and $g(t; x; u)$ are continuously differentiable with respect to u as well. In that case, if $v(t)$ is any admissible control, so is

$$u(t) = (1 - \theta)u_0(t;\epsilon) + \theta v(t), \qquad 0 \leqq \theta \leqq 1,$$

and (3.17), or equivalently (3.19), is differentiable in θ, and the derivative at the origin can be expressed as

$$- \int_0^T [z_0(t;\epsilon), f_2(t; x_0(t;\epsilon); u_0(t;\epsilon))\ (v(t) - u_0(t;\epsilon))]\, dt$$

$$+ \epsilon \int_0^T [g_2(t; x_0(t;\epsilon); u_0(t;\epsilon)), v(t) - u_0(t;\epsilon)]\, dt \tag{3.20}$$

and must be nonnegative. Here

$$f_2(t; x; u) = \nabla_u f(t; x; u),$$

$$g_2(t; x; u) = \nabla_u g(t; x; u),$$

$f_2(\cdot)$ being $n \times p$ and $g_2(\cdot)$ being $p \times 1$. Let E be the subset of $[0, T]$ on which

$$[f_2(t; x_0(t;\epsilon); u_0(t;\epsilon))^* z_0(t;\epsilon) - \epsilon g_2(t; x_0(t;\epsilon); u_0(t;\epsilon)), v(t) - u_0(t;\epsilon)]$$

is positive. Then let us define

$$w(t) = \begin{cases} v(t) & \text{on } E \\ u_0(t;\epsilon) & \text{on the complement of } E. \end{cases}$$

Then $w(t)$ is an admissible control by virtue of our assumptions, and we must have that

$$\int_0^T [f_2(t; x_0(t; \epsilon); u_0(t; \epsilon))^* z_0(t; \epsilon) - \epsilon g_2(t; x_0(t; \epsilon); u_0(t; \epsilon)), w(t) - u_0(t; \epsilon)]\, dt$$

is nonpositive, which is clearly violated unless E is of measure zero. Hence we have that

$$[f_2(t; x_0(t; \epsilon); u_0(t; \epsilon))^* z_0(t; \epsilon) - \epsilon g_2(t; x_0(t; \epsilon); u_0(t; \epsilon)), v(t) - u_0(t; \epsilon)] \leqq 0 \tag{3.21}$$

for every admissible control almost everywhere in $[0, T]$. In particular, then

$$\max_{u \in Q} H(\epsilon; z_0(t; \epsilon); x_0(t; \epsilon); u) = H(\epsilon; z_0(t; \epsilon); x_0(t; \epsilon); u_0(t; \epsilon)) \tag{3.22}$$

almost everywhere in $[0, T]$, as is readily proved by contradiction. We have thus obtained a maximum principle which we may term the epsilon maximum principle. We note that (3.22) can be proved under weaker assumptions.

3.2. Behavior at $\epsilon = 0$; the maximum principle. Let us next examine the limiting form of the necessary conditions as ϵ goes to zero. Let us assume as before the existence of an optimal solution for the epsilon problem and use the same notation $x_0(t; \epsilon)$, $u_0(t; \epsilon)$, $z_0(t; \epsilon)$ as before.

Here we shall treat only the least complicated case (cf. [1]). Suppose then that $u_0(t; \epsilon)$ converges for each t to a function $u_0(t)$, and suppose that $u_0(t)$ is an admissible control. (This has been shown to hold for linear systems with quadratic criteria in §2, but can be shown to hold for more general systems such as the "bilinear" case when

$$f(t; x; u) = A(t)x + B(t)u + (C(t)u)x,$$

where $C(t)$ is a linear transformation mapping E_p into the space of $n \times n$ matrices and $C(t)$ is continuous in t, and $g(\cdot)$ is as in §2, for example.) Since the functions $x_0(t; \epsilon)$ are equicontinuous (by virtue of (3.1a) and the fact that the integral of $z_0(t; \epsilon)$ over any subinterval of $[0, T]$ goes to zero), we can certainly choose a sequence of ϵ's such that $x_0(t; \epsilon)$ converges uniformly in t to, say, $x_0(t)$. Because of our continuity assumptions,

$$x_0(t) = x_0 + \int_0^t \lim f(t; x_0(t; \epsilon); u_0(t; \epsilon))\, dt + \lim \int_0^t z_0(t; \epsilon)\, dt$$

$$= x_0 + \int_0^t f(t; x_0(t); u_0(t))\, dt,$$

and it readily follows that $x_0(t)$, $u_0(t)$ is an optimal solution to the control problem using (3.11). From (3.14), which is a linear equation for $z_0(t;\epsilon)$, it readily follows that $\lim_{\epsilon\to 0}(1/\epsilon)z_0(t;\epsilon)$ exists, and if we denote the limit by $\phi(t)$ we actually have that

$$\dot{\phi}(t) = -f_1(t; x_0(t); u_0(t))^*\phi(t) + g_1(t; x_0(t); u_0(t)), \quad \phi(T) = 0. \tag{3.23}$$

Also it follows (by dividing through by ϵ in (3.13) and taking limits) that

$$\int_0^T [\phi(t), \dot{h}(t) - f_1(t; x_0(t); u_0(t))h(t)]\, dt - \int_0^T [g_0(t; x_0(t); u_0(t)), h(t)]\, dt = 0, \tag{3.24}$$

which is of course equivalent to (3.23). Now from (3.19) we have that for every ϵ and every admissible control $u(t)$,

$$\int_0^T \frac{1}{\epsilon} H(\epsilon; z_0(t;\epsilon); x_0(t;\epsilon); u(t); t)\, dt \leqq \int_0^T \frac{1}{\epsilon} H(\epsilon; z_0(t;\epsilon); x_0(t;\epsilon); u_0(t;\epsilon))\, dt.$$

Hence by taking limits on both sides as ϵ goes to zero, we obtain

$$\int_0^T \tilde{H}(\phi(t); x_0(t); u(t); t)\, dt \leqq \int_0^T \tilde{H}(\phi(t); x_0(t); u_0(t); t)\, dt, \tag{3.25}$$

where

$$\tilde{H}(\phi; x; u; t) = [\phi, f(t; x; u)] - g(t; x; u). \tag{3.26}$$

Again under the additional conditions as in obtaining (3.22), we also obtain (by dividing through by ϵ and taking limits)

$$\max_{u\in Q} \tilde{H}(\phi(t); x_0(t); u; t) = \tilde{H}(\phi(t); x_0(t); u_0(t); t)$$

almost everywhere. In other words we have obtained the maximum principle as announced in [1]. It is clear that the partial derivatives with respect to ϵ used therein are all obtained by simply dividing by ϵ and letting ϵ go to zero.

3.3. Existence of optimal solutions to the epsilon problem. Finally we shall show that an optimal solution to the epsilon problem exists under the following mild condition:

(H) If $u_n(t)$ is a sequence of admissible controls converging weakly to $u_0(t)$, then $u_0(t)$ is also admissible, and further for each absolutely continu-

ous $x(t)$ we have the semicontinuity properties that

$$\int_0^T g(t; x(t); u_0(t))\, dt \leqq \liminf \int_0^T g(t; x(t); u_n(t))\, dt, \tag{3.27}$$

$$\int_0^T \| \dot{x}(t) - f(t; x(t); u_0(t)) \|^2\, dt \leqq \liminf \int_0^T \| \dot{x}(t) - f(t; x(t); u_n(t)) \|^2\, dt. \tag{3.27a}$$

Let us assume (H) and let $x_n(t)$, $u_n(t)$ be an admissible sequence such that

$$\lim \frac{1}{2\epsilon} \int_0^T \| z_n(t) \|^2\, dt + \int_0^T g(t; x_n(t); u_n(t))\, dt = h(\epsilon), \tag{3.28}$$

where

$$z_n(t) = \dot{x}_n(t) - f(t; x_n(t); u_n(t)), \qquad x_n(0) = x_0. \tag{3.29}$$

Since $h(\epsilon)$ is finite, we can clearly choose the sequence so that the first term in (3.28) converges also. Now the functions $x_n(t)$ are uniformly bounded. For if

$$m_n(t) = \| x_n(t) \|^2,$$

then we have

$$\tfrac{1}{2}\dot{m}_n(t) = [\dot{x}_n(t), x_n(t)] = [z_n(t), x_n(t)] + [f(t, x_n(t), u_n(t)), x_n(t)]$$

and

$$|[z_n(t), x_n(t)]| \leqq \| z_n(t) \| (1 + m_n(t)),$$

$$|[f(t; x_n(t), u_n(t)); x_n(t)]| \leqq C(1 + m_n(t)),$$

so that

$$\dot{m}_n(t) \leqq 2 \| z_n(t) \| m_n(t) + M_1 m_n(t) + M_2 \| z_n(t) \| + M_3$$

from which, using (3.28), the usual arguments yield boundedness of $x_n(t)$ independent of n and t. In turn, since $f(\cdot)$ is continuous, $\int_0^T \| \dot{x}_n(t) \|^2\, dt$ is bounded. Since

$$\| x_n(t_2) - x_n(t_1) \|^2 = \left\| \int_{t_1}^{t_2} \dot{x}_n(t)\, dt \right\|^2 \leqq (t_2 - t_1) \int_0^T \| \dot{x}_n(t) \|^2\, dt,$$

the $x_n(t)$ are actually equicontinuous, and hence we can choose a subsequence that converges uniformly to, say, $x_0(t)$. Again, a further subsequence

can be chosen so that $\dot{x}_n(t)$ converges weakly to, say, $y(t)$. But

$$x_0(t) - x_0 \lim_n \int_0^t \dot{x}_n(s)\, ds = \int_0^t \dot{x}_n(s)\, ds = \int_0^t y(s)\, ds$$

implies that $x_0(t)$ equals $y(t)$ almost everywhere. Now for any k, because of the existence of continuous derivatives in x for both $f(\cdot)$ and $g(\cdot)$, we have that

$$\int_0^T g(t; x_0(t); u_k(t))\, dt = \lim_n \int_0^T g(t; x_n(t); u_k(t))\, dt \tag{3.30}$$

uniformly in k. Again,

$$\int_0^T \| \dot{x}_0(t) - f(t; x_0(t); u_k(t)) \|^2\, dt \leqq \lim_n \inf \int_0^T \| \dot{x}_n(t) - f(t; x_n(t); u_k(t)) \|^2\, dt \tag{3.31}$$

independently of k in the sense that the right side equals

$$\lim_n \inf \int_0^T \| \dot{x}_n(t) \|^2\, dt + \int_0^T \| f(t; x_0(t); u_k(t)) \|^2\, dt - 2 \int_0^T [x_0(t), f(t; x_0(t); u_k(t))]\, dt. \tag{3.32}$$

Since the $u_k(t)$ are uniformly bounded we may pick a subsequence that converges weakly to $u_0(t)$, say, which by (H) is admissible. We shall show that by virtue of (H) the functions $x_0(t)$, $u_0(t)$ provide an optimal solution to the epsilon problem. For by (H) we have

$$\int_0^T \| x_0(t) - f(t; x_0(t); u_0(t)) \|^2\, dt \leqq \lim_k \inf \int_0^T \| x_0(t) - f(t; x_0(t); u_k(t)) \|^2\, dt$$

and

$$\int_0^T g(t; x_0(t); u_0(t))\, dt \leqq \lim_k \inf \int_0^T g(t; x_0(t); u_k(t))\, dt,$$

and by virtue of (3.32), (3.31) and (3.30), it follows that

$$\frac{1}{2\epsilon} \int_0^T \| x_0(t) - f(t; x_0(t); u_0(t)) \|^2\, dt + \int_0^T g(t; x_0(t); u_0(t))\, dt \leqq h(\epsilon),$$

or, in other words, $x_0(t)$, $u_0(t)$ is optimal. Condition (H) can be weakened but our aim has not been to state the weakest conditions. Condition (H)

is clearly satisfied if $f(t; x(t); u(t))$ is linear in $u(t)$ (see the following section for such a case).

4. Time-optimal problems. In order to illustrate how the method applies to time-optimal problems, we shall consider such a problem for a somewhat specialized case where the dynamics are specified by

$$\dot{x}(t) = f(t; x(t)) + B(t)u(t) \tag{4.1}$$

with the state function being $n \times 1$, the control $u(t)$ being $p \times 1$ and $f(\cdot)$, $B(\cdot)$ continuous and $f(\cdot)$ continuously differentiable with respect to x, as before. The time-optimal problem we shall consider is that of finding a solution to (4.1) with

$$\begin{aligned} x(0) &= x_0, \\ x(T) &= x_1 \end{aligned} \tag{4.1a}$$

for the smallest possible T, the controls $u(\cdot)$ being now subject to the constraint

$$\| u(t)\| \leqq 1. \tag{4.2}$$

We proceed as follows. First, the epsilon problem is now phrased as that of minimizing

$$h(\epsilon; x(\cdot); u(\cdot)) = \frac{1}{2\epsilon}\int_0^T \| \dot{x}(t) - f(t; x(t)) - B(t)u(t)\|^2 dt + T \tag{4.3}$$

over positive numbers T, the functions $x(t)$ being required to be absolutely continuous with derivative square integrable over finite intervals and $x(0) = x_0$, $x(T) = x_1$, and the controls $u(\cdot)$ being subject to (4.2). As is usual, we assume that there exists at least one solution of (4.1) (with admissible control) satisfying (4.1a) for some finite T. Then the minimal time for the time-optimal problem, denoted T_0, is finite and

$$\inf h(\epsilon; x(\cdot); u(\cdot)) = h(\epsilon) \leqq T_0 \tag{4.4}$$

for every $\epsilon > 0$. Now because condition (H) of the previous section is satisfied, and of course also (3.1a), it readily follows that the infimum in (4.4) is attained. Let us denote the optimal upper limit by T_ϵ, let $x_\epsilon(t)$, $u_\epsilon(t)$ denote an optimal solution and let

$$z_\epsilon(t) = \dot{x}_\epsilon(t) - f(t; x_\epsilon(t)) - B(t)u_\epsilon(t).$$

Then as in §3, by perturbing $x_\epsilon(t)$ by a smooth function vanishing outside the compact interval of $(0, T_\epsilon)$, it follows that we must have that $z_\epsilon(t)$ satisfies

$$\dot{z}_\epsilon(t) + f_1(t; x_\epsilon(t))z_\epsilon(t) = 0, \qquad 0 < t < T_\epsilon, \tag{4.5}$$

where

$$f_1(t; x) = \nabla_x f(t; x).$$

Also because of the "pointwise" nature of the constraint (4.2) we can, by defining (analogous to (3.16))

$$H(\epsilon; z; x; u; t) = [z, f(t; x) + B(t)u], \qquad 0 < t < T_\epsilon, \tag{4.6}$$

proceed to obtain (as in (3.19) and (3.22)) the epsilon maximum principle that

$$\max_{\|u\| \leqq 1} H(\epsilon; z_\epsilon(t); x_\epsilon(t); u; t) = H(\epsilon; z_\epsilon(t); x_\epsilon(t); u_\epsilon(t); t) \quad \text{a.e.,} \quad 0 < t < T_\epsilon. \tag{4.7}$$

In fact it follows, because (4.1) is linear in $u(\cdot)$, that

$$\max_{\|u\| \leqq 1} [z_\epsilon(t), B(t)u] = [z_\epsilon(t), B(t)u_\epsilon(t)] \tag{4.8}$$

and hence that

$$u_\epsilon(t) = \frac{B(t)^* z_\epsilon(t)}{\| B(t)^* z_\epsilon(t) \|}, \tag{4.9}$$

where the denominator is not zero.

Next let us consider the situation as ϵ goes to zero. Now because (4.5) is linear in $z(t)$, we can write the solution as

$$z_\epsilon(t) = \psi(T_\epsilon ; t; \epsilon) z_\epsilon$$

for some z_ϵ, where $\psi(T_\epsilon ; T_\epsilon ; \epsilon)$ is the identity. Now because the functions $x_\epsilon(t)$ are equicontinuous, we can find a strongly pointwise convergent subsequence converging to $x_0(t)$ which is also absolutely continuous in $[0, T]$ and we know that T_ϵ increases monotonically to T_0. We can clearly find a subsequence of

$$\frac{z_\epsilon}{\| z_\epsilon \|}$$

that converges to a unit vector z_0. It follows that

$$\frac{z_\epsilon(t)}{\| z_\epsilon \|}$$

for a suitable subsequence converges for each t to $z_0(t)$, where $z_0(t)$ satisfies

$$\dot{z}_0(t) + f_1(t; x_0(t)) z_0(t) = 0, \qquad z_0(T) = z_0, \tag{4.10}$$

and $u_\epsilon(t)$ converges to

$$u_0(t) = \frac{B(t)^* z_0(t)}{\| B(t)^* z_0(t) \|}, \tag{4.11}$$

where the denominator is not zero. By using the weak convergence of $u_\epsilon(t)$ alone we know that, if we define (4.6) as

$$H(z; x; u; t) = [z, f(t; x) + B(t)u], \qquad 0 < t < T_0, \tag{4.12}$$

then

$$\max_{\|u\| \leq 1} H(z_0(t); x_0(t); u; t) = H(z_0(t); x_0(t); u_0(t); t) \quad \text{a.e.}, \quad 0 < t < T_0, \tag{4.13}$$

from which (4.11) would follow also.

We note that a similar development is possible in the infinite-dimensional case except that there is a major difficulty that

$$\frac{z_\epsilon}{\| z_\epsilon \|}$$

will now converge weakly only, in general, and the limit can be zero. If the state and control functions have their ranges in Hilbert spaces, for example, and (see also [4])

$$f(t; x) + B(t)u = Ax + Bu,$$

where A is the infinitesimal generator of a strongly continuous semigroup $T(t)$, (as in (3.1)), (4.5) becomes

$$\dot{z}_\epsilon(t) + A^* z_\epsilon(t) = 0, \qquad 0 < t < T_\epsilon,$$

so that

$$z_\epsilon(t) = T^*(T_\epsilon - t) z_\epsilon,$$

but as is known (see [5]), (4.11) is in general false. On the other hand, $u_\epsilon(t)$, which is necessarily of the form

$$\frac{B^* T^*(T_\epsilon - t) z_\epsilon}{\| T^*(T_\epsilon - t) z_\epsilon \|},$$

always yields an approximating sequence of controls.

REFERENCES

[1] A. V. Balakrishnan, *On a new computing technique in optimal control theory and the maximum principle*, Proc. Nat. Acad. Sci. U.S.A., 59 (1968), pp. 373–375.

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

[3] E. AXELBAND, *The optimal control of linear distributed parameter systems*, Doctoral thesis, Department of Engineering, University of California, Los Angeles, 1965.

[4] A. V. BALAKRISHNAN, *Optimal control problems in Banach spaces*, this Journal, 3(1965), pp. 109–127.

[5] J. L. LIONS, *Control problems in systems described by partial differential equations*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 251–272.

[6] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1(1962), pp. 76–84.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

ON THE NECESSITY OF A CERTAIN CONVEXITY CONDITION FOR LOWER CLOSURE OF CONTROL PROBLEMS*

PAVOL BRUNOVSKÝ†

1. Introduction. The existence problem of optimal controls has been investigated by many authors (see, e.g., [1]–[9]). This problem is closely connected with a property of control problems, which we shall call "lower closure" in this paper. The proof of this property is usually the crucial part of the proof of any existence theorem. It is based on a convexity assumption, which has subsequently been relaxed. The most general assumption of this type is that of Cesari (cf. [7], [8]). It is of some interest to know whether this assumption might be further relaxed. The main purpose of this paper is to give some insight into this problem. It is proved that in an important class of control problems Cesari's condition is actually necessary for lower closure, but that this is not true in general. As a by-product, some slight relaxations of the continuity part of the sufficient conditions for lower closure are obtained; by its applications we could obtain some improvements of the existence theorems.

2. Definitions. Let R^n denote the n-dimensional Euclidean space with the Euclidean norm denoted by $|x|$ for $x \in R^n$, $n = 1, 2, \cdots$. Further, denote $\rho(X, x) = \inf |x' - x|$, $N(x, \delta) = \{x' \mid |x' - x| < \delta\}$, $N(X, \delta) = \{x \mid \rho(X, x) < \delta\}$, $\rho(X, X') = \inf \{|x - x'| \mid x \in X, x' \in X'\}$, cl X and co X the closure and convex hull of X, respectively, for $x, x' \in R^n$, X, X' subsets of R^n.

DEFINITION 1. A mapping F of $D \subset R^m$ into the set of nonempty subsets of R^n will be called *β-continuous* if, for every $x_0 \in D$ and $\epsilon > 0$, there is a $\delta > 0$ such that for every $x \in N(x_0, \delta) \cap D$, $F(x) \in N(F(x_0), \epsilon)$. F will be called *α-continuous* if it is β-continuous and, moreover, for every $x_0 \in D$ and $\epsilon > 0$ there is a $\delta > 0$ such that for every $x \in N(x_0, \delta) \cap D$, $F(x_0) \subset N(F(x), \epsilon)$ is valid.

DEFINITION 2. F will be called *$\tilde{\beta}$-continuous* if, for every $x \in D$, $F(x) = \bigcap_{\delta > 0} \text{cl } F(N(x, \delta) \cap D)$. F will be called *$\tilde{\alpha}$-continuous* if it is $\tilde{\beta}$-continuous and, moreover, for every $x \in D$, $z \in F(x)$ and every sequence $\{x_k\}$, $x_k \to x$, $x_k \in D$, there is a sequence $z_k \in F(x_k)$ such that $z_k \to z$.

Both β-continuity and $\tilde{\beta}$-continuity are usually called upper semi-

* Received by the editors August 28, 1967, and in revised form January 22, 1968.

† Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Czechoslovakia. At present at Center for Control Sciences, University of Minnesota, Minneapolis, Minnesota 55455. The final version of this paper was prepared with the support of the National Aeronautics and Space Administration under Grant NGR 24-005-063.

continuity; α-continuity coincides with continuity of F in the Hausdorff set topology.

Obviously, if F is $\tilde{\beta}$-continuous, its values are closed sets. Further, β-continuity (α-continuity) implies $\tilde{\beta}$-continuity ($\tilde{\alpha}$-continuity); if the values of F are all contained in a compact subset of R^n, also the converse implications are true. It is also easy to prove that $\tilde{\beta}$-continuity is equivalent with the following property: $x_k \in D$, $k = 0, 1, \cdots$, $Z_k \in F(x_k)$, $k = 1, 2, \cdots$, $x_k \to x_0$, $z_k \to z_0$ implies $z_0 \in F(x_0)$ (if D is closed, in other words, the graph of F is closed). α-continuity coincides with continuity in Hausdorff set topology.

DEFINITION 3. The set-valued function F with closed values will be called *measurable* (*Borel-measurable*) if, for every closed set $Z \subset R^n$, the set $\{x \mid F(x) \subset Z\}$ is measurable (a Borel set).

A control problem (f^0, f, U) is given by a differential equation

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

$x = (x^1, \cdots, x^n) \in R^n$, $u = (u^1, \cdots, u^m) \in R^m$, $f = (f^1, \cdots, f^n)$, a scalar cost function $f^0(t, x, u)$ and a control domain $U(t, x)$. Here $U(t, x)$ is a set-valued function defined on a closed domain $D \subset R^{n+1}$ (such that $D = \operatorname{cl} \operatorname{int} D$), its values being closed subsets of R^m, and f and f^0 are defined on the set $\hat{D} = \{(t, x, u) \mid (t, x) \in D, u \in U(t, x)\}$. The pair (f, U) will be called the *control system*.

A pair of functions $u(t)\colon [t_1, t_2] \to R^m$, $x(t)\colon [t_1, t_2] \to R^n$ will be called the *control-trajectory pair* (CT-pair) of the control system (f, U) if $u(t)$ is measurable, $x(t)$ is absolutely continuous, $u(t) \in U(t, x(t))$ and $x(t)$ is a solution of the differential system (1) with $u = u(t)$ on $[t_1, t_2]$. $u(t)$ is called the *control* and $x(t)$ the corresponding *trajectory*.

Denote $\tilde{x} = (x^0, x) \in R^{n+1}$, $\tilde{f} = (f^0, f)$ and for a given CT-pair $\langle u(t), x(t)\rangle$, $x^0(t) = \int_{t_1}^{t} f^0(s, x(s), u(s))\, ds$.

DEFINITION 4. The control problem (f^0, f, U) will be called *lower closed* if it has the following property: Given a sequence $x_k(t)$, $u_k(t)$ of CT-pairs on a common interval $[t_1, t_2]$ such that the $\tilde{x}_k(t)$ converge uniformly towards an absolutely continuous function $\tilde{x}(t)$, then there is a control $u(t)$ on $[t_1, t_2]$ such that $x(t)$ is its corresponding trajectory and

$$\int_{t_1}^{t} f^0(s, x(s), u(s))\, ds \leqq x^0(t) \tag{2}$$

for all $t \in [t_1, t_2]$.

In optimal control problems, $I(u) = \int_{t_1}^{t_2} f^0(s, x(s), u(s))\, ds$ represents the cost functional which has to be minimized in a certain class of admissible

CT-pairs Ω (e.g., the class of CT-pairs $\langle u, x\rangle$ such that $x(t_1)$ and $x(t_2)$ are fixed). The proof of any existence theorem of optimal controls usually proceeds as follows (or this procedure is hidden in the proof): One finds a minimizing sequence of CT-pairs $\langle u_k, x_k\rangle$ such that

$$I(u_k) \to \inf_{\langle u,x\rangle \in \Omega} I(u).$$

From appropriate assumptions one concludes that from the sequence $\tilde{x}_k(t)$ a subsequence converging to a certain function $\tilde{x}(t)$ can be chosen. Now, the lower closure property allows us to prove that there is a $u(t)$ such that $x(t)$ is the corresponding trajectory and

$$I(u) = x^0(t_2) \leqq \lim_{k\to\infty} \inf x_k^0(t_2) = \lim_{k\to\infty} \inf I(u_k).$$

Thus u is the optimal control (of course, one has to prove that $\langle u, x\rangle$ belongs to Ω).

Let us note that our definition of lower closure is often not general enough, especially in the case of D or U unbounded, since then one is usually able to prove only pointwise convergence of the minimizing sequence $x_k^0(t)$ towards a function $x^0(t)$ of bounded variation (the convergence of $x_k(t)$ being uniform). In this case $x^0(t)$ in (2) has to be replaced by its absolutely continuous part (cf. [8], or [9, p. 25]). Also, the CT-pairs are often not defined on the same interval. However, since this is not essential for our discussion and our attention is concentrated mainly on the necessary conditions for lower closure, we have introduced this simplified definition to avoid unnecessary complications.

Denote

$$Q(t, x) = f(t, x, U(t, x)),$$

$$\tilde{Q}(t, x) = \{(z^0, z) \mid z^0 \geqq f^0(t, x, u),\ z = f(t, x, u),\ u \in U(t, x)\}$$

for $(t, x) \in D$.

CONDITION C. The set $\tilde{Q}(t, x)$ is convex for $(t, x) \in D$.

This is Cesari's convexity condition, the necessity of which for lower closure of control problems we are going to discuss.

3. Auxiliary lemmas. The following lemma is an analogue of Lusin's theorem for set-valued functions. It is a slight generalization of the result of Pliś [11] for set-valued functions with noncompact values.

LEMMA 1. *If the set-valued function F with closed values is measurable on a compact A, then it is asymptotically $\tilde{\alpha}$-continuous on A, i.e., for every $\epsilon > 0$ there is a closed subset A_ϵ of A such that F is $\tilde{\alpha}$-continuous on A_ϵ and* meas $(A - A_\epsilon) < \epsilon$.

PROOF. Denote $N_j = N(0,j) = \{y \mid |y| < j\}$, $E_j = \{x \mid F(x) \cap N_j \neq \varnothing\}$. The E_j are measurable and $\bigcup_{j=1}^{\infty} E_j = A$. For every j there is a finite covering $U_{j\nu}$, $\nu = 1, \cdots, p_j$, of N_j, $U_{j\nu}$ being open with diameter less than j^{-1}. By $K_{j\mu}$, $\mu = 1, \cdots, q_1$, $q_j = 2^{p_j} - 1$, we denote the subsets of the index set $\{1, \cdots, p_j\}$ ordered in an arbitrary manner. Denote

$$E_{j\mu} = \{x \mid F(x) \cap N_j \subset \bigcup_{\nu \in K_{j\mu}} \operatorname{cl} U_{j\nu}, F(x) \cap U_{j\nu} \neq \varnothing\};$$

the $E_{j\mu}$ are measurable and $\bigcup_\mu E_{j\mu} = E_j$.

Let $\epsilon > 0$ and let $\{\epsilon_{j\mu}\}_{\mu,j=1}^{\infty}$ be positive numbers such that $\sum_{\mu,j=1}^{\infty} \epsilon_{j\mu} < \epsilon$. Since the $E_{j\mu}$ are measurable, there are sets $B_{j\mu}$, $C_{j\mu}$ such that $B_{j\mu} \subset E_{j\mu}$, $C_{j\mu} \subset A - E_{j\mu}$ and meas $[A - (B_{j\mu} \cup C_{j\mu})] < \epsilon_{j\mu}$ for all μ. Further, there is a positive constant d_j such that $\rho(B_{j\mu}, C_{j\mu}) > d_j$ for all μ.

Denote $B_\epsilon = \bigcup_{j=1}^{\infty} \bigcup_{\mu=1}^{q_j} [A - (B_{j\mu} \cup C_{j\mu})]$, $A_\epsilon = A - B_\epsilon$. We have meas $B_\epsilon < \epsilon$ and

$$A_\epsilon = A - B_\epsilon = A - \bigcup_{j=1}^{\infty} \bigcup_{\mu=1}^{q_j} [A - (B_{j\mu} \cup C_{j\mu})]$$

$$= A - \bigcup_{j=1}^{\infty} \bigcup_{\mu=1}^{q_j} [E_{j\mu} - B_{j\mu}] \subset \bigcup_{j=1}^{\infty} \bigcup_{\mu=1}^{q_j} B_{j\mu}$$

since

$$A = \bigcup_{j=1}^{\infty} \bigcup_{\mu=1}^{q_j} E_{j\mu}.$$

Further, B_ϵ is open and, consequently, A_ϵ is closed.

To prove the $\tilde{\alpha}$-continuity of F on A_ϵ suppose first that F is not $\tilde{\beta}$-continuous on A_ϵ. Then there is a sequence $\{x_i\}_{i=1}^{\infty}$, $x_i \in A_\epsilon$, $x_i \to x_0$, $z_i \in F(x_i)$, $z_i \to z_0$, $z_0 \notin F(x_0)$. Since $F(x_0)$ is closed, there is a $\delta > 0$ such that $\rho(F(x_0), z_0) > \delta$. There is an i_0 such that for $i > i_0$, $|z_i - z_0| < \delta/2$. Further, there are j_0, μ_0 such that $F(x_0) \cap N_{j_0} \neq \varnothing$, $j_0^{-1} < \delta/2$, $x_0 \subset E_{j_0\mu_0}$ and, consequently, $N(z_0, \delta/2) \cap U_{i_0\mu_0} = \varnothing$. From this it follows $x_i \notin E_{j_0\mu_0}$, $|x_0 - x_i| > d_{j_0}$, for $i > i_0$, which is inconsistent with $x_i \to x_0$. Hence, F is $\tilde{\beta}$-continuous on A.

Suppose further that there is an $x_0 \in A_\epsilon$, $z_0 \in F(x_0)$, a sequence $x_i \to x_0$ such that for no sequence $\{z_i\}$, $z_i \in F(x_i)$, $z_i \to z_0$ is valid. Then passing to a subsequence, if necessary, we may suppose that there is a $\delta > 0$ such that $\rho(F(x_i), z_0) > \delta$. There is a j_0 such that $z_0 \in F(x_0) \cap N_{j_0}$, $j_0^{-1} < \delta$. Then, $x_0 \in E_{j_0\mu_0}$ for some μ_0, while $x_i \notin E_{j_0\mu_0}$, $i = 1, 2, 3, \cdots$. Hence, $|x_i - x_0| > d_{j_0}$, contrary to the assumption. This completes the proof.

Remark. A $\tilde{\beta}$-continuous set-valued function F is measurable. This follows from the fact that $\tilde{\beta}$-continuity of F implies that $\{x \mid F(x) \cap C \neq \varnothing\}$ is closed for every compact C and $\{x \mid F(x) \subset Z\} = A - \bigcup_{j=1}^{\infty} \{x \mid F(x) \cap$

$C_j \neq \varnothing\}$, $\{C_j\}_{j=1}^{\infty}$ being an arbitrary denumerable covering of $R^n - Z$ by compact sets C_j such that $C_j \cap Z \neq \varnothing$. Thus, from Lemma 1 it follows that a $\tilde{\beta}$-continuous set-valued function F on a compact A is asymptotically $\tilde{\alpha}$-continuous on A.

LEMMA 2. *Let $E(y)$ be a set-valued function on a closed domain $D \subset R^n$ with values being closed convex subsets of R^r. Let $E(y)$ be $\tilde{\alpha}$-continuous in D. Let $g:\{(y, v) \mid y \in D, v \in E(y)\} \to R^n$ be continuous and denote $F(y) = g(y, E(y))$. Then, given points $y_0 \in D$ and $z_0 \in F(y_0)$, there is a continuous function $\phi:D \to R^r$ such that $\phi(y_0) = z_0$ and $\phi(y) \in F(y)$, $y \in D$.*

Proof. Let v_0 be an arbitrary point of $E(y_0)$ such that $g(y_0, v_0) = z_0$. Then, according to [12, Theorem 3.1] there is a continuous function $\psi:D \to R^r$ such that $\psi(y_0) = v_0$ and $\psi(y) \in E(y)$ for $y \in D$. It is obvious that the function $\phi(y) = g(y, \psi(y))$ satisfies the desired properties.

The following example shows that if the convexity assumption in this lemma is dropped, it need not be valid even if F is connected. Let $m = r = n = 2$ and let g be the identical mapping. Identify the points of R^2 with complex numbers. Denote

$$F(0) = \{z \mid |z| \leqq 1\},$$

$$F(y) = (F(0) - \{z \mid |z| > 1, \operatorname{Re} z \in (-|y|, |y|)\}) \cdot e^{(i \operatorname{Arg} y)/2},$$

$$E(y) = F(y), \quad y \in R^2.$$

It is evident that no continuous function ϕ may be found so that $\phi(0) = 0$, $\phi(y) \in F(y)$.[1]

Remark. Lemma 2 can be used in the theory of contingent equations (orientor fields; cf. [5], [11]). Let in Lemma 2, $m = n + 1, y = (t, x)$, dim $x = n$. Then, $F(t, x)$ may be considered as an orientor field in R^n. From Lemma 2 it follows that through every point $(t_0, x_0) \in D$ passes a trajectory with an arbitrary given initial direction $z_0 \in F(t_0, x_0)$. (This trajectory is identical with the trajectory of the differential equations $\dot{x} = \phi(t, x)$.) In particular, this gives an existence theorem of solutions for fields of nonconvex orientors.

LEMMA 3. *Let $D \subset R^m$ be measurable and $f:D \to R^n$ be a measurable function. Then, for every Borel set $E \subset R^{m+n}$ the set $P_E = \{x \mid (x, f(x)) \in E\}$ is measurable.*

Proof. Denote $\mathfrak{S}$ the set of all subsets of R^{m+n} such that P_E is measurable. Then, $\mathfrak{S}$ clearly contains all sets of the form $A \times B$, $A \subset R^m$ being measurable, $B \subset R^n$ being a Borel set, and $\mathfrak{S}$ is a σ-algebra. Thus, $\mathfrak{S}$ contains any Borel set of R^{m+n}.

LEMMA 4. Let $D \subset R^l \times R^m$ *be a set such that* $D_x = \{y \mid (x, y) \in D\}$ *is closed for every x, and D_x is a measurable set-valued function of x,*

[1] Let us note that the counterexample [12, Example 6.1] does not fit to our case since Φ in it is not $\tilde{\alpha}$-continuous according to our definition.

$P = \{x \mid (x, y) \in D$ for some $y\}$ be compact. Let $f(x, y): D \to R^n$ be a function such that f is measurable in x for every fixed y and continuous in y for every fixed x. Then, for every $\epsilon > 0$ there is a closed set $F \subset P$ such that f is jointly continuous in x and y on $D \cap (F \times R^m)$ and meas $(P - F) < \epsilon$.

This lemma is a slight modification of Goodman's extension of Scorza-Dragoni's theorem ([13], cf. also [14]), the modification being in somewhat relaxed conditions on D. The detailed proof of this modification will be given in [15].

4. Main results.

THEOREM 1. *Suppose that*:

(i) *f^0, f are continuous in u for every (t, x) fixed and Borel-measurable in (t, x) for every u fixed over $\hat{D}$;*

(ii) *$U(t, x)$ is closed for every $(t, x) \in D$ and $U(t, x)$ is Borel-measurable on D;*

(iii) *$\tilde{Q}(t, x)$ is $\tilde{\alpha}$-continuous on D;*

(iv) *there is an $\tilde{\alpha}$-continuous set-valued function $E(t, x)$ on D, its values being convex subsets of some R^r, and a continuous mapping $g: \{(t, x, v) \mid (t, x) \in D, v \in E(t, x)\} \to R^n$ such that $Q(t, x) = g(t, x, E(t, x))$ for $(t, x) \in D$.*

Then, the condition C *is necessary for lower closure of the control problem (f^0, f, U). If* (i) *and* (ii) *are satisfied and, moreover, either*

(v) *$\tilde{Q}(t, x)$ is β-continuous on D,*

or

(vi) *$\tilde{Q}(t, x)$ is $\tilde{\beta}$-continuous and there is a nonnegative function $\phi(\xi)$, $\xi \geqq 0$, such that $\lim_{\xi \to \infty} \xi^{-1}\phi(\xi) = \infty$ and $f^0(t, x, u) \geqq \phi(|f(t, x, u)|)$ for $|f(t, x, u)|$ sufficiently large,*

then the condition C *is sufficient for lower closure of the control problem (f^0, f, U).*

Proof of necessity. If the condition C is violated, then there is a point $(t_0, x_0) \in \operatorname{int} D$ and $u_1, u_2 \in U(t_0, x_0)$ such that

$$\tfrac{1}{2}[\tilde{f}(t_0, x_0, u_1) + \tilde{f}(t_0, x_0, u_2)] \notin \tilde{Q}(t_0, x_0). \tag{3}$$

Denote $\tilde{x}_0 = (0, x_0)$. By Lemma 2, there are continuous functions $\tilde{\phi}_i(t, x)$, $\tilde{\phi} = (\phi^0, \phi)$, $i = 1, 2$, such that $\tilde{\phi}_i(t_0, x_0) = \tilde{f}(t_0, x_0, u_i)$ and $\tilde{\phi}_i(t, x) \in \tilde{Q}(t, x)$ for $(t, x) \in D$, $i = 1, 2$. There is a $\gamma > 0$ such that the set $A = \{(t, x) \mid |t - t_0| \leqq \gamma, |x - x_0| \leqq \gamma\} \subset D$.

Denote

$$\tilde{\psi}_k(t, x) = \begin{cases} \tilde{\phi}_1(t, x) & \text{for } t \in [t_0 + 2j\gamma k^{-1}, t_0 + (2j + 1)\gamma k^{-1}), \\ & \qquad j = 0, 1, \cdots, \tfrac{1}{2}k - 1, \\ \tilde{\phi}_2(t, x) & \text{for } t \in [t_0 + (2j + 1)\gamma k^{-1}, t_0 + (2j + 2)\gamma k^{-1}), \\ & \qquad j = 0, 1, \cdots, \tfrac{1}{2}k - 1, \end{cases}$$

$k = 2, 4, \cdots$. Since the $\tilde{\psi}_k(t, x)$ satisfy Carathéodory's conditions in A, there are solutions $\tilde{x}_k(t)$ of the systems

$$\dot{\tilde{x}} = \tilde{\psi}_k(t, x) \tag{4}$$

starting at $(t_0, \tilde{x}_0)$. Evidently, the $\tilde{\psi}_k(t, x)$ are uniformly bounded. Hence, there is a $t_1 > t_0$ such that $(t, \tilde{x}_k(t)) \in A$ for $t \in [t_0, t_1]$ and the $\tilde{x}_k(t)$ are equiabsolutely continuous on $[t_0, t_1]$. From this it follows that there is a subsequence of $\{\tilde{x}_k(t)\}$ which tends uniformly towards a certain absolutely continuous function $\tilde{x}(t)$ on $[t_0, t_1]$. Without loss of generality we may suppose that this is the original sequence.

Denote

$$\tilde{\phi}(t, x) = \tfrac{1}{2}[\tilde{\phi}_1(t, x) + \tilde{\phi}_2(t, x)].$$

We have for $t \in [t_0, t_1]$,

$$\begin{aligned}
\dot{\tilde{x}}(t) - \tilde{x}_0 - \int_{t_0}^{t} \tilde{\phi}(s, x(s))\, ds \\
&= \lim_{k\to\infty} \int_{t_0}^{t} [\tilde{\psi}_k(s, x_k(s))\, ds - \tilde{\phi}(s, x(s))\, ds \\
&= \lim_{k\to\infty} \int_{t_0}^{t} [\tilde{\psi}_k(s, x_k(s)) - \tilde{\psi}_k(s, x(s))]\, ds \\
&\qquad + \lim_{k\to\infty} \int_{t_0}^{t} [\tilde{\psi}_k(s, x(s)) - \tilde{\phi}(s, x(s))]\, ds.
\end{aligned} \tag{5}$$

The first term tends to zero, since the $\tilde{\psi}_k(t, x)$ are continuous in x, uniformly in $(t, x) \in A$ and k, and $x_k(t) \to x(t)$ uniformly in t. For the second term of (5), which we denote by I_k, we have

$$\begin{aligned}
I_k = \int_{t_0}^{t} [\tilde{\psi}_k(s, x(s)) - \tilde{\phi}(s, x(s))]\, ds = \sum_{j=0}^{j_0} \Bigg[\int_{t_0+2j\gamma k^{-1}}^{t_0+(2j+1)\gamma k^{-1}} \tilde{\phi}_1(s, x(s))\, ds \\
+ \int_{t_0(2j+1)\gamma k^{-1}}^{t_0+(2j+2)\gamma k^{-1}} \tilde{\phi}_2(s, x(s))\, ds - \int_{t_0+2j\gamma k^{-1}}^{t_0+(2j+2)\gamma k^{-1}} \tilde{\phi}(s, x(s))\, ds \\
+ \int_{t_0+(2j_0+2)\gamma k^{-1}}^{t} [\tilde{\psi}_k(s, x(s)) - \tilde{\phi}(s, x(s))]\, ds,
\end{aligned}$$

where j_0 is such that $(2j_0 + 2)\gamma k^{-1} < t < (2j_0 + 4)\gamma k^{-1}$. By the mean value theorem we have

$$\begin{aligned}
|I_k| = \frac{2\gamma}{k} \sum_{j=0}^{j_0} \Big| \frac{1}{2} [\tilde{\phi}_1(\sigma_{2j}, x(\sigma_{2j})) \\
+ \tilde{\phi}_2(\sigma_{2j+1}, x(\sigma_{2j+1})) - \phi(\sigma_j', x(\sigma_j', x(\sigma_j')))] \\
+ 2\gamma k^{-1} \max_{s\in[t_0+(2j+2)\gamma k^{-1}, t]} |\tilde{\psi}_k(s, x(s)) - \tilde{\phi}(s, x(s))|,
\end{aligned}$$

where $\sigma_i \in [t_0 + i\gamma k^{-1}, t_0 + (i + 1)\gamma k^{-1}]$, $\sigma_j' \in [t_0 + 2j\gamma k^{-1}, t_0$

$+ (2j + 2)\gamma k^{-1}]$. From the fact that ϕ_1, ϕ_2 are uniformly continuous in A and $\tilde{\psi}_k$, $\tilde{\phi}$ are bounded in A, it follows that $I_k \to 0$ for $k \to \infty$. Hence,

$$\dot{\tilde{x}}(t) = \tilde{x}_0 + \int_{t_0}^{t} \tilde{\phi}(s, x(s))\, ds.$$

This means that $\tilde{x}(t)$ is a trajectory of the system

$$\dot{\tilde{x}} = \tilde{\phi}(t, x(t)).$$

There are measurable functions $u_k(t)$ such that $\tilde{\psi}_k(t, x_k(t)) = \tilde{f}(t, x_k(t), u_k(t))$. To prove it note first that $U(t, x_k(t))$ is measurable. This follows from Theorem 1 (ii), Lemma 3 and $\{t \mid U(t, x_k(t)) \in Z\} = \{t \mid (t, x_k(t)) \in \{(t, x) \mid U(t, x) \subset Z\}\}$. By a similar argument $f(t, x_k(t), u)$ is measurable for every fixed u. Now, by Lemmas 1 and 4 for every $\epsilon > 0$, there is a compact subset I_ϵ of $[t_0, t_0 + \gamma]$ such that $\tilde{f}$ is jointly continuous in t, x, u, $U(t, x_k(t))$ is $\tilde{\beta}$-continuous on I_ϵ and meas $([t_0, t_0 + \gamma] - I_\epsilon) < \epsilon$. Using the formalism of [8, pp. 384–385] we can construct the desired measurable $u_k(t)$ on I_ϵ. Since $\epsilon > 0$ is arbitrary, this proves the existence of the desired $u_k(t)$ on $[t_0, t_0 + \gamma]$.

Now we can consider two cases in which (3) can be satisfied.

Case 1. $\frac{1}{2}[f(t_0, x_0, u_1) + f(t_0, x_0, u_2)] \notin Q(t_0, x_0)$.

Case 2. $\frac{1}{2}[f^0(t_0, x_0, u_1) + f^0(t_0, x_0, u_2)] < \min \{f^0(t_0, x_0, u) \mid f(t_0, x_0, u) = \frac{1}{2}f(t_0, x_0, u_1) + f(t_0, x_0, u_2)],\ u \in U(t, x)\}$.

In Case 1, from Theorem 1 (iii) and the continuity of ϕ it follows that $\phi(t, x) \notin Q(t, x)$ for (t, x) from a certain neighborhood B of (t_0, x_0). There is a $\tau > 0$ such that $(t, x(t)) \in B$ for $[t_0, t_0 + \tau]$. Since $\dot{x}(t) = \phi(t, x(t)) \notin Q(t, x(t))$ for a.e. $t \in [t_0, t_0 + \tau]$, $x(t)$ is not a trajectory of the control systems (f, U).

In Case 2 from Theorem 1 (iii) it follows again that

$$\phi^0(t, x) < \min \{f^0(t, x, u) \mid f(t, x, u) = \phi(t, x)\} \tag{6}$$

for (t, x) from a certain neighborhood of (t_0, x_0) and that there is a $\tau > 0$ such that $(t, x(t)) \in B$ for $t \in [t_0, t_0 + \tau]$. If $x(t)$ is a trajectory of (f, U), then there is a control $u(t)$ such that $\phi(t, x(t)) = f(t, x(t), u(t))$ for $t \in [t_0, t_0 + \tau]$. Then for every $t \in [t_0, t_0 + \tau]$, and any such $u(t)$, we have from (6) that

$$\begin{aligned}\int_{t_0}^{t_0+t} f^0(s, x(s), u(s))\, ds &> \int_{t_0}^{t_0+t} \phi^0(s, x(s))\, ds \\ &= \lim_{k\to\infty} \int_{t_0}^{t_0+t} \psi_k{}^0(s, x_k(s))\, ds \\ &\geqq \lim_{k\to\infty} \int_{t_0}^{t_0+t} f^0(s, x_k(s), u_k(s))\, ds,\end{aligned}$$

which completes the proof of the necessity part of the theorem.

Proof of sufficiency. Since the proof of the sufficiency part of the theorem is essentially covered by [8] and [9], we shall point out only the differences which arise from the slightly generalized assumptions.

First let us note that both from (v) and (vi) of Theorem 1 it follows that $\tilde{Q}(t, x)$ has property (Q) of [8]. If (v) is valid, this follows from the fact that cl $N(\tilde{Q}(t, x), \epsilon)$ is convex and, consequently, if $Q(t', x') \in N(Q(t, x), \epsilon)$ for all $(t', x') \in N((t, x), \delta) \cap D$, then cl co $\tilde{Q}((N(t, x), \delta) \cap D) \subset$ cl $N(\tilde{Q}(t, x), \epsilon)$.

If (vi) is valid, this follows from [9, Proposition 3] or [16].[2]

As in [8, p. 394] we may construct the auxiliary control problem, denoting $\tilde{u} = (u^0, u^1, \cdots, u^m)$, $\tilde{U}(t, x) = \{\tilde{u} = (u^0, u) \mid u \in U(t, x), u^0 \geqq f^0(t, x, u)\}$, $\hat{f} = \{\hat{f}^0, \hat{f}^1, \cdots, \hat{f}^n\}$ with $\hat{f}_0 = u_0$.

Now, if $U(t, x)$ was $\tilde{\beta}$-continuous and f^0, f continuous, we could obtain the desired result from [8, Closure Theorem II], but the set $\tilde{Q}(t, x)$ being the same for the auxiliary as for the original problem, this does not concern the proof of the fact that, for the limit trajectory $x(t)$ of a uniformly convergent sequence of trajectories $\tilde{x}_k(t)$, $\dot{\tilde{x}}(t) \in \tilde{Q}(t, x(t))$ a.e. on $[t_1, t_2]$. Thus, the only thing we have to prove is the existence of a measurable $\tilde{u}(t)$ such that

$$\dot{\tilde{x}}(t) = \tilde{f}(t, x(t), \tilde{u}(t)) \quad \text{a.e.} \tag{6}$$

As in the proof of the necessity part of the theorem we can prove that $U(t, x(t))$ is measurable and $\hat{f}(t, x(t), u)$ is measurable for every fixed u. Thus, by Lemmas 1 and 4 there is a closed subset I_ϵ of $[t_1, t_2]$ such that $\hat{f}(t, x(t), u)$ is jointly continuous in t, u on the set $G_\epsilon = \{(t, u) \mid t \in I_\epsilon, u \in U(t, x(t))\}$, $U(t, x(t))$ is $\tilde{\beta}$-continuous on I_ϵ and meas $([t_1, t_2] - I_\epsilon) < \epsilon$.

Now, let $t_k \to t$, $t_k \in I_\epsilon$, $\tilde{u}_k \to \tilde{u}$, $\tilde{u}_k \in \tilde{U}(t_k, x(t_k))$. Then, from the $\tilde{\beta}$-continuity of $U(t, x(t))$ on I_ϵ, the $\tilde{\beta}$-continuity of $\tilde{Q}$ and the continuity of $x(t)$, it follows that $u \in U(t, x(t))$, $u^0 \geqq f^0(t, x(t), u)$, i.e., $\tilde{u} \in \tilde{U}(t, x(t))$.

Hence, $\tilde{U}(t, x(t))$ is $\tilde{\beta}$-continuous on I_ϵ.

Since $\hat{f}$ is continuous on G_ϵ and $\tilde{U}(t, x(t))$ is $\tilde{\beta}$-continuous on I_ϵ, we can apply again the procedure of [8, pp. 384–385] to prove the existence of a measurable $\tilde{u}(t)$ satisfying (6) on I_ϵ. $\epsilon > 0$ being arbitrary, this proves the existence of $u(t)$ with desired properties on the whole $[t_1, t_2]$.

The necessity part of Theorem 1 seems to be of little value because the convexity condition C is replaced by another condition (iv). However, the following important corollary shows that in particular the theorem is valid if $U(t, x)$ is convex for every (t, x). The importance of this corollary lies in the fact that the convexity condition of U is usually natural from

[2] The author is grateful to Professor Czeslaw Olech who called his attention towards this fact.

the point of view of applications, which could not be said about the condition C. On the other hand, the counterexample of §3 indicates that the convexity assumption (iv) cannot be completely dropped.

COROLLARY 1. *Let* (i) *and* (iii) *of Theorem* 1 *be satisfied and let* f; f *be continous on* $\hat{D}$, $U(t,x)$ *be* $\tilde{\alpha}$*-continuous on* D, $U(t, x)$ *being convex for every* $(t, x) \in D$. *Then the condition* C *is necessary for lower closure of the control problem* (f^0, f, U).

For the proof it has only to be noted that the assumptions of Theorem 1 are satisfied by $r = m + 1, v = (v^0, u), E(t, x) = (-\infty, \infty) \times U(t, x)$,

$$g(t, x, u) = \begin{cases} (f^0(t, x, u), f(t, x, u)) & \text{if } v^0 < f^0(t, x, u), \\ (v^0, f(t, x, u)) & \text{if } v^0 \geqq f^0(t, x, u). \end{cases}$$

COROLLARY 2. *Let* (i), (ii) *and* (iii) *of Theorem* 1 *be valid. Further, let* $Q(t, x)$ *be convex for every* $(t, x) \in D$ *and let the mapping* $h: G \to R^1$, $G = \{(t, x, \hat{v}) \mid (t, x) \in D, \hat{v} \in Q(t, x)\} \subset R^{n+1}$ *defined by* $h(t, x, \hat{v}) = \min \{f^0(t, x, u) \mid f(t, x, u) = \hat{v}\}$ *be continuous on* G. *Then, the condition* C *is necessary for the lower closure of the problem* (f^0, f, U).

Again it is easy to verify that the assumptions of Theorem 1 are satisfied by $r = n + 1, E(t, x) = (-\infty, \infty) \times Q(t, x), v = (v^0, \hat{v})$,

$$g(t, x, v) = \begin{cases} (h(t, x, \hat{v}), \hat{v}) & \text{if } v^0 \leqq h(t, x, \hat{v}), \\ (v^0, \hat{v}) & \text{if } v^0 > h(t, x, \hat{v}). \end{cases}$$

5. Two counterexamples. The purpose of the following example is to show that the $\tilde{\alpha}$-continuity assumptions of Theorem 1 cannot be in general dropped. More precisely it shows that even if U, Q are compact, convex and β-continuous and the condition C is violated on a set of positive measure, then the problem can be lower closed.

Example 1. Let $n = m = 1, D = [0, 1], f(x, u) = u, f^0(x, u) = -u^2$. Denote G the union of open intervals $(a_{kj}, b_{kj}), k = 1, 2, 3, \cdots, j = 1, 2, \cdots, 2^{k-1}$, of length 2^{-2k} which are placed as follows. The center of (a_{11}, b_{11}) is $\frac{1}{2}$. The centers of $(a_{kj}, b_{kj}), j = 1, \cdots, 2^{k-1}$, coincide with centers of the equal segments which remain from $[0, 1]$ after removing the intervals $a_{\gamma j}, b_{\gamma j}, \gamma = 1, \cdots, k-1, j = 1, \cdots, 2^{\gamma-1}$. G is open and $F = [0, 1] - G$ is closed, meas G = meas $F = \frac{1}{2}$.

Define

$$U(x) = \begin{cases} \{-1 + 2(x - a_{kj})/(b_{kj} - a_{kj})\} & \text{for } x \in (a_{kj}, b_{kj}), \\ & k = 1, 2, \cdots, \ j = 1, 2, \cdots, 2^{k-1}, \\ [-1, 1] & \text{for } x \in F. \end{cases}$$

Evidently the condition C is violated on F. We shall prove that the problem (f^0, f, U) is lower closed.

Since $U(x)$ is compact for every $x \in [0, 1]$ and β-continuous on $[0, 1]$,

every limit of a uniformly convergent sequence of trajectories is a trajectory (cf. [1]).

(a) If $u(t), x(t)$ is a CT-pair on $[t_0, t_1]$ and for some $t' \in [t_0, t_1]$, $x(t') \in F$, then $x(t) = x(t')$ for $t \in [t', t_1]$ and, consequently, $u(t) = 0$ for a.e. $t \in [t', t_1]$. For the proof, suppose the contrary. Then there is a $t'' \in [t', t_1]$ such that $x(t'') \neq x(t')$. Let $x(t'') > x(t')$. Then there is an a_{kj} such that $x(t'') \leqq a_{kj} < x(t'')$. This is impossible, since $\dot{x}(t)$ has to be negative if $x(t) > a_{kj}$ and $x(t)$ is sufficiently close to a_{kj}. The case $x(t'' < x(t')$ may be treated similarly, a being replaced by b.

(b) From (a) it follows that for every CT-pair $u(t)$, $x(t)$, we have $u(t) = w(x(t))$ for a.e. t, where

$$w(x) = \begin{cases} 0 & \text{for } x \in F, \\ -1 + 2(x - a_{kj})/(b_{kj} - a_{kj}) & \text{for } x \in (a_{kj}, b_{kj}). \end{cases}$$

(c) Let $x(t)$ be a trajectory starting at a point (t_0, x_0) with $x_0 \in (a_{kj}, b_{kj})$. Then from (a) and (b) it follows that for $t \geqq t_0 + \frac{1}{2}(b - a) \cdot \ln [\frac{1}{2}(b - a)|x_0 - \frac{1}{2}(a + b)|^{-1}]$, $x(t) = a_{kj}$ if $x_0 < \frac{1}{2}(a_{kj} + b_{kj})$, and $x(t) = b_{kj}$ if $x_0 > \frac{1}{2}(a_{kj} + b_{kj})$; if $x_0 = \frac{1}{2}(a_{kj} + b_{kj})$, then $x(t) = x_0$ for $t \geqq t_0$.

(d) Let $x_k(t)$ be a sequence of trajectories which tends uniformly to a trajectory $x(t)$ on $\langle t_0, t_1 \rangle$. We have to prove

$$\lim_{k \to \infty} \int_{t_0}^{t_1} f^0(t, x_k(t), w(x_k(t)))\, dt = \int_{t_0}^{t_1} f^0(t, x(t), w(x(t)))\, dt. \tag{7}$$

First suppose $x(t_0) \in F$. Then, divide the sequence $x_k(t)$ into two subsequences: the first consisting of those $x_k(t)$ with $x_k(t_0) \in F$, the second consisting of the remaining elements. For the first subsequence, (7) holds, since the integrands vanish for all k. As for the second one, the integrands are bounded and by (c) they are nonzero only on an interval the length of which tends to zero. If $x(t_0) \in G$, then (7) follows from the continuity of w on G, (a) and (c).

The following example shows that Cesari's property (Q) of $\tilde{Q}$ is not necessary for lower closure even if (i), (ii) and (iii) of Theorem 1 are satisfied (of course, then (v) and (vi) have to be violated, since otherwise this property follows from the condition C). Recall that $\tilde{Q}$ is said to have property (Q) if $\tilde{Q}(t, x) = \bigcap_{\delta > 0} \operatorname{cl} \operatorname{co} \tilde{Q}((N(t, x), \delta) \cap D)$. In this example $\tilde{Q}$ has property (Q) nowhere in D.

Example 2. Let $n = 2$, $m = 1$, $D = R^2$, $U = R^1$, $f^1(x, u) = u^1$, $f^2(x, u) = x^1 u^1$, $f^0(x, u) \equiv 0$. That $\tilde{Q}$ has property (Q) nowhere in D follows from the fact that for every $x \in R^2$ and $\delta > 0$, $\operatorname{cl} \operatorname{co} \tilde{Q}(N(x, \delta)) = (0, \infty) \times R^2$ while $\tilde{Q}(x) = (0, \infty) \times \{(z^1, z^2) | z^2 = x^1 z^1\}$. Let $x(t)$ be a trajectory. Then we have $d(x^2 - \frac{1}{2}(x^1)^2)/dt = 0$ so that the graph of it is a part of some

parabola

$$x^2 = \tfrac{1}{2}(x^1)^2 + c, \tag{8}$$

where $c \in (-\infty, \infty)$. On the other hand, every $x(t)$ such that $x^1(t)$ is absolutely continuous and $x^2(t) = \frac{1}{2}[x^1(t)]^2 + c$ is a trajectory. Now, it is obvious that the graph of a limit of a uniformly convergent sequence of trajectories is a part of a parabola (8). Thus if this limit is absolutely continuous, it is a trajectory. Since $f^0 \equiv 0$, this proves the lower closure of the problem (f^0, f, U).

REFERENCES

[1] A. F. Filippov, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[2] ———, *Differential equations with multi-valued discontinuous right-hand side*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 65–68.

[3] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[4] E. Roxin, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[5] T. Ważewski, *On an optimal control problem*, Proc. Conference on Differential Equations and Their Applications, Publ. house of the Czech. Acad. of Sciences, Prague, 1962. See also a series of papers in Bull. Acad. Polon. Sci. Sér. Math. Astronom. Phys., 9 (1961), pp. 861–863, 865–867, 869–872; 10 (1962), pp. 11–15, 17–21.

[6] L. Cesari, *An existence theorem in problems of optimal control*, this Journal, 3 (1965), pp. 7–22.

[7] ———, *Existence theorems for optimal solutions in Pontryagin and Lagrange problems*, this Journal, 3 (1965), pp. 475–498.

[8] ———, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints I, II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–430.

[9] C. Olech, *Existence theorems for optimal problems with vector-valued cost function*, Tech. rep. 67-6, Center for Dynamical Systems, Brown University, Providence, 1967.

[10] S. C. Zaremba, *Sur les équations au paratingent*, Bull. Sci. Math., 60 (1936), pp. 139–160.

[11] A. Plíś, *Remark on measurable set-valued functions*, Bull. Acad. Polon. Sci. Sér. Math. Astronom. Phys., 9 (1961), pp. 857–859.

[12] E. Michael, *Continuous selections I*, Ann. of Math., 63 (1956), pp. 361–381.

[13] C. S. Goodman, *On a theorem of Scorza-Dragoni and its application to optimal control*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 222–233.

[14] M. Q. Jacobs, *Remark on some recent extensions of Filippov's implicit functions lemma*, this Journal, 5 (1967), pp. 622–627.

[15] P. Brunovský, *Scorza-Dragoni's theorem for set-valued functions and its applications to control problems*, Matematicky Časopis Sloven. Akad. Vied., to appear.

[16] L. Cesari, *Existence theorems for optimal controls of the Mayer type*, this Journal, to appear.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

L_p-STABILITY OF LINEAR TIME-VARYING FEEDBACK SYSTEMS*

CHI-TSONG CHEN†

1. Introduction. Consider the single-loop feedback system shown in Fig. 1, where G is a linear time-invariant nonanticipative system and N is a memoryless element. It is well known that if N is a constant gain and if the Nyquist criterion is satisfied, then for any class of input function u, the output y will be in the same class of function [1]. For example, if the input is bounded, so is the output; if the input is of finite energy, so is the output. If N is a nonlinear element, then a stability has to be defined with respect to a specific class of input [2]–[5]. In this paper we consider the case where N is a time-varying memoryless element; more specifically, N is characterized by $x(t) = k(t)e(t)$, where $k(t)$ is the time-varying gain, x and e are the output and input of N, respectively. Clearly, this feedback system is linear. Let $g(t)$ be the impulse response of G and $h(t, \tau)$ be the impulse response of the entire feedback system. $h(t, \tau)$ is, by definition, the output response at time t due to a δ-function input applied at time τ. Suppose the feedback path in Fig. 1 is disconnected for a moment and a δ-function input is applied at time τ; then the output of G will be $k(\tau)g(t - \tau)$ for $t \geqq \tau$. Now if the feedback path is connected, then $k(\tau)g(t - \tau)$ will go to the output and the input of the entire feedback system as shown in Fig. 2. Hence we have

$$(1) \qquad h(t,\tau) = \begin{cases} k(\tau)g(t-\tau) - \displaystyle\int_{\tau}^{t} h(t,v)k(\tau)g(v-\tau)\,dv & \text{for } t \geqq \tau, \\ 0 & \text{for } t < \tau. \end{cases}$$

If $h(t, \tau)$ can be solved from (1) and if $\displaystyle\int_{-\infty}^{t} |h(t, \tau)|\,d\tau < c < \infty$ for all t, then in the feedback system any bounded input generates a bounded output [6]. However, solving for $h(t, \tau)$ from (1) is very difficult, if not impossible; hence this approach is not feasible. Furthermore, it is desirable to state the stability condition in terms of the open loop system.

The time-varying gain $k(t)$ considered in this paper is assumed to be of the form $k_0 + \tilde{k}(t)$, where k_0 is a positive constant. It will be shown that if the gain deviation $\tilde{k}(t)$ is bounded and absolutely integrable, then this gain deviation can be neglected in considering stabilities of the feedback sys-

* Received by the editors August 16, 1967, and in revised form November 21, 1967.

† College of Engineering, State University of New York at Stony Brook, Stony Brook, Long Island, New York 11790. This research was supported by the United States Air Force Office of Scientific Research under Grant AF-AFOSR-542-67.

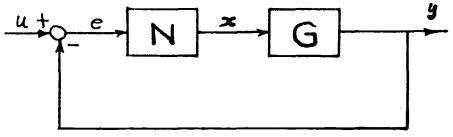


FIG. 1. *Continuous feedback system*

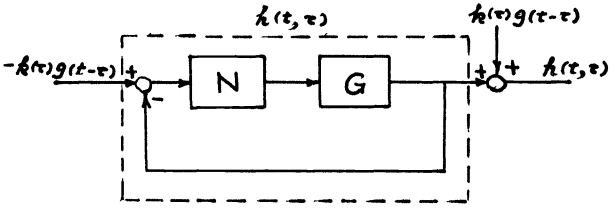


FIG. 2. *The relationship between $g(t)$ and $h(t, \tau)$*

tem. A system with sector condition on $k(t)$, $0 \leqq k(t) \leqq k_1$, is studied in [7], [8].

2. Statement of the problem and some lemmas. Given the feedback control system shown in Fig. 1, N is a memoryless time-varying element and G is a nonanticipative linear time-invariant system. It is assumed that G is characterized by a convolution integral. More precisely, let g be the impulse response of G; then the input x and output y of G are related by

$$y(t) = z(t) + \int_0^t g(t - \tau)x(\tau)\, d\tau, \qquad t \geqq 0, \tag{2}$$

where $z(t)$ is the zero-input response of G. It is assumed that G satisfies the following conditions:

(A1) For all initial conditions, the zero-input response z is an element of $L_p[0, \infty)$, where p is a fixed number in $[1, \infty]$.

(A2) The impulse response $g(t)$ is of the form

$$g(t) = r + g_1(t), \qquad t \geqq 0, \tag{3}$$

where r is a nonnegative constant and $g_1(t)$ is bounded on $[0, \infty)$, is an element of $L_1[0, \infty)$ and tends to zero as t tends to infinity. It is also assumed that g_1 is either continuous or of bounded variation.

Let $G(s)$ be the Laplace transform of g, i.e.,

$$G(s) \triangleq \int_0^\infty g(t)e^{-st}\, dt.$$

Here $G(s)$ is not necessarily a rational function of s. Let e and x be the input and output of the time-varying element N. It is assumed that N is charac-

terized by

$$x(t) = k(t)e(t)$$

and $k(t)$ is of the form

$$k(t) = k_0 + \tilde{k}(t),$$

where k_0 is a positive constant and $\tilde{k}(t)$ is a continuous function of t, $\tilde{k}$ being the gain deviation from k_0.

A feedback system with $\tilde{k}(t) \equiv 0$ is called a "constant gain" feedback system. Let $\bar{h}$ be the impulse response of a constant gain feedback system with $k(t) = k_0$. Then

$$\bar{h}(t) = \begin{cases} k_0\, g(t) - k_0 \int_0^t g(t-\tau)\bar{h}(\tau)\, d\tau & \text{for} \quad t \geqq 0, \\ 0 & \text{for} \quad t < 0. \end{cases} \tag{4}$$

Before proceeding we need the following lemmas.

LEMMA 1. *Let $f_1(t)$ and $f_2(t)$ be any real-valued functions defined on $[0, \infty)$. If f_2 is an element of $L_p[0, \infty)$, $1 \leqq p \leqq \infty$, and if f_1 is an element of $L_1[0, \infty)$ and $L_\infty[0, \infty)$ and tends to zero as t tends to infinity, then the convolution*

$$(f_2 * f_1)(t) \triangleq \int_0^t f_1(t-\tau) f_2(\tau)\, d\tau$$

tends to zero as t tends to infinity.

Proof. Let p' be such that $1/p' + 1/p = 1$. By using the Hölder inequality we obtain

$$\begin{aligned} &|(f_2 * f_1)(t)| \\ &\quad \leqq \int_0^t |f_1(t-\tau)|^{1/p'} |f_1(t-\tau)|^{1/p} |f_2(\tau)|\, d\tau \\ &\quad \leqq \left(\int_0^t |f_1(t-\tau)|\, d\tau\right)^{1/p'} \left(\int_0^t |f_1(t-\tau)|\, |f_2(\tau)|^p\, d\tau\right)^{1/p}. \end{aligned} \tag{5}$$

Since $|f_2|^p$ is an element of $L_1[0, \infty)$, it is shown in [9] that $(f_1 * |f_2|^p)(t)$ tends to zero as $t \to \infty$. Hence from (5) we conclude that $f_2 * f_1$ tends to zero as $t \to \infty$.

LEMMA 2. *Let ω, ϕ, ψ, and θ be real-valued functions defined on $[0, \infty)$ and either continuous or of bounded variation. Let*

$$\omega(t) \leqq \phi(t) + \psi(t) \int_t^\infty \theta(\tau)\omega(\tau)\, d\tau \quad \textit{for all} \quad t \in [0, \infty).$$

If $\theta(t)$ and $\psi(t)$ are nonnegative for all $t \in [0, \infty)$, then

$$\omega(t) \leqq \phi(t) + \psi(t) \int_t^\infty \phi(\tau)\theta(\tau) \exp \int_t^\tau \psi(v)\theta(v)\, dv\, d\tau$$

for all $t \in [0, \infty)$.

This is a modified version of the Bellman-Gronwall lemma [13]. This can be proved by defining $\nu(t) \triangleq \int_t^\infty \theta(\tau)\omega(\tau)\, d\tau$ and $\nu(\infty) \triangleq 0$.

3. Main result. We state the main result as a theorem.

THEOREM 1. *If the feedback system shown in Fig. 1 satisfies the assumptions* (A1) (A2) *and if, in addition, satisfies*

(a) $\inf_{\mathrm{Re}\, s \geqq 0} |1 + k_0 G(s)| > 0$,

(b) $\tilde{k}$ *is a continuous function and an element of* $L_1[0, \infty)$ *and* $L_\infty[0, \infty)$, *then*

(i) *the output y is an element of $L_p[0, \infty)$ if the input u and the zero-input response z are elements of $L_p[0, \infty)$, $1 \leqq p \leqq \infty$;*

(ii) *for any initial state,*

$$\lim_{t \to \infty} |y(t) - \bar{y}(t)| = 0,$$

where $\bar{y}$ is the output of the corresponding constant gain feedback system.

In the proof, we shall use a comparison technique and the following facts: Let $h(t, \tau)$ be the impulse response function of a nonanticipative linear time-varying system. Then (a) any bounded input generates a bounded output if and only if $\int_{-\infty}^t |h(t, \tau)|\, d\tau \leqq c < \infty$ for all t (see [6]); (b) if $\int_{-\infty}^t |h(t, \tau)|\, d\tau \leqq c_1 < \infty$ for all t and $\int_\tau^\infty |h(t, \tau)|\, dt \leqq c_2 < \infty$ for all τ, and if, in addition, the input is an element of $L_p[0, \infty)$, $1 \leqq p \leqq \infty$, then so is the output [11].

Proof of Theorem 1. Consider the constant gain feedback system with $k(t) = k_0$. It is shown in [1] that if the assumption (A1), (A2) and the condition (a) in the theorem are satisfied, then the impulse response of the constant gain feedback system, $\bar{h}(t)$, is an element of $L_1[0, \infty)$ and $L_\infty[0, \infty)$ and tends to zero as $t \to \infty$. It follows that the output of the constant gain feedback system $\bar{y}$,

$$\bar{y}(t) = z(t) + \int_0^t \bar{h}(t - \tau)\, [u(\tau) - z(\tau)]\, d\tau, \qquad t \geqq 0, \tag{6}$$

is an element of $L_p[0, \infty)$ if u and z are elements of $L_p[0, \infty)$, $1 \leqq p \leqq \infty$.

The output of the time-varying feedback system is given by

$$y(t) = z(t) + k_0 \int_0^t g(t-v)\,[u(v) - y(v)]\,dv + \int_0^t g(t-v)\tilde{k}(v)\,[u(v) - y(v)]\,dv. \tag{7}$$

Under some manipulation and using (4) and (6), (7) can be written as

$$y(t) = \bar{y}(t) + k_0^{-1} \int_0^t \bar{h}(t-\tau)\tilde{k}(\tau)\,[u(\tau) - y(\tau)]\,d\tau. \tag{8}$$

Define $\tilde{y}(t) \triangleq y(t) - \bar{y}(t)$, $\tilde{u}(t) \triangleq u(t) - \bar{y}(t)$; then

$$\tilde{y}(t) = k_0^{-1} \int_0^t \bar{h}(t-\tau)\tilde{k}(\tau)\,[\tilde{u}(\tau) - \tilde{y}(\tau)]\,d\tau. \tag{9}$$

This is a key equation in our proof. Instead of studying $y(t)$ directly, we shall compare the output of the feedback system and the output of the corresponding constant gain feedback system. Since $\bar{y}$ is an element of L_p, if we succeed in proving that $\tilde{y}$ is an element of L_p, then by the Minkowski inequality, we can conclude that y is an element of L_p (if u and z are elements of L_p).

Consider the linear time-varying feedback system shown in Fig. 3, where the impulse response function of $\bar{H}$ is $\bar{h}(t)$ which is defined in (4). Clearly Fig. 3 is a block diagram representation of (9). Let $\tilde{h}(t,\tau)$ be the impulse response of the feedback system shown in Fig. 3. Then, similarly to (1),

$$\tilde{h}(t,\tau) = \begin{cases} k_0^{-1}\tilde{k}(\tau)\bar{h}(t-\tau) - \int_\tau^t \tilde{h}(t,v)k_0^{-1}\tilde{k}(\tau)\bar{h}(v-\tau)\,dv & \text{for } t \geqq \tau, \\ 0 & \text{for } t < \tau, \end{cases} \tag{10}$$

and $\tilde{y}$ and $\tilde{u}$ are related by

$$\tilde{y}(t) = \int_\tau^t \tilde{h}(t,\tau)\tilde{u}(\tau)\,d\tau. \tag{11}$$

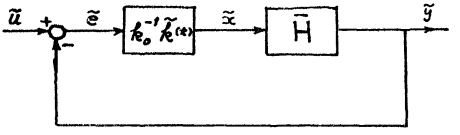


FIG. 3. *Linear time-varying feedback system*

We shall show now that (a) $\int_{-\infty}^{t} |\tilde{h}(t,\tau)|\,d\tau \leqq c_1 < \infty$ for all t, and (b) $\int_{\tau}^{\infty} |\tilde{h}(t,\tau)|\,dt \leqq c_2 < \infty$ for all τ.

(a) From (9), if $\tilde{u} \in L_\infty[0,\infty)$, then

$$|\tilde{y}(t)| \leqq k_0^{-1} \int_0^t |\bar{h}(t-\tau)|\,|\tilde{k}(\tau)|\,|\tilde{u}(\tau)|\,d\tau + k_0^{-1} \int_0^t |\bar{h}(t-\tau)|\,|\tilde{k}(\tau)|\,|\tilde{y}(\tau)|\,d\tau \tag{12}$$

$$\leqq k_0^{-1}\bar{h}_m\,\tilde{u}_m \int_0^t |\tilde{k}(\tau)|\,d\tau + k_0^{-1}\bar{h}_m \int_0^t |\tilde{k}(\tau)|\,|\tilde{y}(\tau)|\,d\tau,$$

where $\bar{h}_m \triangleq \sup_t |\bar{h}(t)|$, $\tilde{u}_m \triangleq \sup_t |\tilde{u}(t)|$. Applying the Bellman-Gronwall lemma [10], [12, p. 35] and using the assumption that $\tilde{k} \in L_1[0,\infty)$, it can readily be shown that $\tilde{y}(t)$ is bounded for all t. Equivalently, in (11), any bounded $\tilde{u}$ gives a bounded $\tilde{y}$. Hence we conclude that $\int_{-\infty}^{t} |\tilde{h}(t,\tau)|\,d\tau \leqq c_1 < \infty$ for all t.

(b) From (10),

$$|\tilde{h}(t,\tau)| \leqq k_0^{-1}\,|\tilde{k}(\tau)|\,|\bar{h}(t-\tau)| + k_0^{-1}\bar{h}_m \int_\tau^\infty |\tilde{h}(t,s)|\,|\tilde{k}(\tau)|\,ds. \tag{13}$$

Taking the integration with respect to t,

$$\int_\tau^\infty |\tilde{h}(t,\tau)|\,dt \leqq k_0^{-1}\,|\tilde{k}(\tau)| \int_\tau^\infty |\bar{h}(t-\tau)|\,dt + k_0^{-1}h_m\,|\tilde{k}(\tau)| \int_\tau^\infty\int_\tau^\infty |\tilde{h}(t,s)|\,ds\,dt. \tag{14}$$

Changing the order of integration [14, p. 234] and noting that $\tilde{h}(t,s) = 0$ for $t < s$, (14) becomes

$$\int_\tau^\infty |\tilde{h}(t,\tau)|\,dt \leqq k_0^{-1}\,|\tilde{k}(\tau)|\,c_h + k_0^{-1}\bar{h}_m\,|\tilde{k}(\tau)| \int_\tau^\infty \left(\int_s^\infty |\tilde{h}(t,s)|\,dt\right) ds, \tag{15}$$

where $c_h \triangleq \int_0^\infty |\bar{h}(t)|\,dt$. Define

$$f(\tau) \triangleq \int_\tau^\infty |\tilde{h}(t,\tau)|\,dt. \tag{16}$$

Then (15) becomes

$$f(\tau) \leqq k_0^{-1} \, |\tilde{k}(\tau)| \, c_h + k_0^{-1}\bar{h}_m \, |\tilde{k}(\tau)| \int_\tau^\infty f(s)\, ds. \tag{17}$$

Now applying Lemma 2, we obtain

$$\begin{aligned} |f(\tau)| &\leqq k_0^{-1} \, |\tilde{k}(\tau)| \, c_h \\ &\quad + k_0^{-1}\bar{h}_m \, |\tilde{k}(\tau)| \int_\tau^\infty k_0^{-1} \, |\tilde{k}(s)| \, c_h \exp \int_\tau^s k_0^{-1}\bar{h}_m \, |\tilde{k}(l)| \, dl \, ds. \end{aligned} \tag{18}$$

Since $\tilde{k}$ is bounded and absolutely integrable, from (18) we conclude that $|f(\tau)|$ is bounded on $(-\infty, \infty)$. It follows that $\int_\tau^\infty |\tilde{h}(t,\tau)|\, dt \leqq c_2 < \infty$ for all τ.

We have proved that $\tilde{h}(t,\tau)$ is absolutely integrable with respect to t and τ; hence from (11) we conclude that $\tilde{y} \in L_p$ if $\tilde{u} \in L_p$. Now $y = \tilde{y} + \bar{y}$; hence $y \in L_p$ if $u \in L_p$ and $z \in L_p$. This proves the first part of Theorem 1. From (8) we have

$$|y(t) - \bar{y}(t)| \leqq k_0^{-1}\bar{k}_m \int_0^t |\bar{h}(t-\tau)| \, [|u(\tau)| + |y(\tau)|] \, d\tau. \tag{19}$$

Since $\bar{h}$ is an element of L_1 and L_∞ and tends to zero as $t \to \infty$, and since u and y are elements of L_p, by applying Lemma 1, we know that the right-hand side of (19) tends to zero as $t \to \infty$. Consequently,

$$\lim_{t\to\infty} |y(t) - \bar{y}(t)| = 0.$$

The frequency domain interpretation of $\inf_{\mathrm{Re}\, s \geqq 0} |1 + k_0 G(s)| > 0$ is given in [1]. With slight modification Theorem 1 can be easily extended to multiple loop feedback systems. Furthermore, if N is a nonlinear time-varying element and is characterized by $x = \phi(e,t)$ and $\phi(e,t) = k_0 e + \tilde{\phi}(e,t)$, and if $|\tilde{\phi}(e,t)| \leqq \tilde{k}(t)|e|$, then all the results in the paper still apply to this class of nonlinear time-varying feedback systems.

REFERENCES

[1] C. A. Desoer, *On the general formulation of the Nyquist criterion*, IEEE Trans. Circuit Theory, CT-12 (1965), pp. 230–234.

[2] V. M. Popov, *Absolute stability of nonlinear systems of automatic control*, Automat. Remote Control, 22 (1962), pp. 857–875.

[3] Y. Z. Tsypkin, *Frequency criteria for the absolute stability of nonlinear sampled-data systems*, Ibid., 25 (1964), pp. 261–267.

[4] C. T. Chen, *On the stability of nonlinear sampled-data feedback systems*, J. Franklin Inst., 280 (1965), pp. 316–324.

[5] I. W. SANDBERG, *On the boundedness of solutions of nonlinear functional equations*, Bell Syst. Tech. J., 44 (1965), pp. 439–453.

[6] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

[7] I. W. SANDBERG, *On generalizations and extensions of the Popov criterion*, IEEE Trans. Circuit Theory, CT-13 (1966), pp. 117–118.

[8] C. T. CHEN, *On the stability of sampled-data feedback systems with time-varying gain*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 624–625.

[9] ———, *On the stability of feedback control systems with perturbation gain*, Notes on System Theory, vol. VII, Electronics Research Laboratory, University of California, Berkeley, 1965, pp. 47–56; J. Franklin Inst., to appear.

[10] G. SANSONE AND R. CONTI, *Nonlinear Differential Equations*, Macmillan, New York, 1964.

[11] R. E. EDWARDS, *Functional Analysis, Theory and Applications*, Holt, Rinehart and Winston, New York, 1965.

[12] R. BELLMAN, *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1953.

[13] G. S. JONES, *Fundamental inequalities for discrete and discontinuous functional equations*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 43–57.

[14] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1963.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

OPTIMAL CONTROL OF PARTIALLY OBSERVABLE DIFFUSIONS*

WENDELL H. FLEMING†

Summary. The problems considered are stochastic analogues of the problem of Lagrange in calculus of variations. The response to the control is assumed to be a diffusion process, and the controls admitted are based on partial observations of the current states of the response. The problem can then be phrased as one of optimally controlling the coefficients of linear second order parabolic equations. An existence theorem in the class of bounded, measurable controls and necessary conditions in terms of conditional expectations are obtained.

1. Introduction. The problem of Lagrange in calculus of variations is to find the extrema of some variational integral subject to end conditions and to a system of ordinary differential equations as side conditions. Similar problems occur in the theory of optimal control, in which one usually has inequality constraints on the control variables of the problem. Lagrange multipliers are introduced, which are functions of time satisfying a linear system of differential equations dual to the linearized state equations. One then has the classical necessary conditions of calculus of variations for an extremum, and for optimal control problems, Pontryagin's maximum principle (see, for example, [11]).

We are interested here in stochastic extremum problems like the problem of Lagrange, in which one has as side conditions a system of stochastic (rather than ordinary) differential equations. The problem of finding appropriate necessary conditions for an extremum in such problems has been settled only in certain cases. In this paper the stochastic differential equations have the form (2.1′) below. The controls Y admitted are based on observing at each time t the current state $\xi(t)$ of the solution, which is a diffusion process called the response to the control Y. (By diffusion is meant a Markov process with continuous sample paths.)

When $\xi(t)$ is completely observable, we called the problem Markovian (see [4], [6]) and used a dynamic programming method which reduced the problem to the solution of a nonlinear second order parabolic equation (see (5.9) below). In the present paper we suppose that $\xi(t)$ is only partially observable. The problem can be formulated in terms of linear second order parabolic equations, whose coefficients involve the control function Y. By varying Y we vary the coefficients of the backward operator of the response process, and from that a necessary condition for a minimum is deduced (§4, Theorem 1). Under some further assumptions we obtain in §5 stronger

* Received by the editors August 7, 1967.

† Mathematics Department, Brown University, Providence, Rhode Island 02912. This research was supported by the National Science Foundation under Grant GP-6733.

necessary conditions and a theorem about the existence of a minimizing control. These necessary conditions amount to solving simultaneously boundary problems for the forward and backward operators of the response process together with a minimum condition of the Pontryagin type involving conditional expectations. Similar necessary conditions have been derived independently by Mieri [15], by a formal calculation involving the semigroups generated by the backward and forward operators.

When the response is partially observable the controller does better by remembering past observations up to time t than by observing only the current state of the response at time t. By admitting only controls based on current observations we are considering a somewhat artificial problem, except in two extreme cases (no observations and complete observations of the response). However, if this problem can be solved, then at least an upper bound is obtained for the minimum in the more difficult problem where controls based on past observations are admitted. By an extension of the method, another approximation to this minimum can be obtained using controls based on observations at a fixed finite set of times (see §7).

A different approach to a stochastic variational calculus is to look for Lagrange multiplier processes which satisfy a system of stochastic differential equations dual (in some sense) to the linearized system equations (2.1′). If the matrix σ in (2.1′) is constant, then the dual system consists of ordinary (not stochastic) differential equations. In that case, Kushner [12] derived the relevant variational formula, provided moreover that the response process is stopped at a fixed time τ. When σ is not constant, well-known difficulties are encountered, but some partial results have been found (see [7]). This approach and the one in the present paper are compared in §6.

2. Probabilistic formulation. Let us first formulate the minimum problem in terms of stochastic processes, and then in §3 turn it into a problem about parabolic partial differential equations.

Let $T > 0$ be fixed throughout, and let t, $x = (x_1, \cdots, x_n)$, $y = (y_1, \cdots, y_p)$ denote, respectively, points of the interval $[0, T]$, of Euclidean R^n, and of Euclidean R^p. Let f, σ be functions on $[0, T] \times R^n \times R^p$ with, respectively, values in R^n and $n \times m$ matrices as values. Let $(\cdot)_t$ denote partial derivative in the first variable t, and $(\cdot)_x$, $(\cdot)_y$ gradients in the second set of variables x and the third set y. We make the following assumptions.

(i) f, σ are of class $C^{(2)}$.

(ii) If $H \subset R^p$ is any compact set, then f, σ, f_x, σ_x are bounded on $[0, T] \times R^n \times H$.

(iii) Let $a = \frac{1}{2}\sigma\sigma^*$, where * denotes matrix transpose. Then there exists

$c > 0$ such that

$$\sum_{i,j=1}^{n} a_{ij}(t, x, y)\mu_i\mu_j \geqq c \mid \mu \mid^2$$

for every $\mu \in R^n$.

Condition (ii) will prove to be no real limitation, since we shall restrict (t, x) to a compact cylinder $\bar{Q}$ in §3. Condition (iii) is essential to treating the problem as one about parabolic equations, since it assures that the operators A^Y in §3 are uniformly parabolic. (Certain results have been found in [5] and [6] without condition (iii) by a combination of probabilistic and partial differential equations methods.)

Let $\Omega = (\Omega, \mathfrak{B}, P)$ be a probability space on which a random variable ξ_0 (with values in R^n) and an m-dimensional Brownian motion process $w = (w_1, \cdots, w_m)$ independent of ξ_0 are defined. Let Y be a function from $[0, T] \times R^n$ onto R^p which is bounded and Lipschitz. (For brevity we say that a function is Lipschitz if it satisfies a Lipschitz condition, and Hölder (or Hölder continuous) if it satisfies a Hölder condition.) Let

$$f^Y(t, x) = f(t, x, Y(t, x)), \qquad \sigma^Y(t, x) = \sigma(t, x, Y(t, x)).$$

These functions are also Lipschitz. Hence the system of stochastic differential equations

$$\xi(t) = \xi_0 + \int_0^t f^Y[r, \xi(r)]\, dr + \int_0^t \sigma^Y[r, \xi(r)]\, dw(r) \tag{2.1}$$

determines uniquely a diffusion process ξ on $[0, T]$ satisfying $E\{\mid \xi(t)\mid^2\} < \infty$ for each t [2, Chap. 6]. We say that Y is a *control based on observing current states*, and ξ, the *response* to the control Y given the initial data $\xi(0) = \xi_0$. Equation (2.1) is of course the integrated form of the system of stochastic differential equations

$$d\xi = f^Y\, dt + \sigma^Y\, dw \tag{2.1'}$$

with the initial data.

Using results about generalized solutions of linear parabolic equations with discontinuous coefficients, one can still define a response process ξ when Y is merely bounded and measurable, at least when $\sigma = \sigma(t, x)$ and (iii) holds. This is done in [10] and [19] under slightly more restrictive assumptions. While the existence of a response ξ is of interest in connection with §5, we shall not give details.

2.1. Controls based on partially observed current states. Let us now suppose that the controller can observe at time t not the state $\xi(t)$ of the response but only $\hat{\xi}(t)$, which is a vector consisting of certain components

of $\xi(t)$:

$$\hat{\xi}(t) = (\xi_1(t), \cdots, \xi_l(t)), \qquad 0 \leqq l \leqq n.$$

When $l = n$, the states are completely observable, and the problem is of the type we called Markovian in [4], [6]. We agree that $l = 0$ means that the controller has no information and chooses a nonrandom function of time. In many problems there are observations $\eta(t)$ governed by stochastic differential equations of the form

$$d\eta = \tilde{f}(t, \xi, \eta)\, dt + \tilde{\sigma}(t, \xi, \eta)\, d\tilde{w},$$

where $\tilde{w}$ is a Brownian motion independent of w. The pairs (η, ξ) form a (vector) diffusion process; and if we regard $(\eta(t), \xi(t))$ as the state of the system at time t, this is of the type being considered here.

Let us admit only controls of the form $Y(t, \hat{x})$, $\hat{x} = (x_1, \cdots, x_l)$, which are bounded and Lipschitz on $[0, T] \times R^l$. Let $\hat{\mathcal{Y}}$ denote the class of all such controls Y. If $K \subset R^p$, let $\hat{\mathcal{Y}}_K$ consist of those $Y \in \hat{\mathcal{Y}}$ which have values in K.

Let Σ be a closed subset of the strip $[0, T] \times R^n$ called the terminal set. We shall make a specific choice for Σ immediately below. Let τ denote the least time $t \in [0, T]$ such that $(t, \xi(t)) \in \Sigma$. We assume that the controller can observe the stopping time τ.

Let $K \subset R^p$ be closed and convex (K is the "control region"), and let $L(t, x, y)$ be a real-valued function of class $C^{(2)}$. We are interested in necessary conditions for

$$J(Y) = E \int_0^\tau L[t, \xi(t), Y(t, \hat{\xi}(t))]\, dt$$

to be minimum among controls $Y \in \hat{\mathcal{Y}}_K$.

3. The backward and forward boundary problems. Let us choose the terminal set Σ to lie on the boundary of a cylinder Q, just as in [4], [6]. Let $B \subset R^n$ be open with compact closure $\bar{B} = B \cup \partial B$. Moreover, let ∂B be locally representable by functions with Hölder continuous second order partial derivatives. Let

$$Q = (0, T) \times B,$$

$$\Sigma = [0, T] \times \partial B \cup \{T\} \times B,$$

$$\bar{Q} = Q \cup \partial Q = Q \cup \Sigma \cup \{0\} \times B.$$

Thus, the stopping time is the first positive t when $(t, \xi(t))$ reaches ∂Q, starting at time 0 in B. We have chosen such a terminal set since several of the results needed about parabolic equations are available in the literature

for bounded cylindrical domains. There would not appear to be any difficulty extending the results to a domain $Q \subset [0, T] \times R^n$ such that $\partial Q = B_0 \cup S \cup B_T$, where B_0, B_T are open subsets of the hyperplanes $t = 0$, $t = T$, respectively, and S (the lateral boundary) is a compact set locally representable by

$$x_j = h(t, x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n)$$

for suitable j as in [9, pp. 64–65]. In case Q is unbounded, one needs growth conditions on f, σ, L as $|x| \to \infty$; for instance, f, σ, L, f_x, σ_x, L_x bounded when y is restricted to any compact set H would suffice.

Let $\mathfrak{F}_0$ denote the set of all real-valued functions ϕ on the cylinder $\bar{Q}$ such that:

(i) ϕ and ϕ_x are Hölder continuous on $\bar{Q}$;

(ii) the partial derivatives ϕ_t, $\phi_{x_i x_j}$, $i, j = 1, \cdots, n$, are continuous on $\bar{Q} - \{T\} \times \partial B$ and square integrable on Q;

(iii) $\phi(t, x) = 0$ for all $(t, x) \in \Sigma$.

Given $Y \in \hat{\mathfrak{Y}}$, let (as in §2)

$$g^Y(t, x) = g(t, x, Y(t, x)),$$

where g is any function of (t, x, y). Let A^Y denote the backward operator of the response process ξ:

$$A^Y\phi = \phi_t + a^Y \cdot \phi_{xx} + f^Y \cdot \phi_x,$$

where

$$a^Y \cdot \phi_{xx} = \sum_{i,j=1}^{n} a_{ij}^Y \phi_{x_i x_j}.$$

Since Y is Lipschitz, the coefficients of A^Y are Lipschitz. The boundary problem

$$\begin{aligned} A^Y\psi + L^Y &= 0 \quad \text{in} \quad Q, \\ \psi &= 0 \quad \text{on} \quad \Sigma, \end{aligned} \tag{3.1}$$

has a unique solution $\psi \in \mathfrak{F}_0$. This is a slight modification of a standard existence theorem [9, p. 65] for smooth solutions to (3.1). In the Appendix we review this and other known results about parabolic equations (in the notation there, $a^Y = \alpha, f^Y = \beta, L^Y = \gamma$).

From the theory of Markov processes [3, especially Chap. 13],

$$E \int_0^\tau A^Y\psi(t, \xi(t))\, dt = E\psi\,|_0^\tau = -E\psi(0, \xi_0), \tag{3.2}$$

since $\psi(\tau, \xi(\tau)) = 0$. Let the distribution of the initial state ξ_0 be a prob-

ability measure π_0 on B. Then from (3.2),

$$J(Y) = \int_B \psi(0, x)\, d\pi_0(x). \tag{3.3}$$

Our minimum problem can now be restated: find $Y \in \hat{\mathcal{Y}}_K$ such that (3.3) is minimum, where ψ is the solution of the boundary problem (3.1). Necessary conditions for a minimum will be found using this boundary problem and a dual one for the forward operator. The method is nonprobabilistic; the reader who may wish to avoid probabilistic considerations altogether may take (3.1)–(3.3) as defining $J(Y)$.

In stating the dual problem let us for the moment assume that the initial distribution π_0 has a smooth density q_0 and that the control Y is smooth (say of class $C^{(3)}$). Then the adjoint is given by

$$(A^Y)^* q = -q_t + (a^Y q)_{xx} - (f^Y q)_x .$$

Let $q(t, x)$ solve the boundary problem

$$\begin{aligned} (A^Y)^* q &= 0 \quad \text{in} \quad Q, \\ q(t, x) &= 0 \quad \text{for} \quad x \in \partial B, \qquad 0 \leqq t \leqq T, \\ q(0, x) &= q_0(x) \quad \text{for} \quad x \in B. \end{aligned} \tag{3.4}$$

If $\phi \in \mathcal{F}_0$, then

$$\int_Q (A^Y \phi) q \, dt \, dx = -\int_B \phi(0, x) q_0(x) \, dx. \tag{3.4'}$$

The remaining terms in Green's identity [9, p. 27] disappear since $q = \phi = 0$ for $x \in \partial B$ and $\phi = 0$ when $t = T$.

Let us now merely suppose that π_0 is a probability measure on B and that $Y \in \hat{\mathcal{Y}}$. Then not all of the indicated derivatives in (3.4) need exist. However, (3.4) has a weak solution q in the following sense. Let q be a function integrable on Q. Then q is a weak solution of (3.4) if, for every $\phi \in \mathcal{F}_0$ such that $A^Y \phi$ is bounded,

$$\int_Q (A^Y \phi) q \, dt \, dx = -\int_B \phi(0, x)\, d\pi_0(x). \tag{3.4''}$$

The existence and uniqueness of the weak solution q are known. Moreover, q has the following properties (Appendix 2). Let $Q_\delta = (\delta, T) \times B$.

1. For any $\delta > 0$, q is Hölder continuous on $\bar{Q}_\delta$ and $|q_x|^2$ is integrable over Q_δ.
2. $q(t, x) = 0$ for $0 < t \leqq T$ and $x \in \partial B$.
3. $q(t, x) > 0$ for $0 < t \leqq T$ and $x \in B$.

In probabilistic terms, $q(t, \cdot)$ is the density of the random variable

$\xi'(t)$, where ξ' is the response process ξ stopped at Σ; namely,

$$\xi'(t) = \begin{cases} \xi(t) & \text{if } 0 \leqq t \leqq \tau, \\ x_\infty & \text{if } \tau < t \leqq T, \end{cases}$$

where x_∞ is an "ideal point" to which $\xi'(t)$ jumps when Σ is first reached.

The necessary conditions for a minimum will involve conditional expectations. Let

$$\hat{x} = (x_1, \cdots, x_l), \quad \hat{\hat{x}} = (x_{l+1}, \cdots, x_n),$$

$$\hat{q}(t, \hat{x}) = \int q(t, \hat{x}, \hat{\hat{x}})\, d\hat{\hat{x}}, \qquad 0 < t \leqq T, \tag{3.5}$$

where the integral is over R^{n-l} and we have set $q(t, x) = 0$ for $x \notin \bar{B}$ Let $\hat{Q}$ denote the projection onto $(t, \hat{x})$-space of the cylinder Q. The function $\hat{q}$ is continuous, and $\hat{q}(t, \hat{x}) > 0$ on $\hat{Q}$. Let G be any function continuous on $(0, T) \times \bar{B}$. For every $(t, \hat{x}) \in \hat{Q}$ let

$$EG(t, \cdot) \mid \hat{x} = \int G(t, \hat{x}, \hat{\hat{x}})\, \frac{q(t, \hat{x}, \hat{\hat{x}})}{\hat{q}(t, \hat{x})}\, d\hat{\hat{x}}. \tag{3.6}$$

The integral is over R^{n-l}; for $x \notin \bar{B}$, $G(t, x)$ is arbitrary since $q(t, x) = 0$ there. The function defined by (3.6) is continuous on $\hat{Q}$ and

$$|\, EG(t, \cdot) \mid \hat{x} \,| \leqq \max_{\bar{B}} |\, G(t, x) \,|.$$

If $\hat{\xi}'$ consists of the first l components of the stopped process ξ', then $\hat{q}(t, \cdot)$ is the density of $\hat{\xi}'(t)$ and $EG(t, \cdot) \mid \hat{x}$ the conditional expectation of $G(t, \xi'(t))$ given that $\hat{\xi}'(t) = \hat{x}$.

4. Necessary conditions for a minimum. Let us suppose that Y_0 makes $J(Y)$ a minimum on $\mathfrak{Y}_K$. Let ψ^0 be the solution of the boundary problem (3.1) when $Y = Y_0$; and let

$$\Phi(t, x, y) = a \cdot \psi^0_{xx} + f \cdot \psi_x{}^0 + L, \tag{4.1}$$

$$\hat{\Phi}(t, \hat{x}, y) = E\Phi(t, \cdot, y) \mid \hat{x}. \tag{4.2}$$

In this section we prove a necessary condition which corresponds to the vanishing of the first variation in classical calculus of variations. When $\sigma = \sigma(t, x)$, a stronger result corresponding to Pontryagin's maximum principle (§5, Theorem 2) is proved later.

As usual $(\cdot)_y$ denotes gradient in the variables $y = (y_1, \cdots, y_p)$. Let q^0 be the weak solution of (3.4″), and $\hat{q}^0$ given by (3.5), when $Y = Y_0$.

LEMMA. *If* $Y_0 + Z \in \mathfrak{Y}_K$, *then*

$$\int_{\hat{Q}} \hat{\Phi}_y(t, \hat{x}, Y_0) \cdot Z\, \hat{q}^0\, dt\, d\hat{x} \geqq 0.$$

Proof. Since K is convex the control $Y_0 + \epsilon Z$ is in $\hat{\mathfrak{Y}}_K$ for $0 \leqq \epsilon \leqq 1$. Write $A^\epsilon = A^{Y_0+\epsilon Z}$, etc., and ψ^ϵ for the corresponding solution in $\mathfrak{F}_0$ of (3.1). Then

$$A^0(\psi^\epsilon - \psi^0) + (A^\epsilon - A^0)\psi^\epsilon + L^\epsilon - L^0 = 0,$$

$$0 \leqq J(Y_0 + \epsilon Z) - J(Y_0) = \int_B [\psi^\epsilon(0, x) - \psi^0(0, x)]\, d\pi_0(x).$$

If we take $\phi = \psi^\epsilon - \psi^0$ and $A^Y = A^0$ in (3.4″), then

$$(4.3) \qquad 0 \leqq \int_Q [(a^\epsilon - a^0)\cdot\psi^\epsilon_{xx} + (f^\epsilon - f^0)\cdot\psi_x{}^\epsilon + L^\epsilon - L^0]q^0\, dt\, dx.$$

The functions ψ^ϵ_{xx}, $\psi_x{}^\epsilon$ have uniformly bounded square integrals over Q and, as $\epsilon \to 0$, tend to ψ^0_{xx}, $\psi_x{}^0$ uniformly on any compact subset of Q (even on compact subsets of $\bar{Q} - \{T\} \times \partial B$) (see Appendix 1). Moreover, if g denotes a, f, or L, then $\epsilon^{-1}(g^\epsilon - g^0)$ tends uniformly to $g_y \cdot Z$. Hence, if we divide by ϵ in (4.3) and let $\epsilon \to 0$, then

$$0 \leqq \int_Q \Phi_y(t, x, Y_0)\cdot Z\, q^0\, dt\, dx.$$

But since Y_0 and Z are functions of $(t, \hat{x})$, the right side equals the expression in the lemma by taking conditional expectations.

For each $y \in K$ let $K(y)$ denote the contingent to K at y. It is the convex cone of all z such that $y + \epsilon z \in K$ for $0 \leqq \epsilon \leqq h(z)$, $h(z) > 0$.

THEOREM 1. *If Y_0 minimizes* (3.3) *among all controls $Y \in \hat{\mathfrak{Y}}_K$, then for every $(t, \hat{x}) \in \hat{Q}$,*

$$\hat{\Phi}_y(t, \hat{x}, Y_0(t, \hat{x}))\cdot z \geqq 0 \quad \textit{for all} \quad z \in K(Y_0(t, \hat{x})).$$

Proof. Suppose not. Then for some $(t_1, \hat{x}_1) \in \hat{Q}$, $y_1 = Y_0(t_1, \hat{x}_1)$, and $z_1 \in K(y_1)$, we have $\hat{\Phi}_y(t, \hat{x}_1, y_1)\cdot z_1 < 0$. We may assume that $y_1 + z_1 \in K$. Now Φ_y is continuous on $[0, T) \times \bar{B} \times R^p$ and hence $\hat{\Phi}_y$ is continuous on $\hat{Q} \times R^p$. Therefore, there exist a neighborhood U of $(t_1, \hat{x}_1)$ and $\theta > 0$ such that $\hat{\Phi}_y(t, \hat{x}, y)\cdot z < 0$ for all $(t, \hat{x}) \in U$ and $|y - y_1| < \theta$, $|z - z_1| < \theta$.

For each $\eta \in R^p$, let $P(\eta)$ be the point of K nearest η. Then

$$|P(\eta_2) - P(\eta_1)| \leqq |\eta_2 - \eta_1|, \qquad P(\eta) = \eta \quad \text{if} \quad \eta \in K.$$

If $z = P(y + z_1) - y$, then $y + z \in K$; and since $P(y_1 + z_1) = y_1 + z_1$, $z - z_1 = P(y + z_1) - P(y_1 + z_1) + y_1 - y$, $|z - z_1| \leqq 2|y - y_1|$. We now take U small enough that

$$|Y_0(t, \hat{x}) - Y_0(t_1, \hat{x}_1)| < \theta/2$$

for all $(t, \hat{x}) \in U$, and take $y = Y_0(t, \hat{x})$, z as above. Let g be Lipschitz

with $0 \leqq g(t, \hat{x}) \leqq 1$, $g \not\equiv 0$, and $g(t, \hat{x}) = 0$ outside U. If $Z = gz$, then

$$\int_{\hat{Q}} \hat{\Phi}_y(t, \hat{x}, Y_0) \cdot Zq^0 \, dt \, d\hat{x} < 0,$$

contradicting the lemma.

COROLLARY. *If $K = R^p$ (no control constraints), then $E\Phi_y(t, \cdot, Y_0(t, \hat{x})) \mid \hat{x} = 0$ for every $(t, \hat{x}) \in \hat{Q}$.*

Proof. The contingent to R^p at any point is R^p.

5. Measurable controls. Let us suppose now that $\sigma = \sigma(t, x)$ and admit controls which are merely bounded, measurable functions of $(t, \hat{x})$. If Y is such a control, then

$$A^Y\phi = \phi_t + a \cdot \phi_{xx} + f^Y \cdot \phi_x,$$

where $a(t, x)$ is a $C^{(2)}$ function (the same for all choices of Y) and f^Y is bounded and measurable. Let $\mathfrak{F}_1$ denote the set of all ϕ which satisfy (i), (iii) in the definition (see §3) of $\mathfrak{F}_0$, and in addition:

(ii′) the partial derivatives ϕ_t, $\phi_{x_i x_j}$, $i, j = 1, \cdots, n$, are square integrable on Q.

The boundary problem (3.1) now has a unique solution $\psi \in \mathfrak{F}_1$ (Appendix 1). We take (3.3) as the definition of $J(Y)$. The forward boundary problem (3.4″) has a weak solution q with the same properties 1–3 as in §3. Moreover, (3.4″) remains correct for any $\phi \in \mathfrak{F}_1$ such that $A^Y\phi$ is bounded (Appendix 2). The conditional expectation $EG(t, \cdot) \mid \hat{x}$ is defined exactly as before. Let us suppose that Y_0 is minimizing; let ψ^0 be the corresponding solution in $\mathfrak{F}_1$ of (3.1), and

$$\Psi(t, x, y) = f \cdot \psi_x^0 + L. \tag{5.1}$$

Since ψ_x^0 is continuous on $\bar{Q}$, Ψ is continuous on $\bar{Q} \times R^p$.

THEOREM 2. *Let Y_0 minimize (3.3) among all bounded, measurable controls $Y(t, \hat{x})$ with values in K. Then for almost every $(t, \hat{x}) \in \hat{Q}$, $E\Psi(t, \cdot, y) \mid \hat{x}$ is minimum on K when $y = Y_0(t, \hat{x})$.*

Proof. Let $\{y_1, y_2, \cdots\}$ be a countable dense subset of K, and

$$\hat{Q}_k = \{(t, \hat{x}) \in \hat{Q} : \hat{\Psi}(t, \hat{x}, y_k) < \hat{\Psi}(t, \hat{x}, Y_0(t, \hat{x}))\},$$

where

$$\hat{\Psi}(t, \hat{x}, y) = E\Psi(t, \cdot, y) \mid \hat{x}.$$

It suffices to show that $\hat{Q}_k$ has measure 0 for each $k = 1, 2, \cdots$. Suppose that this is false for some k. Consider the control Y given by

$$Y(t, \hat{x}) = \begin{cases} y_k & \text{for } (t, \hat{x}) \in \hat{D}, \\ Y_0(t, \hat{x}) & \text{otherwise,} \end{cases}$$

where $\hat{D} \subset \hat{Q}_k$ has the following properties:

(i) there is a cylinder $\hat{V} = [t_1, t_2] \times \hat{\Gamma}$ such that $\hat{D} \subset \hat{V} \subset \hat{Q}$ and $|\hat{V}| \leqq 2|\hat{D}|$, where $|\cdot|$ denotes Lebesgue measure;

(ii) for $(t, \hat{x}) \in \hat{D}$, $\hat{\Psi}(t, \hat{x}, y_k) \leqq \hat{\Psi}(t, \hat{x}, Y_0) - b$, where $b > 0$;

(iii) $|\hat{\Gamma}|$ is small enough (see (5.8) below).

Let $\psi \in \mathfrak{F}_1$ be the solution of (3.1) corresponding to Y, and $\phi = \psi - \psi^0$. Then

$$0 \leqq J(Y) - J(Y_0) = -\int_Q (A^0\phi)q^0 \, dt \, dx, \tag{5.2}$$

$$-A^0\phi = (f^Y - f^0)\cdot\psi_x + L^Y - L^0, \tag{5.3}$$

where $A^0 = A^{Y_0}$, etc. Let

$$D = \{(t, x) \in Q : (t, \hat{x}) \in \hat{D}\}, \qquad \Gamma = \{x \in B : \hat{x} \in \hat{\Gamma}\}.$$

Their measures are bounded by a constant times the measures of $\hat{D}$, $\hat{\Gamma}$, respectively, since Q is a bounded set. The right side of (5.3) is bounded and equal to 0 except on D. Hence,

$$\int_Q |A^0\phi|^2 \, dt \, dx \leqq c_1 |D| \tag{5.4}$$

for suitable c_1. By Sobolev's lemma,

$$\left(\int_B |\phi_x|^r \, dx\right)^{1/r} \leqq c_2 \left(\int_B |\phi_{xx}|^2 \, dx\right)^{1/2}, \quad r = \frac{2n}{n-2}, \tag{5.5}$$

while by (5.4) and (A2) in Appendix 1,

$$\int_Q |\phi_{xx}|^2 \, dt \, dx \leqq c_3 |D|. \tag{5.6}$$

Since $f^Y - f^0$ is bounded and q^0 is bounded for $t \geqq t_1 > 0$, we have, using Hölder's inequality,

$$\left|\int_Q (f^Y - f^0)\cdot\phi_x q^0 \, dt \, dx\right| \leqq c_4 \int_D |\phi_x| \, dt \, dx$$

$$\leqq c_4 |\Gamma|^{1/r'} \int_{t_1}^{t_2} \left(\int_\Gamma |\phi_x|^r \, dx\right)^{1/r} dt,$$

where $(r')^{-1} + r^{-1} = 1$ and thus $1 < r' < 2$. From (5.5), (5.6) and the Cauchy-Schwarz inequality,

$$\left|\int_Q (f^Y - f^0)\cdot\phi_x q^0 \, dt \, dx\right| \leqq c_5 |\Gamma|^\alpha |V|^{1/2} |D|^{1/2} \leqq c_6 |\hat{\Gamma}|^\alpha |\hat{D}|, \tag{5.7}$$

where $\alpha = (r')^{-1} - 2^{-1} > 0$. Now $0 < b_1 \leqq \hat{q}^0(s, \hat{x})$ on $\hat{V}$, for some b_1,

Let us choose $\hat{\Gamma}$ small enough that

$$c_6 \mid \hat{\Gamma} \mid^{\alpha} < b_1 b. \tag{5.8}$$

From (5.3),

$$-A^0\phi = \Psi(t, x, Y) - \Psi(t, x, Y_0) + (f^Y - f^0)\cdot\phi_x .$$

Since

$$\int_Q \Psi \mid_{Y_0}^{Y} q^0 \, dt \, dx = \int_{\hat{Q}} \hat{\Psi} \mid_{Y_0}^{Y} \hat{q}^0 \, dt \, d\hat{x} \leqq -b_1 b \mid \hat{D} \mid,$$

we find by (5.7) and (5.8) that

$$-\int_Q (A^0\phi) q^0 \, dt \, dx < 0,$$

contrary to (5.2). This proves Theorem 2.

Remarks. If Y_0 is continuous, then the conclusion of Theorem 2 holds for all $(t, \hat{x}) \in \hat{Q}$. In that case the sets $\hat{Q}_k$ are open, of Lebesgue measure 0, hence empty.

The necessary conditions in Theorem 2 involve the solutions ψ^0, q^0 of the boundary problems (3.1), (3.4″) together with a minimum condition of Pontryagin type. These conditions resemble those of Pontryagin's maximum principle in ordinary control theory; however, we must now solve backward and forward boundary problems for linear parabolic equations instead of initial and final value problems for ordinary differential equations.

If $\xi(t)$ can be completely observed, then the minimum condition in Theorem 2 becomes $f\cdot\psi_x^0 + L$ = minimum on K when $y = Y_0(t, x)$. This is equivalent to saying that ψ^0 also solves the nonlinear boundary problem

$$\begin{aligned} \psi_t^0 + a\cdot\psi_{xx}^0 + \min_{y \in K} [f\cdot\psi_x^0 + L] &= 0 \quad \text{in} \quad Q, \\ \psi^0 &= 0 \quad \text{on} \quad \Sigma. \end{aligned} \tag{5.9}$$

This equation can be obtained formally from Bellman's principle of optimality in dynamic programming. Since the equation is uniformly parabolic, the solutions of (5.9) are smooth, and the dynamic programming formalism can be made precise.

The problem with complete observations was studied from the point of view of (5.9) in [4], [6]. However, the dynamic programming approach seems less promising when $\xi(t)$ is only partially observable. The principle of optimality then leads (formally) not to a partial differential equation, but to some more complicated equation for a function of t, $\hat{x}$, π, where π is

a probability measure on B representing the conditional distribution of $\xi'(t)$, given that $\hat{x}$ is observed at time t.

Under some reasonable additional conditions $J(Y)$ has a minimum among measurable controls.

THEOREM 3. *Let L be convex in y and f linear in y. Moreover, let the control region K be compact and convex. Then* (3.3) *has a minimum among all measurable controls Y with values in K.*

Proof. Let $Y_1, Y_2, \cdots$ be a minimizing sequence. By taking subsequences we may assume that Y_k tends weakly in $L^\infty(\hat{Q})$ to a limit Y_0, which has values in K since K is compact and convex. As usual $L^\infty(\hat{Q})$ is the space of bounded measurable functions on $\hat{Q}$. Let

$$L_k(t, x) = L(t, x, Y_k(t, x)), \quad k = 0, 1, 2, \cdots$$

and let $h(t, x)$ be any nonnegative continuous function. By a standard semicontinuity theorem (Appendix 3, with $F = hL$),

$$\int_Q hL_0 \, dt \, dx \leqq \liminf_{k\to\infty} \int_Q hL_k \, dt \, dx. \tag{5.10}$$

By taking a further subsequence we may assume that L_k tends weakly in $L^\infty(Q)$ to a limit L^*. The right side of (5.10) equals $\int_Q hL^* \, dt \, dx$. Since this is true for each h, we have $L_0 \leqq L^*$ almost everywhere in Q. Let $A^k = A^{Y_k}$, and ψ^k the corresponding solution in $\mathfrak{F}_1$ of (3.1) for $Y = Y_k$. Let $\psi^* \in \mathfrak{F}_1$ satisfy

$$A^0\psi^* + L^* = 0.$$

By the maximum principle for parabolic equations, $\psi^0 \leqq \psi^*$ since $L_0 \leqq L^*$ almost everywhere. The second order coefficients of A^k and A^0 are the same, and the first order coefficients f^k tend to f^0 weakly since f is linear in y. Moreover, L_k tends to L^* weakly. This implies (Appendix 1) that $\psi^k(0, x)$ tends to $\psi^*(0, x)$ uniformly on $\bar{B}$, from which

$$\int_B \psi^0(0, x) \, d\pi_0(x) \leqq \int_B \psi^*(0, x) \, d\pi_0(x) = \lim_{k\to\infty} \int_B \psi^k(0, x) \, d\pi_0(x).$$

This shows that Y_0 is a minimizing control.

It is an interesting open problem to determine what further continuity or smoothness properties a minimizing control Y_0 must have. Simple examples show that Y_0 may in general have discontinuities. However, under a stronger convexity condition on L we can prove that Y_0 is continuous on $\hat{Q}$. In fact, we can show that the following theorem holds.

THEOREM 4. *Let f be linear in y, and let*

$$\sum_{i,j=1}^{p} L_{y_i y_j} \mu_i \mu_j \geqq b \mid \mu \mid^2$$

for all $\mu \in R^p$, *where* $b > 0$. *Then* Y_0 *is Hölder continuous on any compact subset of* $\hat{Q}$.

Proof. Let us define $\hat{\Psi}$ as in the proof of Theorem 2. By differentiating under the integral sign,

$$\hat{\Psi}_y = \frac{\int \Psi_y \, q^0 \, d\hat{x}}{\hat{q}^0} .$$

Let $\hat{E} \subset \hat{Q}$ be compact. For any $\delta > 0$, $\Psi_y q^0$ is Hölder on $[\delta, T] \times \bar{B} \times K$, and $\hat{q}^0$ is Hölder and positive on $\hat{E}$ (we set $\Psi_y q^0 = 0$ for $x \notin B$). Hence $\hat{\Psi}_y$ is Hölder on $\hat{E} \times K$. Further differentiation under the integral sign gives

$$\sum_{i,j=1}^{p} \hat{\Psi}_{y_i y_j} \mu_i \mu_j \geqq b \mid \mu \mid^2$$

for all μ. In particular, $\hat{\Psi}$ is strictly convex in y, and $Y_0(t, \hat{x})$ is the unique $y \in K$ at which $\hat{\Psi}(t, \hat{x}, y)$ is minimum. The proof of [6, Lemma 2.1] shows that Y_0 is Hölder on $\hat{E}$.

For completely observable diffusions we proved a slightly stronger result about Y_0 [6, Theorem 2.2] by the same method.

6. Lagrange multipliers. If in §3 and §4 we let

$$\lambda(t) = \psi_x(t, \xi(t)), \qquad \mu(t) = \psi_{xx}(t, \xi(t)), \tag{6.1}$$

then $\Phi = \mu \cdot a + \lambda \cdot f + L$. The processes λ, μ have the role of Lagrange multipliers in Theorems 1 and 2. In case $a = a(s, x)$ the multipliers μ are irrelevant and in §5 we could replace Φ by Ψ.

A quite different way to define multipliers, which resembles the usual method for the Lagrange problem in calculus of variations and control theory, is the following. Let us take $\sigma = \text{const.}$, and a fixed stopping time ($\tau \equiv T$). Let us also make the (rather strong) assumption that Y is of class $C^{(2)}$ in $\hat{x}$, and define a process Λ by the system of ordinary differential equations:

$$\frac{d\Lambda}{dt} = -\Lambda(f_x + f_y \, Y_{\hat{x}}) - (L_x + L_y \, Y_{\hat{x}}), \tag{6.2}$$

$$0 \leqq t \leqq T, \quad \Lambda(T) = 0.$$

Here $Y_{\hat{x}}$ is the gradient in the variables $x_1, \cdots, x_l$.

The multipliers λ in (6.1) can be obtained from these by taking conditional expectations with respect to the complete past of w and ξ_0. More pre-

cisely, let $\mathcal{B}_s$ denote the σ-algebra (of subsets of our probability space Ω) generated by the random variables $w(r)$ for $0 \leqq r \leqq s$ and ξ_0. Then

$$\lambda(s) = E\Lambda(s) \mid \mathcal{B}_s, \qquad 0 \leqq s \leqq T, \tag{6.3}$$

with probability 1 for each s. This can be proved by substantially the same reasoning as in [12, §4]. Since $\tau \equiv T$, ψ is now a solution of (3.1) in the strip $[0, T] \times R^n$ with the Cauchy data $\psi = 0$ on the hyperplane $\{T\} \times R^n$.

Unfortunately the multipliers Λ involve the partial derivatives of Y, which in many examples the optimal control does not possess everywhere. In certain instances one can replace Λ by $\tilde{\Lambda}$, defined by

$$\frac{d\tilde{\Lambda}}{dt} = -(\tilde{\Lambda} f_x + L_x), \qquad 0 \leqq t \leqq T, \qquad \tilde{\Lambda}(T) = 0, \tag{6.4}$$

which are the usual equations in calculus of variations for the multipliers. Obviously $\Lambda = \tilde{\Lambda}$ if there are no observations ($l = 0$, $Y = Y(t)$). Moreover, $\tilde{\Lambda}$ will serve as a set of Lagrange multipliers whenever we can put $\tilde{\Lambda}$ instead of Λ in (6.3). Let us show that this is so in the case when information is complete ($\hat{x} = x$), there are no control constraints, and Y is optimal.

Theorem 1 then states that

$$\lambda \cdot f_y + L_y = 0. \tag{6.5}$$

For $s < t$ let $M(s, t)$ be the fundamental matrix of the linear system $d\eta/dt = f_x \eta$ (actually $M = M(s, t, \omega)$, $\omega \in \Omega$). By subtracting (6.2) from (6.4) we obtain

$$\tilde{\Lambda}(s) - \Lambda(s) = \int_s^T (\Lambda f_y + L_y) Y_x M(s, t)\, dt.$$

All terms under the integral sign are $\mathcal{B}_t$-measurable except $\Lambda(t)$. Hence, upon taking conditional expectations we obtain

$$E(\tilde{\Lambda}(s) - \Lambda(s)) \mid \mathcal{B}_s = E\left\{\int_s^T [(E\Lambda \mid \mathcal{B}_t) f_y + L_y] Y_x M\, dt\right\} \mid \mathcal{B}_s.$$

The right side is 0 by (6.3) and (6.5). Thus with probability 1,

$$E\tilde{\Lambda}(s) \mid \mathcal{B}_s = E\Lambda(s) \mid \mathcal{B}_s = \lambda(s)$$

in this case.

Kushner [12] took as controls and variations processes u, v depending explicitly on the Brownian past, rather than functions of the states as we have done. In his variational formula the multipliers $\tilde{\Lambda}$ rather than Λ appear. In [7] we derived a generalization of Kushner's formula when $\sigma = \sigma(t, x)$, again with fixed stopping time. In doing so, (6.2) and (6.4) must be replaced by certain stochastic differential equations which in-

involve $\int \cdots dw$ after time is reversed. While the partial differential equation method above works for fixed or variable stopping time τ, the author knows how to derive a variational formula in terms of the multipliers Λ (or $\tilde{\Lambda}$) only when τ is fixed.

7. Observations at a finite set of times. We have assumed that the controller knows $\hat{\xi}(t)$ at time t, but does not remember past observations $\hat{\xi}(r)$ for $r < t$. A more difficult problem is the one in which controls may be based on past observations $\hat{\xi}(r)$ for $0 \leqq r \leqq t$. If the controller could solve that minimum problem, then he would generally do better than by using an optimal control based on current observations. However, a solution to the problem we have considered gives at least an upper bound for the minimum attainable in the more difficult one.

Another problem of practical interest to which our method applies in principle (although the calculations involved are quite difficult) is the following. Suppose that $\hat{\xi}$ is observed only at fixed times $t_0, t_1, \cdots, t_N$ and that at any time t the control is based on the observations $\hat{\xi}(t_j)$ for $t_j \leqq t$. For $j = 1, \cdots, N$ let $t_j = jT/N$, $t_0 = 0$,

$$Q_j = (t_{j-1}, t_j) \times B,$$

$$S_j = [t_{j-1}, t_j] \times \partial B.$$

For brevity let us write

$$x^j = (x_0, \cdots, x_{j-1}), \qquad \hat{x}^j = (\hat{x}_0, \cdots, \hat{x}_{j-1}),$$

where $x_i \in R^n$, $\hat{x}_i \in R^l$. Let $Y_j(t, x^j)$ be Lipschitz in all variables appearing. For $t_{j-1} \leqq t < t_j$ we apply at time t the control

$$u(t) = Y_j(t, \hat{\xi}(t_0), \cdots, \hat{\xi}(t_{j-1})).$$

The problem is to choose $Y = (Y_1, \cdots, Y_N)$ such that the expected value

$$J(Y) = E \int_0^\tau L[t, \xi(t), u(t)]\, dt$$

is minimum, where the response ξ is determined from the stochastic differential equations

$$d\xi = f(t, \xi, u)\, dt + \sigma(t, \xi, u)\, dw$$

with $\xi(0) = \xi_0$.

We define functions $\psi_1, \cdots, \psi_N$ working backward recursively from $j = N$ as follows. The function $\psi_j(t, x^j, x)$ is defined for $(t, x) \in \bar{Q}_j$, $x_i \in \bar{B}$ for $i \leqq j - 1$:

$$
\begin{aligned}
\psi_N(T, x^N, x) &= 0, \\
A^j\psi_j + L^j &= 0 \quad \text{in} \quad Q_j, \\
\psi_j &= 0 \quad \text{on} \quad S_j, \\
\psi_j(t_j, x^j, x) &= \psi_{j+1}(t_j, x^j, x, x),
\end{aligned} \tag{7.1}
$$

where $A^j = A^{Y_j}$ is applied in the variables (t, x). In probabilistic terms,

$$
\psi_j(s, x^j, x) = E\int_s^\tau L\, dt \,|\, \xi'(t_i) = x_i \quad \text{for} \quad i < j, \quad \xi'(s) = x,
$$

where ξ' is the stopped response process. Therefore, if $\xi(0)$ has distribution π_0,

$$
J(Y) = \int_B \psi_1(0, x, x)\, d\pi_0(x).
$$

By the methods we have used one can obtain necessary conditions for a minimum corresponding to Theorems 1 and 2. For instance, suppose that $\sigma = \sigma(t, x)$ and that $Y_0 = (Y_{01}, \cdots, Y_{0N})$ minimizes. For $t_{j-1} \leqq t < t_j$ let

$$
\begin{aligned}
\Psi_j(t, x^j, x, y) &= f\cdot\psi_{jx}^0 + L, \\
\hat{\Psi}_j(t, \hat{x}^j, y) &= E\Psi(t, \cdot, \cdot, y) \,|\, \hat{x}^j,
\end{aligned} \tag{7.2}
$$

where ψ_j^0 is the corresponding solution of (7.1). Then $\hat{\Psi}_j(t, \hat{x}^j, y)$ is minimum when $y = Y_{0j}(t, \hat{x}^j)$. In the proof, which we shall not give, one needs to verify that each ψ_j in (7.1) is Lipschitz in all variables from the fact that Y_0 has this property.

It would be interesting to be able to answer the following general question, of which the preceding are particular cases. Consider any control problem defined by stochastic differential equations and a loss criterion, of the kind in §2, with controls u based on certain observations of past responses. If u_0 is a minimizing control process, do there exist multiplier processes λ, μ such that

$$
E\{(\mu\cdot a + \lambda\cdot f + L) \,|\, \text{observations up to } t\}
$$

is minimum when $y = u_0(t)$?

Appendix 1. We shall summarize known results about the boundary problem

$$
\begin{aligned}
\phi_t + \alpha(t, x)\cdot\phi_{xx} + \beta(t, x)\cdot\phi_x + \gamma(t, x) &= 0 \quad \text{in} \quad Q, \\
\phi &= 0 \quad \text{on} \quad \Sigma,
\end{aligned} \tag{A1}
$$

where the cylinder Q and Σ are as in §3. Let us assume that (A1) is uniformly parabolic and that the entries α_{ij} of the symmetric matrix α are Lipschitz. For the moment also assume that β, γ are Hölder and that $\gamma(T, x) = 0$ for $x \in \partial B$. Then the boundary problem (A1) has a solution ϕ in Q such that ϕ, ϕ_t, ϕ_x, $\phi_{x_i x_j}$ are Hölder in $\bar{Q}$ [9, p. 65]. We need the following a priori estimates. Suppose that

$$|\alpha(t, x)| + \frac{|\alpha(t', x') - \alpha(t, x)|}{|t' - t| + |x' - x|} \leqq M_1, \quad |\beta(t, x)| \leqq M_2$$

for all (t, x), $(t', x') \in \bar{Q}$, and that $c > 0$ is a lower bound for the characteristic values of the matrices $\alpha(t, x)$. Then

$$\int_Q [\phi_t^2 + |\phi_{xx}|^2 + |\phi_x|^2]\, dt\, dx + \int_B [\phi(t, x)]^2\, dx$$

$$\text{(A2)} \qquad \leqq C_1 \int_Q \gamma^2\, dt\, dx,$$

$$|\phi_{xx}|^2 = \sum_{i,j=1}^{n} \phi_{x_i x_j}^2.$$

For $0 < \delta < 1$ and $\zeta = \phi$ or ϕ_x,

$$\text{(A3)} \qquad |\zeta(t, x)| + \frac{|\zeta(t', x') - \zeta(t, x)|}{[|t' - t| + |x' - x|^2]^{\delta/2}} \leqq C_2 \|\gamma\|_0,$$

$$\|\gamma\|_0 = \max_{\bar{Q}} |\gamma(t, x)|.$$

The numbers C_1, C_2 depend only on Q, c, M_1, M_2, and (in the case of C_2) δ. See [16, §2, Theorem 8], [9, p. 191]. Estimates similar to (A3) are proved by different methods in [13, I §5, II §1].

In §3 and §5 we used results about (A1) under weaker assumptions on β and γ. If β, γ are merely bounded and measurable, take sequences β^k, γ^k tending almost everywhere to β, γ, respectively, as $k \to \infty$, such that $|\beta^k| \leqq M_2$, $|\gamma^k| \leqq M_2$ for some M_2 and for each k, β^k, γ^k are Hölder with $\gamma^k(T, x) = 0$ for $x \in \partial B$. The estimates (A2), (A3) hold with ϕ replaced by the corresponding solution ϕ^k of (A1). From this one can conclude that (A1) has a solution $\phi \in \mathfrak{F}_1$ (see §5 for notation). This solution ϕ is unique; see for example [16, §5, Theorem 3]. Moreover, ϕ^k and ϕ_x^k tend uniformly on $\bar{Q}$ to ϕ, ϕ_x, while ϕ_t^k, ϕ_{xx}^k tend weakly in $L^2(Q)$ to ϕ_t, ϕ_{xx} as $k \to \infty$. If in (A1) we put $\tilde{\gamma}$ instead of γ and let $\tilde{\phi}$ denote the corresponding solution in $\mathfrak{F}_1$, then $\tilde{\gamma} \geqq \gamma$ implies $\tilde{\phi} \geqq \phi$ (see for example, [4, p. 135]).

In a similar way, one can deduce from (A2) and (A3) the following. Let β^k, γ^k be any uniformly bounded sequence tending weakly in $L^\infty(Q)$ to β, γ. Then the corresponding ϕ^k tends to ϕ in the sense just described.

If β, γ are Hölder, then one can say more. Let $E \subset Q - \{T\} \times \partial B$ be compact, and let h be a $C^{(3)}$ function with value 1 on E and 0 on $\{T\} \times \partial B$. Then

$$(*) \qquad (h\phi)_t + \alpha \cdot (h\phi)_{xx} + \beta \cdot (h\phi)_x = -h\gamma + (h_t + \alpha \cdot h_{xx} + \beta \cdot h_x)\phi + 2\alpha \cdot h_x \phi_x .$$

Let $D\psi$ denote ψ or any of its derivatives ψ_t, ψ_{x_i}, $\psi_{x_i x_j}$. Since the right side of $(*)$ is Hölder on $\bar{Q}$ and 0 on $\{T\} \times \partial B$, $D(h\phi)$ is Hölder on $\bar{Q}$. In particular, ϕ_t, $\phi_{x_i x_j}$ are continuous on $\bar{Q} - \{T\} \times \partial B$.

Similarly, let α^k, β^k, γ^k be any sequences tending uniformly to α, β, γ such that α^k satisfies a uniform Lipschitz condition and β^k, γ^k, a uniform Hölder condition on $\bar{Q}$. Let ϕ^k be the corresponding solution of (A1). If h and E are above, then by [9, Theorem 6, p. 65] there are uniform bounds and Hölder estimates for $D(h\phi^k)$. Therefore, $D\phi^k$ tends to $D\phi$ uniformly on compact subsets of $\bar{Q} - \{T\} \times \partial B$.

Appendix 2. Here we discuss weak solutions of the adjoint equation. Let us write (A1) as $A\phi + \gamma = 0$, and assume that the coefficients α_{ij} of A are Lipschitz, the β_i bounded, measurable. These assumptions are satisfied in §3 and §4 if we take $\alpha = a^Y$, $\beta = f^Y$ (in that case β is Lipschitz) and in §5 if we take $\alpha = a$, $\beta = f^Y$.

Let $\mathfrak{L}$ denote the linear functional on $L^\infty(Q)$ such that, for all $\gamma \in L^\infty(Q)$,

$$\mathfrak{L}\gamma = \int_B \phi(0, x)\, d\pi_0(x),$$

where $A\phi + \gamma = 0$, $\phi \in \mathfrak{F}_1$. By (A3), $\mathfrak{L}$ is bounded. We need to show that

$$(A4) \qquad \mathfrak{L}\gamma = \int_Q \gamma q \, dt \, dx = -\int_Q (A\phi) q \, dt \, dx,$$

where $q \in L^1(Q)$ and q has properties 1—3 listed in §3. Clearly q is uniquely determined by (A4). The existence of q follows from recent results about weak solutions of parabolic equations in the "divergence form"

$$(A5) \qquad -\frac{\partial q}{\partial t} + \sum_i \frac{\partial}{\partial x_i}\left(\sum_j \alpha_{ij} \frac{\partial q}{\partial x_j} + \alpha_i q\right) = 0 \quad \text{in} \quad Q$$

with $q(t, x) = 0$ for $x \in \partial B$, $0 < t \leqq T$. In our case,

$$\alpha_i = \sum_j \frac{\partial \alpha_{ij}}{\partial x_j} - \beta_i$$

is bounded, measurable. For $k = 1, 2, \cdots$, let α^k, β^k be smooth functions on $[0, T] \times R^n$ which are uniformly bounded, such that the α^k satisfy a

uniform Lipschitz condition and α^k, β^k, $\partial\alpha_{ij}^k/\partial x_j$ tend, respectively, to α, β, $\partial\alpha_{ij}/\partial x_j$ almost everywhere in Q as $k \to \infty$. Let G^k be the Green's function for Q [9, p. 82] and the operator A^k (with α, β in (A1) replaced by α^k, β^k); and let

$$q^k(t, x) = \int_B G^k(0, x_0; t, x)\, d\pi(x_0), \qquad 0 < t \leqq T. \tag{A6}$$

Then q^k satisfies the equation corresponding to (A5) with α_{ij} replaced by α_{ij}^k and α_i by

$$\alpha_i{}^k = \sum_j \frac{\partial\alpha_{ij}^k}{\partial x_j} - \beta_i{}^k .$$

Moreover, $q^k \geqq 0$ with equality for $x \in \partial B$. Now A^k has a (unique) fundamental solution F^k satisfying

$$0 \leqq F^k(s, x_0; t, x) \leqq \frac{C}{(t-s)^{n/2}} \exp\left(-\frac{b|x - x_0|^2}{t-s}\right), \quad s < t, \tag{A7}$$

for suitable positive constants C, b. From the construction of F^k [9, pp. 14–16] or [16, §4.1] one sees that these constants depend on uniform bounds and Hölder estimates for α_{ij}^k, but only on a uniform bound for $\beta_i{}^k$. Since $0 \leqq G^k \leqq F^k$, the integral of q^k over Q is uniformly bounded, and q^k is uniformly bounded on Q_δ for any $\delta > 0$. By [13, I §6] there are a uniform Hölder estimate for q^k on $\bar{Q}_\delta$ and a uniform bound for $\int_{Q_\delta} |q_x^k|^2\, dt\, dx$. For a subsequence q^k tends uniformly on each $\bar{Q}_\delta$ to a limit q, which is integrable on Q and has properties 1 and 2 in §3. Let γ be any smooth function with compact support in Q, and ϕ^k the solution of

$$\begin{aligned} A^k\phi^k + \gamma &= 0 \quad \text{in} \quad Q, \\ \phi^k &= 0 \quad \text{on} \quad \Sigma. \end{aligned}$$

For each $k = 1, 2, \cdots$,

$$\int_Q \gamma q^k\, dt\, dx = \int_B \phi^k(0, x)\, d\pi_0(x).$$

Since $\phi^k(0, x)$ tends uniformly on $\bar{B}$ to $\phi(0, x)$, the right side tends to $\mathfrak{L}\gamma$. On the other hand,

$$\lim_{k\to\infty} \int_Q \gamma q^k\, dt\, dx = \int_Q \gamma q\, dt\, dx.$$

Thus (A4) holds if γ is smooth and has compact support in Q. Any bounded measurable γ is the limit almost everywhere in Q of a uniformly bounded

sequence of such functions γ^k. From (A3) $\mathfrak{L}\gamma^k$ tends to $\mathfrak{L}\gamma$, from which (A4) holds for all $\gamma \in L^\infty(Q)$ by Lebesgue's convergence theorem.

Similar reasoning shows that

$$\int_{t_1}^{t_2}\int_B (A\phi)q \, dt \, dx = \int_B \phi q \,|_{t_1}^{t_2} \, dx, \qquad 0 < t_1 < t_2 \leqq T.$$

Therefore q is a weak solution of (A5) in the sense of [1] or [13]. By a Harnack inequality for nonnegative weak solutions [1, Theorem 3], $q(t, x) > 0$ for all $(t, x) \in Q$ (§3, property 3).

Appendix 3. For completeness we include a proof of the following known result. We consider sets $V \subset R^{n+1}$, $K \subset R^p$ such that V is open with compact closure $\bar{V}$ and K is compact, convex. In the proof of Theorem 3 in §5, we take $V = Q$.

THEOREM. *Let F be continuous on $\bar{V} \times K$ and convex on K for each $v \in V$. If $Y_1, Y_2, \cdots$ have values in K and tend weakly in $L^\infty(V)$ to Y_0, then*

$$\liminf_{k\to\infty} \int_V F(v, Y_k(v)) \, dv \geqq \int_V F(v, Y_0(v)) \, dv.$$

Proof. We may assume that $F(v, Y_k(v))$ tends weakly in $L^\infty(V)$ to a limit $F^*(v)$. Let v_0 be any point at which both Y_0 and F^* are approximately continuous; almost all points of V have this property [17, p. 132]. Since F is convex in y, there is a linear function G such that $G(y) \leqq F(v_0, y)$ for all $y \in K$ with equality when $y = Y_0(v_0)$. If Δ denotes an $(n+1)$-dimensional cube such that $v_0 \in \Delta \subset V$, and $|\Delta|$, its Lebesgue measure, then

$$F(v, Y_0(v_0)) = \lim_{|\Delta|\to 0} \frac{1}{|\Delta|} \int_\Delta F(v, Y_0(v)) \, dv = \lim_{|\Delta|\to 0} \frac{1}{|\Delta|} \int_\Delta G(Y_0(v)) \, dv.$$

Since F is uniformly continuous,

$$G(Y_k(v)) \leqq F(v_0, Y_k(v)) \leqq F(v, Y_k(v)) + \chi(\Delta)$$

for all $v \in \Delta$, where $\chi(\Delta)$ tends to 0 as $|\Delta| \to 0$. But $F(v, Y_k(v))$ tends weakly to $F^*(v)$, and since G is linear, $G(Y_k(v))$ tends weakly to $G(Y_0(v))$. Thus

$$\int_\Delta G(Y_0(v)) \, dv \leqq \int_\Delta F^*(v) \, dv + |\Delta| \, \chi(\Delta)$$

for all such Δ. Therefore,

$$F(v_0, Y_0(v_0)) \leqq F^*(v_0).$$

Since this is true for almost all $v_0 \in V$, the theorem is proved.

REFERENCES

[1] D. G. Aronson and J. Serrin, *Local behavior of solutione of quasilinear parabolic equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 81–122.
[2] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.
[3] E. B. Dynkin, *Markov Processes*, Springer, Berlin, 1965.
[4] W. H. Fleming, *Some Markovian optimization problems*, J. Math. Mech., 12 (1963), pp. 131–140.
[5] ———, *The Cauchy problem for degenerate parabolic equations*, Ibid., 13 (1964), pp. 987–1008.
[6] ———, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279.
[7] ———, *Stochastic Lagrange multipliers*, Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.
[8] W. H. Fleming and M. Nisio, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.
[9] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.
[10] I. V. Girsanov, *On Ito's stochastic integral equation*, Soviet Math. Dokl., 2 (1962), pp. 506–509.
[11] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
[12] H. J. Kushner, *On the stochastic maximum principle: fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.
[13] O. A. Ladyzhenskaya and N. N. Ural'ceva, *Boundary problems for linear and quasilinear parabolic equations, I, II*, Amer. Math. Soc. Transl. (2), 47 (1965), pp. 217–299.
[14] J. L. Lions, *On some optimization problems for linear parabolic equations*, Functional Analysis and Optimization, E. R. Caianiello, ed., Academic Press, New York, 1966.
[15] A. Z. Mieri, *A new approach to the general problem of optimal filtering and control of stochastic systems*, Doctoral thesis, Department of Engineering, University of California, Berkeley, 1967.
[16] O. A. Oleinik, A. M. Il'in and A. S. Kalashnikov, *Linear equations of second order of parabolic type*, Russian Math. Surveys, 17 (1962), no. 3, pp. 1–143.
[17] S. Saks, *Theory of the Integral*, Hafner, New York, 1937.
[18] A. V. Skorokhod, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Massachusetts, 1965.
[19] N. V. Krylov, *On quasidiffusion processes*, Teor. Verojatnost. i Primenen, 11 (1966), pp. 424–443.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

OPTIMAL CONTROLS FOR SYSTEMS WITH TIME LAG*

A. HALANAY†

In this paper Pontryagin's maximum principle will be established for some systems with time lag in cases where the control is also delayed. Results of Kharatishvili [1] (see also [1a, Chap. 4, §27]), Friedman [2] and Ožiganova [3] are obtained as particular cases. In the proof the abstract multipliers rule of Hestenes [4] will be applied.[1]

1. The problem. Consider a system of the form

$$x(t) = x_0(b) + \int_{t_0}^{t} \left[\int_{-\tau}^{0} f(t, s, \sigma, x(s), \cdots, x(s-\tau_k), x(s+\sigma), u(s), \cdots, u(s-\tau_k), u(s+\sigma), b)\, d\sigma \right] ds,$$

$$x_{t_0} = \phi_{t_0}, \quad u_{t_0} = \eta_{t_0}, \quad x(T) = \omega(b), \quad \tau_j > 0, \quad j = 1, \cdots, k,$$

and the functionals

$$I_\gamma(u, b) = g_\gamma(b) + \int_{t_0}^{T} \left[\int_{-\tau}^{0} L_\gamma(s, \sigma, x(s), \cdots, x(s-\tau_k), x(s+\sigma), u(s), \cdots, u(s-\tau_k), u(s+\sigma), b)\, d\sigma \right] ds,$$

$$\gamma = 0, 1, \cdots, m.$$

Here x, x_0, ω, $f(t, s, \sigma, x, y_1, \cdots, y_k, z, v_1, \cdots, v_k, w, b)$ are n-vectors; x is the state variable, u the control function and b the control parameters; x_{t_0} is defined as $x_{t_0}(\sigma) = x(t_0 + \sigma)$, $\sigma \in [-\tau_M, 0]$, $\tau_M = \max\{\tau_1, \cdots, \tau_k, \tau\}$, and u_{t_0} is defined in the same way; ϕ is a continuous function defined in $[t_0 - \tau_M, t_0]$, which may depend also on b, and η is piecewise continuous in $[t_0 - \tau_M, t_0]$. To allow the initial function ϕ to be free we suppose that the components of b belong to an abstract linear space which may contain the space of continuous functions defined on $[t_0 - \tau_M, t_0]$.

The control couple (u, b) is admissible if u is piecewise continuous in $[t_0 - \tau_M, T]$, $u(t) \in U$, for $t \in [t_0 - \tau_M, T]$, where U is a given set,

* Received by the editors July 6, 1967, and in revised form October 24, 1967.

† Institutul de Matematică, str. M. Eminescu 47, Bucuresti 9, Romania.

[1] Several results in this direction were communicated at the Scientific Meeting on Equations with Deviating Arguments in Moscow at the University of the Friendship of Peoples, May 1966, and at the Interbalkan Congress of Mathematicians in Bucharest, September 1966. Similar results, for usual systems with time lag, were presented by Kharatishvili [7] at the Conference on Control Systems in Los Angeles, January 1967.

$u_{t_0} = \eta_{t_0}$, $b \in B$, where B is a given *open* set, and if the solution of the system defined by $x_{t_0} = \phi_{t_0}$ corresponding to (u, b) satisfies the conditions $x(T) = \omega(b)$, $I_\gamma \leqq 0$ for $1 \leqq \gamma \leqq m'$, $I_\gamma = 0$ for $m' < \gamma \leqq m$; T is fixed. The couple is optimal if it is admissible and minimizes I_0 in the class of admissible couples. We could allow the initial control also to depend on b but in the general case this fact would imply some complications which we wish to avoid here.

We are looking for necessary conditions for a couple to be optimal. In the following we shall suppose that f and L_γ are sufficiently smooth, in order to avoid nonessential technical difficulties.

The system considered admits as special cases the following situations:

(a) Systems with time lag as

$$\dot{x}(t) = \int_{-\tau}^{0} f(t, \sigma, x(t), x(t-\tau_1), \cdots, x(t-\tau_k), x(t+\sigma), u(t), \cdots, u(t-\tau_k), u(t+\sigma), b)\, d\sigma,$$

$$x_{t_0} = \phi, \quad u_{t_0} = \eta, \quad b \in B \subset R^l,$$

$$I_\gamma(u, b) = g_\gamma(b) + \int_{t_0}^{\tau} \left[\int_{-\tau}^{T} L_\gamma(t, \sigma, x(t), \cdots, x(t-\tau_k), x(t+\sigma), u(t), \cdots, u(t-\tau_k), u(t+\sigma), b)\, d\sigma \right] dt.$$

(b) Hereditary processes as

$$x(t) = x_0 + \int_{t_0}^{t} f(t, s, x(s), u(s), b)\, ds,$$

$$I_\gamma(u, b) = g_\gamma(b) + \int_{t_0}^{T} L_\gamma(s, x(s), u(s), b)\, ds.$$

(c) Hereditary processes as

$$x(t) = x_0 + \int_{t-\tau}^{t} f(t, s, x(s), u(s), b)\, ds, \quad x_{t_0} = \phi, \quad u_{t_0} = \eta,$$

$$I_\gamma(u, b) = g_\gamma(b) + \int_{t_0}^{T} L_\gamma(s, x(s), x(s-\tau), u(s), u(s-\tau), b)\, ds.$$

(It is easy to see that this is a special case of (a).)

(d) The most simple systems with time lag as

$$\dot{x}(t) = f(t, x(t), x(t-\tau), u(t), u(t-\tau), b), \quad x_{t_0} = \phi, \quad u_{t_0} = \eta,$$

$$I_\gamma(u, b) = g_\gamma(b) + \int_{t_0}^{T} L_\gamma(t, x(t), x(t-\tau), u(t), u(t-\tau), b)\, dt.$$

The case (a) is obtained from the general one if f does not depend explicitly on t; indeed, we have, by integrating the equation in (a),

$$x(t) = x(t_0, b) + \int_{t_0}^{t} \left[\int_{-\tau}^{0} f(s, \sigma, x(s), \cdots, x(s - \tau_k), x(s + \sigma), u(s), \cdots, u(s - \tau_k), u(s + \sigma), b)\, d\sigma \right] ds.$$

The case (b) is obtained from the general one if f does not depend on σ and on $x(s - \tau_j)$, $x(s + \sigma)$, $u(s - \tau_j)$, $u(s + \sigma)$. The case (c) is seen to be a particular case of (a) by differentiating with respect to t; we obtain

$$\begin{aligned}
\dot{x}(t) &= f(t, t, x(t), u(t), b) - f(t, t - \tau, x(t - \tau), u(t - \tau), b) \\
&\quad + \int_{t-\tau}^{t} \frac{\partial f}{\partial t}(t, s, x(s), u(s), b)\, ds \\
&= f(t, t, x(t), u(t), b) - f(t, t - \tau, x(t - \tau), u(t - \tau), b) \\
&\quad + \int_{-\tau}^{0} \frac{\partial f}{\partial t}(t, t + \sigma, x(t + \sigma), u(t + \sigma), b)\, d\sigma \\
&= -\int_{-\tau}^{0} \left[\frac{1}{\tau} f(t, t, x(t), u(t), b) - \frac{1}{\tau} f(t, t - \tau, x(t - \tau), u(t - \tau), b) \right. \\
&\quad \left. + \frac{\partial f}{\partial t}(t, t + \sigma, x(t + \sigma), u(t + \sigma), b) \right] d\sigma,
\end{aligned}$$

i.e., a particular case of (a).

The general equation considered is a particular case of the so-called Volterra functional equations. For the general theory of such equations see, for example, [5] and [6], where further references can be found.

2. The general result. Let $(\tilde{u}, \tilde{b})$ be an optimal couple, $\tilde{x}$ the corresponding optimal solution. Denote

$$A(t, s, \sigma) = \frac{\partial f}{\partial x}(t, s, \sigma, \tilde{x}(s), \cdots, \tilde{x}(s - \tau_k), \tilde{x}(s + \sigma), \tilde{u}(s), \cdots, \tilde{u}(s - \tau_k), \tilde{u}(s + \sigma), \tilde{b}),$$

$$B_j(t, s, \sigma) = \frac{\partial f}{\partial y_j}(t, s, \sigma, \tilde{x}(s), \cdots, \tilde{x}(s - \tau_k), \tilde{x}(s + \sigma), \tilde{u}(s), \cdots, \tilde{u}(s - \tau_k), \tilde{u}(s + \sigma), \tilde{b}),$$

$$K(t, s, \sigma) = \frac{\partial f}{\partial z}(t, s, \sigma, \tilde{x}(s), \cdots, \tilde{x}(s - \tau_k), \tilde{x}(s + \sigma), \tilde{u}(s), \cdots, \tilde{u}(s - \tau_k), \tilde{u}(s + \sigma), \tilde{b}),$$

$$c_\gamma(s,\sigma) = \frac{\partial L_\gamma}{\partial x}(s,\sigma,\tilde{x}(s),\cdots,\tilde{x}(s+\sigma),\tilde{u}(s),\cdots,\tilde{u}(s+\sigma),\tilde{b}),$$

$$d_{\gamma j}(s,\sigma) = \frac{\partial L_\gamma}{\partial y_j}(s,\sigma,\tilde{x}(s),\cdots,\tilde{x}(s+\sigma),\tilde{u}(s),\cdots,\tilde{u}(s+\sigma),\tilde{b}),$$

$$K_\gamma(s,\sigma) = \frac{\partial L_\gamma}{\partial z}(s,\sigma,\tilde{x}(s),\cdots,\tilde{x}(s+\sigma),\tilde{u}(s),\cdots,\tilde{u}(s+\sigma),\tilde{b}),$$

where f and L_γ are written as functions of $(t, s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)$ and $(s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)$, respectively. Let

$$A(t) = \int_{-\tau}^{0} A(t,t,\sigma)\,d\sigma, \qquad B_j(t) = \int_{-\tau}^{0} B_j(t,t,\sigma)\,d\sigma,$$

$$A(t,s) = \int_{-\tau}^{0} A(t,s,\sigma)\,d\sigma, \qquad B_j(t,s) = \int_{-\tau}^{0} B_j(t,s,\sigma)\,d\sigma,$$

$$c_\gamma(t) = \int_{-\tau}^{0} c_\gamma(t,\sigma)\,d\sigma, \qquad d_{\gamma j}(t) = \int_{-\tau}^{0} d_{\gamma j}(t,\sigma)\,d\sigma.$$

THEOREM 1. *If $(\tilde{u}, \tilde{b})$ is optimal, there exist multipliers $\lambda_0, \lambda_1, \cdots, \lambda_{m+n}$ with the following properties:*

(a) $\lambda_0^2 + \cdots + \lambda_{m+n}^2 \neq 0$, $\lambda_\gamma \geqq 0$ *for* $0 \leqq \gamma \leqq m'$, $\lambda_\gamma = 0$ $(1 \leqq \gamma \leqq m')$ *if* $I_\gamma(\tilde{u}, \tilde{b}) < 0$.

(b) *Let $\tilde{\psi}$ be the solution of the system*

$$\begin{aligned}\dot{\psi}(t) = {} & -\psi(t)A(t) - \sum_j \psi(t+\tau_j)B_j(t+\tau_j) \\ & - \int_t^{t+\tau} \psi(\alpha)K(\alpha,\alpha,t-\alpha)\,d\alpha - \int_t^T \psi(\zeta)\frac{\partial}{\partial\zeta}A(\zeta,t)\,d\zeta \\ & - \sum_j \int_{t+\tau_j}^T \psi(\zeta)\frac{\partial}{\partial\zeta}B_j(\zeta,t)\,d\zeta \\ & - \int_t^{t+\tau}\left(\int_\alpha^T \psi(\zeta)\frac{\partial}{\partial\zeta}K(\zeta,\alpha,t-\alpha)\,d\zeta\right)d\alpha \\ & + \sum_\gamma \lambda_\gamma\left[c_\gamma(t) + \sum_j d_{\gamma j}(t+\tau_j) + \int_t^{t+\tau} K_\gamma(\alpha,t-\alpha)\,d\alpha\right]\end{aligned}$$

defined by the conditions $\psi(t) \equiv 0$ for $t > T$, $\psi_j(T) = -\lambda_{m+j}$, and let

$$\begin{aligned}& H(t,\sigma,x,y_1,\cdots,y_k,z,u,v_1,\cdots,v_k,w,b) \\ & = \tilde{\psi}(t)f(t,t,\sigma,x,y_1,\cdots,y_k,z,u,v_1,\cdots,v_k,w,b) \\ & \quad + \int_s^T \tilde{\psi}(\zeta)\frac{\partial}{\partial\zeta}f(\zeta,t,\sigma,x,y_1,\cdots,y_k,z,u,v_1,\cdots,v_k,w,b)\,d\zeta \\ & \quad - \sum_{\gamma=0}^{m} \lambda_\gamma L_\gamma(t,\sigma,x,y_1,\cdots,y_k,z,u,v_1,\cdots,v_k,w,b).\end{aligned}$$

Then, for all $u \in U$ and for all $t \in [t_0, T]$ in which $\tilde{u}$ is continuous the following inequality (maximum principle) holds:

$$\begin{aligned}
&\int_{-\tau}^{0} H(t, \sigma, \tilde{x}(t), \cdots, \tilde{x}(t-\tau_k), \tilde{x}(t+\sigma), u, \tilde{u}(t-\tau_1), \cdots, \\
&\qquad\qquad \tilde{u}(t-\tau_k), \tilde{u}(t+\sigma), \tilde{b})\, d\sigma \\
&\quad + \int_{-\tau}^{0} H(t+\tau_1, \sigma, \tilde{x}(t+\tau_1), \tilde{x}(t), \cdots, \tilde{x}(t+\tau_1-\tau_k), \tilde{x}(t+\tau_1+\sigma), \\
&\qquad\qquad \tilde{u}(t+\tau_1), u, \cdots, \tilde{u}(t+\tau_1-\tau_k), \tilde{u}(t+\tau_1+\sigma), \tilde{b})\, d\sigma + \cdots \\
&\quad + \int_{-\tau}^{0} H(t+\tau_k, \sigma, \tilde{x}(t+\tau_k), \tilde{x}(t+\tau_k-\tau_1), \cdots, \tilde{x}(t), \tilde{x}(t+\tau_k+\sigma), \\
&\qquad\qquad \tilde{u}(t+\tau_k), \cdots, u, \tilde{u}(t+\tau_k+\sigma), \tilde{b})\, d\sigma \\
&\quad + \int_{-\tau}^{0} H(t-\sigma, \sigma, \tilde{x}(t-\sigma), \cdots, \tilde{x}(t-\sigma-\tau_k), \tilde{x}(t), \tilde{u}(t-\sigma), \cdots, \\
&\qquad\qquad \tilde{u}(t-\sigma-\tau_k), u, \tilde{b})\, d\sigma \\
&\leqq \int_{-\tau}^{0} H(t, \sigma, \tilde{x}(t), \cdots, \tilde{x}(t+\sigma), \tilde{u}(t), \tilde{u}(t-\tau_1), \cdots, \tilde{u}(t+\sigma), \tilde{b})\, d\sigma \\
&\quad + \int_{-\tau}^{0} H(t+\tau_1, \sigma, \tilde{x}(t+\tau_1), \cdots, \tilde{x}(t+\tau_1+\sigma), \tilde{u}(t+\tau_1), \tilde{u}(t), \cdots, \\
&\qquad\qquad \tilde{u}(t+\tau_1+\sigma), \tilde{b})\, d\sigma + \cdots \\
&\quad + \int_{-\tau}^{0} H(t+\tau_k, \sigma, \tilde{x}(t+\tau_k), \cdots, \tilde{x}(t+\tau_k+\sigma), \tilde{u}(t+\tau_k), \\
&\qquad\qquad \tilde{u}(t+\tau_k-\tau_1), \cdots, \tilde{u}(t+\tau_k+\sigma), \tilde{b})\, d\sigma \\
&\quad + \int_{-\tau}^{0} H(t-\sigma, \sigma, \tilde{x}(t-\sigma), \cdots, \tilde{x}(t-\sigma-\tau_k), \tilde{x}(t), \tilde{u}(t-\sigma), \cdots, \\
&\qquad\qquad \tilde{u}(t-\sigma-\tau_k), \tilde{u}(t), \tilde{b})\, d\sigma.
\end{aligned}$$

(c) *We have the transversality conditions:*

$$\begin{aligned}
&\sum \lambda_\beta \frac{\partial g_\beta}{\partial b}(\tilde{b}) - \tilde{\psi}(t_0)\frac{\partial x_0(\tilde{b})}{\partial b} + \tilde{\psi}(T)\frac{\partial \omega(\tilde{b})}{\partial b} \\
&\qquad - \sum \int_{t_0-\tau_j}^{t_0} \tilde{\psi}(t+\tau_j) B_j(t+\tau_j)\frac{\partial \phi(t, \tilde{b})}{\partial b}\, dt \\
&\qquad - \sum \int_{t_0-\tau}^{t_0} \left\{ \int_{t+\tau_j}^{T} \tilde{\psi}(\zeta)\frac{\partial}{\partial \zeta} B_j(\zeta, t)\, d\zeta \right\} \frac{\partial \phi(t, \tilde{b})}{\partial b}\, dt \\
&\qquad - \sum_{\beta} \int_{t_0-\tau}^{t_0} \left[\int_{t_0}^{t+\tau} \lambda_\beta K_\beta(\alpha, t-\alpha)\, d\alpha \right] \frac{d\phi(t, \tilde{b})}{\partial b}\, dt \\
&\qquad - \int_{t_0-\tau}^{t_0} \left[\int_{t_0}^{t+\tau} \tilde{\psi}(\alpha) K(\alpha, \alpha, t-\alpha)\, d\alpha \right] \frac{\partial \phi(t, \tilde{b})}{\partial b}\, dt \\
&\qquad - \int_{t_0-\tau}^{t_0} \left\{ \int_{t_0}^{t+\tau} \left(\int_{\alpha}^{T} \tilde{\psi}(\zeta)\frac{\partial}{\partial \zeta} K(\zeta, \alpha, t-\alpha)\, d\zeta \right) d\alpha \right\} \frac{\partial \phi(t, \tilde{b})}{\partial b}\, dt
\end{aligned}$$

$$- \sum_j \sum_\beta \int_{t_0-\tau_j}^{t_0} d_{\beta j}(t+\tau_j) \frac{\partial \phi(t, \tilde{b})}{\partial b}\, dt$$
$$- \int_{t_0}^{T} \left[\int_{-\tau}^{0} \frac{\partial H(t, s, \tilde{x}(t), \cdots, \tilde{b})}{\partial b}\, ds \right] dt = 0.$$

Remark 1. In the above formulas all functions are supposed vanishing for arguments greater than T. (We remember that functions f and L are defined only for t, s in $[t_0, T]$.) This allows us to write them in a uniform manner.

Remark 2. The transversality conditions become much simpler in the usual case when ϕ and x_0 do not depend on b and also when f and L_γ do not depend on b; this simplest form of the transversality conditions is

$$\sum \lambda_\beta \frac{\partial g_\beta}{\partial b}(\tilde{b}) + \tilde{\psi}(T) \frac{\partial \omega(\tilde{b})}{\partial b} = 0.$$

If $g_\beta \equiv 0$ we get

$$\sum_{j=1}^{n} \lambda_{m+j} \frac{\partial \omega^j(\tilde{b})}{\partial b} = 0.$$

3. Some preliminary computations. Let functions q_γ be defined by the systems

$$\dot{q}_\gamma(t) = -q_\gamma(t)A(t) - \sum_j q_\gamma(t+\tau_j)B_j(t+\tau_j)$$
$$- \int_t^{t+\tau} q_\gamma(\alpha)K(\alpha, \alpha, t-\alpha)\, d\alpha$$
$$- \int_t^T q_\gamma(\zeta) \frac{\partial}{\partial \zeta} A(\zeta, t)\, d\zeta - \sum_j \int_{t+\tau_j}^T q_\gamma(\zeta) \frac{\partial}{\partial \zeta} B_j(\zeta, t)\, d\zeta$$
$$- \int_t^{t+\tau} \left(\int_\alpha^T q_\gamma(\zeta) \frac{\partial}{\partial \zeta} K(\zeta, \alpha, t-\alpha)\, d\zeta \right) d\alpha - c_\gamma(t)$$
$$- \sum_j d_{\gamma j}(t+\tau_j) - \int_t^{t+\tau} K_\gamma(\alpha, t-\alpha)\, d\alpha$$

and the conditions $q_\gamma(t) \equiv 0$ for $t \geqq T$.

Also let functions p_i be defined by the system

$$\dot{p}(t) = -p(t)A(t) - \sum_j p(t+\tau_j)B_j(t+\tau_j)$$
$$- \int_t^{t+\tau} p(\alpha)K(\alpha, \alpha, t-\alpha)\, d\alpha$$
$$- \int_t^T p(\zeta) \frac{\partial}{\partial \zeta} A(\zeta, t)\, d\zeta - \sum_j \int_{t+\tau_j}^T p(\zeta) \frac{\partial}{\partial \zeta} B_j(\zeta, t)\, d\zeta$$
$$- \int_t^{t+\tau} \left(\int_\alpha^T p(\zeta) \frac{\partial}{\partial \zeta} K(\zeta, \alpha, t-\alpha)\, d\zeta \right) d\alpha$$

and the conditions

$$p_i(t) \equiv 0 \text{ for } t > T,$$

$$p_{ij}(T) = \delta_{ij} = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j, \end{cases}$$

where the p_{ij} are the coordinates of the vector p_i.

Define for $\gamma = 0, 1, \cdots, m$,

$$\begin{aligned} &F_\gamma(s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) \\ &\quad = L_\gamma(s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) \\ &\qquad - \sum_j d_{\gamma j}(s, \sigma) y_j - c_\gamma(s, \sigma) x - K_\gamma(s, \sigma) z \\ &\qquad + q_\gamma(s)[f(s, s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) \\ &\qquad - A(s, s, \sigma) x - \sum_j B_j(s, s, \sigma) y_j - K(s, s, \sigma) z] \\ &\qquad + \int_s^T q_\gamma(\zeta) \frac{\partial}{\partial \zeta} [f(\zeta, s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) \\ &\qquad - A(\zeta, s, \sigma) x - \sum_j B_j(\zeta, s, \sigma) y_j - K(\zeta, s, \sigma) z] \, d\zeta, \end{aligned}$$

and for $i = 1, \cdots, n$,

$$\begin{aligned} &F_{m+i}(s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) \\ &\quad = p_i(s)[f(s, s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) \\ &\qquad - A(s, s, \sigma) x - \sum_j B_j(s, s, \sigma) y_j - K(s, s, \sigma) z] \\ &\qquad + \int_s^T p_i(\zeta) \frac{\partial}{\partial \zeta} [f(\zeta, s, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) \\ &\qquad - A(\zeta, s, \sigma) x - \sum_j B_j(\zeta, s, \sigma) y_j - K(\zeta, s, \sigma) z] \, d\zeta. \end{aligned}$$

It is clear that

$$\frac{\partial F_\rho}{\partial x}(s, \sigma, \tilde{x}(s), \cdots, \tilde{x}(s - \tau_k), \tilde{x}(s + \sigma), \tilde{u}(s), \cdots, \tilde{u}(s + \sigma), \tilde{b}) = 0$$

and the same is true for $\partial F_\rho / \partial y_k$ and $\partial F_\rho / \partial z$, $\rho = 0, 1, \cdots, m, m + 1, \cdots, m + n$.

Define further for $\gamma = 0, 1, \cdots, m$,

$$\begin{aligned} G_\gamma(b) &= g_\gamma(b) + q_\gamma(t_0) x_0(b) + \sum_j \int_{t_0 - \tau_j}^{t_0} d_{\gamma j}(s + \tau_j) \phi(s, b) \, ds \\ &\quad + \int_{t_0 - \tau}^{t_0} \left[\int_{t_0}^{s + \tau} K_\gamma(\alpha, s - \alpha) \, d\alpha \right] \phi(s, b) \, ds \end{aligned}$$

$$+\sum_j \int_{t_0-\tau_j}^{t_0} q_\gamma(s+\tau_j)B_j(s+\tau_j)\phi(s,b)\,ds$$
$$+\int_{t_0-\tau}^{t_0}\left[\int_{t_0}^{s+\tau} q_\gamma(\alpha)K(\alpha,\alpha,s-\alpha)\,d\alpha\right]\phi(s,b)\,ds$$
$$+\sum_j \int_{t_0-\tau}^{t_0}\left\{\int_{s+\tau_j}^{T} q_\gamma(\zeta)\frac{\partial}{\partial\zeta}B_j(\zeta,s)\,d\zeta\right\}\phi(s,b)\,ds$$
$$+\int_{t_0-\tau}^{t_0}\left[\int_{t_0}^{s+\tau}\left(\int_{\alpha}^{T} q_\gamma(\zeta)\frac{\partial}{\partial\zeta}K(\zeta,\alpha,s-\alpha)\,d\zeta\right)d\alpha\right]\phi(s,b)\,ds,$$

and for $i = 1, \cdots, n$,

$$G_{m+i}(b) = -\omega^i(b) + p_i(t_0)x_0(b)$$
$$+\sum_j \int_{t_0-\tau_j}^{t_0} p_i(s+\tau_j)B_j(s+\tau_j)\phi(s,b)\,ds$$
$$+\int_{t_0-\tau}^{t_0}\left[\int_{t_0}^{s+\tau} p_i(\alpha)K(\alpha,\alpha,s-\alpha)\,d\alpha\right]\phi(s,b)\,ds$$
$$+\sum_j \int_{t_0-\tau}^{t_0}\left\{\int_{s+\tau_j}^{T} p_i(\zeta)\frac{\partial}{\partial\zeta}B_j(\zeta,s)\,d\zeta\right\}\phi(s,b)\,ds$$
$$+\int_{t_0-\tau}^{t_0}\left[\int_{t_0}^{s+\tau}\left(\int_{\alpha}^{T} p_i(\zeta)\frac{\partial}{\partial\zeta}K(\zeta,\alpha,s-\alpha)\,d\zeta\right)d\alpha\right]\phi(s,b)\,ds.$$

Let

$$J_\rho(u,b) = G_\rho(b) + \int_{t_0}^{T}\left[\int_{-\tau}^{0} F_\rho(s,\sigma,x(s),\cdots,x(s-\tau_k),x(s+\sigma),u(s),\cdots,u(s+\sigma),b)\,d\sigma\right]ds,$$
$$\rho = 0, 1, \cdots, m, m+1, \cdots, m+n.$$

By a direct computation we see that $J_\gamma(u,b) = I_\gamma(u,b)$ for $\gamma = 0, 1, \cdots, m$, and $J_{m+i}(u,b) = x^i(T) - \omega^i(b)$. We have thus transformed the initial problem in such a way that the final conditions $x(T) = \omega(b)$ become new isoperimetric conditions $J_{m+i}(u,b) = 0$ and in the functionals J_ρ the functions F_ρ have the property that $\partial F_\rho/\partial x$, $\partial F_\rho/\partial y_k$, $\partial F_\rho/\partial z$ vanish along the optimal solution considered.

4. Proof of Theorem 1. To prove Theorem 1 we shall use the following abstract multipliers rule due to Hestenes [4]. Let $J_\rho : \mathcal{C} \to R$, $\rho = 0, 1, \cdots, q$, and let $p_0 \in \mathcal{C}$. A set K of vectors $(k^0, \cdots, k^q)$ is a derivative set for J_ρ at p_0 if for an arbitrary set $k_1, k_2, \cdots, k_N$ of vectors from K

there is a function p: $[0, \delta]^N \to \mathcal{C}$ such that $p(0) = p_0$, the functions $f_\rho(\epsilon) = J_\rho(p(\epsilon)) - J_\rho(p_0)$ are continuous on $[0, \delta]^N$, differentiable for $\epsilon = 0$ and $k_j^\rho = \partial f_\rho(0)/\partial \epsilon_j$.

Let p_0 be such that $J_0(p_0) \leqq J_0(p)$ for all $p \in \mathcal{C}$ for which $J_\rho(p) \leqq 0$, $1 \leqq \rho \leqq q'$, $J_\rho(p) = 0$, $q' < \rho \leqq q$ and let K be a derivative set for J_ρ. Then there exist multipliers $\lambda_0, \lambda_1, \cdots, \lambda_q$, $\sum_{i=0}^{q} \lambda_i^2 \neq 0$, $\lambda_\gamma \geqq 0$ for $0 \leqq \gamma \leqq q'$ and $\lambda_\gamma = 0$ if $J_\rho(p_0) < 0$, such that $\sum_{\rho=0}^{q} \lambda_\rho k^\rho \geqq 0$ for all $k \in K$.

Thus, in order to prove the theorem we shall prove that a certain set of vectors is a derivative set for our functionals.

We shall consider the set K of vectors with coordinates k^ρ given by

$$
\begin{aligned}
k^\rho = \int_{-\tau}^{0} & F_\rho(t, \sigma, \tilde{x}(t), \cdots, \tilde{x}(t - \tau_k), \tilde{x}(t + \sigma), u, \tilde{u}(t - \tau_1), \cdots, \\
& \qquad \tilde{u}(t + \sigma), \tilde{b})\, d\sigma \\
- \int_{-\tau}^{0} & F_\rho(t, \sigma, \tilde{x}(t), \cdots, \tilde{x}(t + \sigma), \tilde{u}(t), \cdots, \tilde{b})\, d\sigma \\
+ \int_{-\tau}^{0} & F_\rho(t + \tau_1, \sigma, \tilde{x}(t + \tau_1), \cdots, \tilde{x}(t + \tau_1 - \tau_k), \tilde{x}(t + \tau_1 + \sigma), \\
& \qquad \tilde{u}(t + \tau_1), u, \cdots, \tilde{u}(t + \tau_1 + \sigma), \tilde{b})\, d\sigma \\
- \int_{-\tau}^{0} & F_\rho(t + \tau_1, \sigma, \tilde{x}(t + \tau_1), \cdots, \tilde{u}(t + \tau_1), \tilde{u}(t), \cdots, \\
& \qquad \tilde{u}(t + \tau_1 + \sigma), \tilde{b})\, d\sigma \\
+ \cdots & \\
+ \int_{-\tau}^{0} & F_\rho(t + \tau_k, \sigma, \tilde{x}(t + \tau_k), \cdots, \tilde{x}(t + \tau_k + \sigma), \tilde{u}(t + \tau_k), \cdots, \\
& \qquad u, \tilde{u}(t + \tau_k + \sigma), \tilde{b})\, d\sigma \\
- \int_{-\tau}^{0} & F_\rho(t + \tau_k, \sigma, \tilde{x}(t + \tau_k), \cdots, \tilde{x}(t + \tau_k + \sigma), \tilde{u}(t + \tau_k), \cdots, \\
& \qquad \tilde{u}(t), \tilde{u}(t + \tau_k + \sigma), \tilde{b})\, d\sigma \\
+ \int_{-\tau}^{0} & F_\rho(t - \sigma, \sigma, \tilde{x}(t - \sigma), \cdots, \tilde{x}(t - \sigma - \tau_k), \tilde{x}(t), \tilde{u}(t - \sigma), \cdots, \\
& \qquad \tilde{u}(t - \sigma - \tau_k), u, \tilde{b})\, d\sigma \\
- \int_{-\tau}^{0} & F_\rho(t - \sigma, \sigma, \tilde{x}(t - \sigma), \cdots, \tilde{x}(t - \sigma - \tau_k), \tilde{x}(t), \tilde{u}(t - \sigma), \cdots, \\
& \qquad \tilde{u}(t - \sigma - \tau_k), \tilde{u}(t), \tilde{b})\, d\sigma
\end{aligned}
$$

for all $u \in U$ and all t, points of continuity of $\tilde{u}$, to which we add the vectors with coordinates k^ρ given by $k^\rho = K^\rho h$, where

$$K^\rho = \frac{\partial_\rho G}{\partial b}(\tilde{b}) + \int_{t_0}^{T} \left[\int_{-\tau}^{0} \frac{\partial F_\rho}{\partial b}(t, \sigma, \tilde{x}(t), \cdots, \tilde{u}(t+\sigma), \tilde{b})\, d\sigma \right] dt$$

and h is arbitrary in the space of the parameters.

Let $k_0, k_1, \cdots, k_N$ be a set of vectors from K. We shall suppose that just one of these vectors, say k_0, is of the form $k_0^\rho = K^\rho h$; if more vectors are of this form the reasoning will be the same.

Then let

$$k_j^\rho = \int_{-\tau}^{0} F_\rho(t_j, \sigma, \tilde{x}(t_j), \cdots, u_j, \cdots, \tilde{u}(t_j+\sigma), \tilde{b})\, d\sigma - \cdots$$

$$- \int_{-\tau}^{0} F_\rho(t_j - \sigma, \sigma, \tilde{x}(t_j - \sigma), \cdots, \tilde{u}(t_j), \tilde{b})\, d\sigma$$

(i.e., in k_j we take $t = t_j$ and $u = u_j \in U$).

Suppose $t_1 \leqq t_2 \leqq \cdots \leqq t_N$ and choose δ such that $t_i + N\delta < t_{i+1}$ if $t_i < t_{i+1}$; let $T_1 = t_1$, $T_j = t_j + \epsilon_1 + \cdots + \epsilon_{j-1}$, $0 \leqq \epsilon_j \leqq \delta' < \delta$, and let $M(\epsilon)$ be the complementary set to $\bigcup_j [T_j, T_j + \epsilon_j]$ in $[t_0, T]$. Consider the couple $(u(t, \epsilon), b(\epsilon))$ defined by $u(t, \epsilon) = u_j$ for $t \in [T_j, T_j + \epsilon_j)$, $u(t, \epsilon) = \tilde{u}(t)$ for $t \in M(\epsilon)$, $b(\epsilon) = \tilde{b} + \epsilon_0 h$, $\epsilon = (\epsilon_0, \epsilon_1, \cdots, \epsilon_N)$. Consider then the corresponding system

$$x(t) = x_0(b(\epsilon)) + \int_{t_0}^{t} \left[\int_{-\tau}^{0} f(t, s, \sigma, x(s), \cdots, u(s+\sigma, \epsilon), b(\epsilon))\, d\sigma \right] ds,$$

$$x_{t_0} = \phi, \quad u_{t_0} = \eta,$$

where ϕ which depends on b is taken in the point $b(\epsilon)$. For ϵ_j small enough using the theorems on continuous dependence on parameters which are true for general time-lag systems this system has a solution $x(t, \epsilon)$ defined on $[t_0, T]$ and such that $\lim_{\epsilon \to 0} x(t, \epsilon) = \tilde{x}(t)$. Let $f_\rho(\epsilon) = J_\rho(u(t, \epsilon), b(\epsilon)) - J_\rho(\tilde{u}, \tilde{b})$.

In natural smoothness conditions these functions will be continuous in $[0, \delta']^{N+1}$ and differentiable in the point $\epsilon = 0$.

We have to calculate the partial derivatives in this point.

In order to make clear how such calculation is carried out we shall consider the case $k = 1$ and $\tau_1 < \tau$.

Let $\epsilon_i = 0$ for $i \neq j > 0$; then $T_j = t_j$, $u(t, \epsilon) = u_j$ for $t \in [t_j, t_j + \epsilon_j)$, $u(t, \epsilon) = \tilde{u}(t)$ for $t \notin [t_j, t_j + \epsilon_j)$, $b(\epsilon) = \tilde{b}$. It will follow that $x(t, \epsilon) = x(t)$ for $t \leqq t_j$. We have further

$$
\begin{aligned}
&f_\rho(0, \cdots, \epsilon_j, \cdots, 0) \\
&= \int_{t_j}^{T} \Big\{ \int_{-\tau}^{0} [F_\rho(t, \sigma, x(t, \epsilon), x(t - \tau_1, \epsilon), x(t + \sigma, \epsilon), u(t, \epsilon), \\
&\qquad u(t - \tau_1, \epsilon), u(t + \sigma, \epsilon), \tilde{b}) - F_\rho(t, \sigma, \tilde{x}(t), \tilde{x}(t - \tau_1), \\
&\qquad \tilde{x}(t + \sigma), \tilde{u}(t), \tilde{u}(t - \tau_1), \tilde{u}(t + \sigma), \tilde{b})]\, d\sigma \Big\}\, dt \\
&= \int_{t_j}^{t_j+\epsilon_j} \int_{-\tau}^{0} + \int_{t_j+\epsilon_j}^{t_j+\tau_1} \int_{-\tau}^{0} + \int_{t_j+\tau_1}^{t_j+\tau_1+\epsilon_j} \int_{-\tau}^{0} \\
&\qquad + \int_{t_j+\tau_1+\epsilon_j}^{t_j+\tau} \int_{-\tau}^{0} + \int_{t_j+\tau}^{t_j+\tau+\epsilon_j} \int_{-\tau}^{0} + \int_{t_j+\tau+\epsilon_j}^{T} \int_{-\tau}^{0} \\
&= \int_{t_j}^{t_j+\epsilon_j} \int_{-\tau}^{t_j-t} + \int_{t_j}^{t_j+\epsilon_j} \int_{t_j-t}^{0} + \int_{t_j+\epsilon_j}^{t_j+\tau_1} \int_{-\tau}^{t_j-t} \\
&\qquad + \int_{t_j+\epsilon_j}^{t_j+\tau_1} \int_{t_j-t}^{t_j-t+\epsilon_j} + \int_{t_j+\epsilon_j}^{t_j+\tau_1} \int_{t_j-t+\epsilon_j}^{0} + \int_{t_j+\tau_1}^{t_j+\tau_1+\epsilon_j} \int_{-\tau}^{t_j-t} \\
&\qquad + \int_{t_j+\tau_1}^{t_j+\tau_1+\epsilon_j} \int_{t_j-t}^{t_j-t+\epsilon_j} + \int_{t_j+\tau_1}^{t_j+\tau_1+\epsilon_j} \int_{t_j-t+\epsilon_j}^{0} + \int_{t_j+\tau_1+\epsilon_j}^{t_j+\tau} \int_{-\tau}^{t_j-t} \\
&\qquad + \int_{t_j+\tau_1+\epsilon_j}^{t_j+\tau} \int_{t_j-t}^{t_j-t+\epsilon_j} + \int_{t_j+\tau_1+\epsilon_j}^{t_j+\tau} \int_{t_j-t+\epsilon_j}^{0} + \int_{t_j+\tau}^{t_j+\tau+\epsilon_j} \int_{-\tau}^{t_j-t+\epsilon_j} \\
&\qquad + \int_{t_j+\tau}^{t_j+\tau+\epsilon_j} \int_{t_j-t+\epsilon_j}^{0} + \int_{t_j+\tau+\epsilon_j}^{T} \int_{-\tau}^{0} \\
&= I_1 + I_2 + I_3 + I_4 + I_5 + I_6 + I_7 + I_8 + I_9 \\
&\qquad + I_{10} + I_{11} + I_{12} + I_{13} + I_{14}.
\end{aligned}
$$

We have

$$
\begin{aligned}
I_1(\epsilon_j) = \int_{t_j}^{t_j+\epsilon_j} dt \int_{-\tau}^{t_j-t} &[F_\rho(t, \sigma, x(t, \epsilon), \tilde{x}(t - \tau_1), \tilde{x}(t + \sigma), u_j, \tilde{u}(t - \tau_1), \\
&\tilde{u}(t + \sigma), \tilde{b}) - F_\rho(t, \sigma, \tilde{x}(t), \cdots, \tilde{u}(t + \sigma), \tilde{b})]\, d\sigma;
\end{aligned}
$$

hence

$$
\begin{aligned}
I_1'(0) = \int_{-\tau}^{0} &[F_\rho(t_j, \sigma, \tilde{x}(t_j), \tilde{x}(t_j - \tau_1), \tilde{x}(t_j + \sigma), u_j, \tilde{u}(t_j - \tau_1), \\
&\tilde{u}(t_j + \sigma), \tilde{b}) - F_\rho(t_j, \sigma, \tilde{x}(t_j), \cdots, \tilde{u}(t_j + \sigma), \tilde{b})]\, d\sigma,
\end{aligned}
$$

and further

$$
I_2(\epsilon_j) = \int_{t_j}^{t_j+\epsilon_j} dt \int_{t_j-t}^{0} [F_\rho(t, \sigma, x(t, \epsilon), \tilde{x}(t - \tau_1), x(t + \sigma, \epsilon),
$$

$$u_j, \tilde{u}(t_j - \tau_1), u_j, \tilde{b}) - F_\rho(t, \sigma, \tilde{x}(t), \cdots, \tilde{u}(t + \sigma), \tilde{b})]\, d\sigma,$$

$$I_2'(\epsilon_j) = \int_{-\epsilon_j}^{0} [F_\rho(t_j + \epsilon_j, \sigma, x(t_j + \epsilon_j, \epsilon), \cdots, u_j, \tilde{b}) - F_\rho(t_j + \epsilon_j, \sigma, \tilde{x}(t_j + \epsilon_j), \cdots, \tilde{u}(t_j + \epsilon_j + \sigma), \tilde{b})]\, d\sigma + \int_{t_j}^{t_j+\epsilon_j} dt \frac{\partial}{\partial \epsilon_j} \int_{t_j - t}^{0} [\cdots]\, d\sigma,$$

and it is obvious that $I_2'(0) = 0$.

We have further

$$I_3(\epsilon_j) = \int_{t_j+\epsilon_j}^{t_j+\tau_1} dt \int_{-\tau}^{t_j - t} [F_\rho(t, \sigma, x(t, \epsilon), \tilde{x}(t - \tau_1), \tilde{x}(t + \sigma), \tilde{u}(t), \tilde{u}(t - \tau_1), \tilde{u}(t + \sigma), \tilde{b}) - F_\rho(t, \sigma, \tilde{x}(t), \cdots, \tilde{u}(t + \sigma), \tilde{b}]\, d\sigma,$$

$$I_3'(\epsilon_j) = -\int_{-\tau}^{-\epsilon_j} [F_\rho(t_j + \epsilon_j, \sigma, x(t_j + \epsilon_j, \epsilon), \cdots, \tilde{b}) - F_\rho(t_j + \epsilon_j, \sigma, \tilde{x}(t_j + \epsilon_j), \cdots, \tilde{b})]\, d\sigma + \int_{t_j+\epsilon_j}^{t_j+\tau_1} dt \int_{-\tau}^{t_j - t} \frac{\partial}{\partial \epsilon_j} F_\rho(t, \sigma, x(t, \epsilon), \tilde{x}(t - \tau_1), \cdots, \tilde{b})\, d\sigma.$$

We see that $I_3'(0) = 0$; the first integral is vanishing since $\lim_{\epsilon \to 0} x(t, \epsilon) = \tilde{x}(t)$, and the second since $(\partial F_\rho / \partial x)(t, \sigma, \tilde{x}(t), \cdots, \tilde{b}) \equiv 0$.

We have then

$$I_4(\epsilon_j) = \int_{t_j+\epsilon_j}^{t_j+\tau_1} dt \int_{t_j - t}^{t_j - t + \epsilon_j} [F_\rho(t, \sigma, x(t, \epsilon), \tilde{x}(t - \tau_1), x(t + \sigma, \epsilon), \tilde{u}(t), \tilde{u}(t - \tau_1), u_j, \tilde{b}) - F_\rho(t, \sigma, \cdots, \tilde{u}(t + \sigma), \tilde{b})]\, d\sigma,$$

$$I_4'(0) = \int_{t_j}^{t_j+\tau_1} [F_\rho(t, t_j - t, \tilde{x}(t), \tilde{x}(t - \tau_1), \tilde{x}(t_j), \tilde{u}(t), \tilde{u}(t - \tau_1), u_j, \tilde{b}) - F_\rho(t, t_j - t, \cdots, \tilde{u}(t_j), \tilde{b})]\, dt;$$

$$I_5(\epsilon_j) = \int_{t_j+\epsilon_j}^{t_j+\tau_1} dt \int_{t_j - t + \epsilon_j}^{0} [F_\rho(t, \sigma, x(t, \epsilon), \cdots, \tilde{u}(t), \tilde{u}(t - \tau_1), \tilde{u}(t + \sigma), \tilde{b}) - F_\rho(t, \sigma, \cdots, \tilde{u}(t + \sigma), \tilde{b}]\, d\sigma,$$

$$I_5'(\epsilon_j) = -\int_{t_j+\epsilon_j}^{t_j+\tau_1} [F_\rho(t, t_j - t + \epsilon_j, x(t, \epsilon), \cdots, \tilde{u}\,(t), \tilde{u}(t - \tau_1), \tilde{u}(t_j + \epsilon_j), \tilde{b}) - F_\rho(t, t_j - t + \epsilon_j, \cdots, \tilde{u}(t_j + \epsilon_j), \tilde{b})]\, d\sigma + \int_{t_j+\epsilon_j}^{t_j+\tau_1} dt \int_{t_j - t + \epsilon_j}^{0} \frac{\partial}{\partial \epsilon_j} F_\rho(t, \sigma, x(t, \epsilon), \tilde{x}(t - \tau_1), x(t + \sigma, \epsilon), \tilde{u}(t), \tilde{u}(t - \tau_1), \tilde{u}(t + \sigma), \tilde{b})\, d\sigma,$$

and we deduce that $I_5'(0) = 0$ since $\lim_{\epsilon\to 0} x(t, \epsilon) = \tilde{x}(t)$ and

$$\frac{\partial F_\rho}{\partial x}(t, \sigma, \tilde{x}(t), \cdots, \tilde{b}) \equiv 0, \qquad \frac{\partial F_\rho}{\partial z}(t, \sigma, \tilde{x}(t), \cdots, \tilde{b}) \equiv 0.$$

Furthermore,

$$I_6(\epsilon_j) = \int_{t_j+\tau_1}^{t_j+\tau_1+\epsilon_j} dt \int_{-\tau}^{t_j-t} [F_\rho(t, \sigma, x(t, \epsilon), x(t-\tau_1, \epsilon), x(t+\sigma, \epsilon), \tilde{u}(t), u_j, \tilde{u}(t+\sigma), \tilde{b}) - F_\rho(t, \sigma, \cdots, \tilde{u}(t+\sigma), \tilde{b})]\, d\sigma$$

and

$$I_6'(0) = \int_{-\tau}^{-\tau_1} [F_\rho(t_j+\tau_1, \sigma, \tilde{x}(t_j+\tau_1), \tilde{x}(t_j), \tilde{x}(t_j+\tau_1+\sigma), \tilde{u}(t_j+\tau_1), u_j, \tilde{u}(t_j+\tau_1+\sigma), \tilde{b}) - F_\rho(t_j+\tau_1, \sigma, \tilde{x}(t_j+\tau_1), \tilde{x}(t_j), \tilde{x}(t_j+\tau_1+\sigma), \tilde{u}(t_j+\tau_1), \tilde{u}(t_j), \tilde{u}(t_j+\tau_1+\sigma), \tilde{b})]\, d\sigma.$$

We have then obviously $I_7'(0) = 0$, and in the same way as above,

$$I_8'(0) = \int_{-\tau_1}^{0} [F_\rho(t_j+\tau_1, \sigma, \tilde{x}(t_j+\tau_1), \tilde{x}(t_j), \tilde{x}(t_j+\tau_1+\sigma), \tilde{u}(t_j+\tau_1), u_j, \tilde{u}(t_j+\tau_1+\sigma), \tilde{b}) - F_\rho(t_j+\tau_1, \sigma, \tilde{x}(t_j+\tau_1), \tilde{x}(t_j), \tilde{x}(t_j+\tau_1+\sigma), \tilde{u}(t_j+\tau_1), \tilde{u}(t_j), \tilde{u}(t_j+\tau_1+\sigma), \tilde{b})]d\sigma,$$

$$I_9'(0) = 0,$$

$$I_{10}'(0) = \int_{t_j+\tau_1}^{t_j+\tau} [F_\rho(t, t_j-t, \tilde{x}(t), \tilde{x}(t-\tau_1), \tilde{x}(t_j), \tilde{u}(t), \tilde{u}(t-\tau_1), u_j, \tilde{b}) - F_\rho(t, t_j-t, \tilde{x}(t), \tilde{x}(t-\tau_1), \tilde{x}(t_j), \tilde{u}(t), \tilde{u}(t-\tau_1), \tilde{u}(t_j), \tilde{b})]\, dt,$$

$$I_{11}'(0) = I_{12}'(0) = I_{13}'(0) = I_{14}'(0) = 0.$$

Hence,

$$\frac{\partial f_\rho(0, \cdots, 0)}{\partial \epsilon_j} = \int_{-\tau}^{0} [F_\rho(t_j, \sigma, \tilde{x}(t_j), \tilde{x}(t_j-\tau_1), \tilde{x}(t_j+\sigma), u_j, \tilde{u}(t_j-\tau_1), \tilde{u}(t_j+\sigma), \tilde{b}) - F_\rho(t_j, \sigma, \tilde{x}(t_j), \tilde{x}(t_j-\tau_1), \tilde{x}(t_j+\sigma), \tilde{u}(t_j), \tilde{u}(t_j-\tau_1), \tilde{u}(t_j+\sigma), \tilde{b})]\, d\sigma$$

$$+ \int_{-\tau}^{0} [F_\rho(t_j+\tau_1, \sigma, \tilde{x}(t_j+\tau_1), \tilde{x}(t_j), \tilde{x}(t_j+\tau_1+\sigma),$$

$$\tilde{u}(t_j + \tau_1), u_j, \tilde{u}(t_j + \tau_1 + \sigma), \tilde{b})$$

$$- F_\rho(t_j + \tau_1, \sigma, \tilde{x}(t_j + \tau_1), \tilde{x}(t_j), \tilde{x}(t_j + \tau_1 + \sigma), \tilde{u}(t_j + \tau_1), \tilde{u}(t_j),$$

$$\tilde{u}(t_j + \tau_1 + \sigma), \tilde{b})]\, d\sigma$$

$$+ \int_{t_j}^{t_j+\tau} [F_\rho(t, t_j - t, \tilde{x}(t), \tilde{x}(t - \tau_1), \tilde{x}(t_j), \tilde{u}(t), \tilde{u}(t - \tau_1), u_j, \tilde{b})$$

$$- F_\rho(t, t_j - t, \tilde{x}(t), \tilde{x}(t - \tau_1), \tilde{x}(t_j), \tilde{u}(t), \tilde{u}(t - \tau_1), \tilde{u}(t_j), \tilde{b})]\, dt.$$

The last integral is written (if we do the change $t_j - t = \sigma$)

$$\int_{-\tau}^{0} [F_\rho(t_j - \sigma, \sigma, \tilde{x}(t_j - \sigma), \tilde{x}(t_j - \sigma - \tau_1), \tilde{x}(t_j),$$

$$\tilde{u}(t_j - \sigma), \tilde{u}(t_j - \sigma - \tau_1), u_j, \tilde{b})$$

$$- F_\rho(t_j - \sigma, \sigma, \tilde{x}(t_j - \sigma), \tilde{x}(t_j - \sigma - \tau_1), \tilde{x}(t_j), \tilde{u}(t_j - \sigma),$$

$$\tilde{u}(t_j - \sigma - \tau_1), \tilde{u}(t_j), \tilde{b})]\, d\sigma,$$

and we see that $\partial f_\rho(0)/\partial \epsilon_j = k_j{}^\rho$.

Now let $\epsilon_i = 0$ for $i > 0$; then $u(t, \epsilon) = \tilde{u}(t)$, $b(\epsilon) = \tilde{b} + \epsilon_0 h$. We have

$$f_\rho(\epsilon_0, 0, \cdots, 0) = G_\rho(b(\epsilon)) - G_\rho(\tilde{b})$$

$$+ \int_{t_0}^{T} dt \int_{-\tau}^{0} [F_\rho(t, \sigma, x(t, \epsilon), x(t - \tau_1, \epsilon), x(t + \sigma, \epsilon), \tilde{u}(t), \tilde{u}(t - \tau_1),$$

$$\tilde{u}(t + \sigma), b(\epsilon)) - F_\rho(t, \sigma, \cdots, \tilde{b})]\, d\sigma;$$

hence

$$\frac{\partial f_\rho}{\partial \epsilon_0}(0, \cdots, 0) = \frac{\partial G_\rho}{\partial b}(\tilde{b})h$$

$$+ \int_{t_0}^{T} dt \int_{-\tau}^{0} \frac{\partial F_\rho(t, s, \tilde{x}(t), \tilde{x}(t - \tau_1), \tilde{x}(t + s), \tilde{u}(t), \tilde{u}(t - \tau_1), \tilde{u}(t + s), \tilde{b})}{\partial b}\, ds\, h$$

$$= K^\rho h = k_0{}^\rho.$$

We have thus proved that the set K is a derivative set for our functionals. By the theorem of Hestenes we have the existence of the multipliers λ_ρ such that $\sum \lambda_\rho k^\rho \geqq 0$ for every $k \in K$. We get

$$\sum_{\rho=0}^{m+n} \lambda_\rho \left[\frac{\partial G_\rho}{\partial b}(\tilde{b}) + \int_{t_0}^{T} dt \int_{-\tau}^{0} \frac{\partial F_\rho(t, s, \tilde{x}(t), \cdots, \tilde{u}(t + s), \tilde{b})}{\partial b}\, ds \right] = 0$$

and

$$\sum_{\rho=0}^{m+n} \lambda_\rho \int_{-\tau}^{0} F_\rho(t, \sigma, \tilde{x}(t), \cdots, \tilde{x}(t-\tau_k), \tilde{x}(t+\sigma), u, \tilde{u}(t-\tau_1), \cdots, \tilde{u}(t+\sigma), \tilde{b})\, d\sigma$$
$$+ \sum_{\rho=0}^{m+n} \lambda_\rho \int_{-\tau}^{0} F_\rho(t+\tau_1, \sigma, \tilde{x}(t+\tau_1), \cdots, \tilde{x}(t+\tau_1-\tau_k), \tilde{x}(t+\tau_1+\sigma), \tilde{u}(t+\tau_1), u, \cdots, \tilde{u}(t+\tau_1+\sigma), \tilde{b})\, d\sigma + \cdots$$
$$+ \sum_{\rho=0}^{m+n} \lambda_\rho \int_{-\tau}^{0} F_\rho(t-\sigma, \sigma, \tilde{x}(t-\sigma), \cdots, \tilde{x}(t-\sigma-\tau_k), \tilde{x}(t), \tilde{u}(t-\sigma), \cdots, \tilde{u}(t-\sigma-\tau_k), u, \tilde{b})\, d\sigma$$
$$\geqq \sum_{\rho=0}^{m+n} \lambda_\rho \int_{-\tau}^{0} F_\rho(t, \sigma, \tilde{x}(t), \cdots, \tilde{x}(t+\sigma), \tilde{u}(t), \tilde{u}(t-\tau_1), \cdots, \tilde{u}(t+\sigma), \tilde{b})\, d\sigma$$
$$+ \sum_{\rho=0}^{m+n} \lambda_\rho \int_{-\tau}^{0} F_\rho(t+\tau_1, \sigma, \tilde{x}(t+\tau_1), \cdots, \tilde{x}(t+\tau_1-\tau_k), \tilde{x}(t+\tau_1+\sigma), \tilde{u}(t+\tau_1), \tilde{u}(t), \cdots, \tilde{u}(t+\tau_1+\sigma), \tilde{b})\, d\sigma + \cdots$$
$$+ \sum_{\rho=0}^{m+n} \lambda_\rho \int_{-\tau}^{0} F_\rho(t-\sigma, \sigma, \tilde{x}(t-\sigma), \cdots, \tilde{x}(t), \tilde{u}(t-\sigma), \cdots, \tilde{u}(t), \tilde{b})\, d\sigma.$$

Remembering the definition of F_ρ we obtain

$$\sum_{\rho=0}^{m+n} \lambda_\rho F_\rho = \sum_{\rho=0}^{m} \lambda_\rho \Big\{ L_\rho(t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)$$
$$+ q_\rho(t) f(t, t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)$$
$$+ \int_s^T q_\rho(\zeta) \frac{\partial}{\partial \zeta} f(\zeta, t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)\, d\zeta \Big\}$$
$$+ \sum_{i=1}^{m} \lambda_{m+i} \Big\{ p_i(t) f(t, t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)$$
$$+ \int_s^T p_i(\zeta) \frac{\partial}{\partial \zeta} f(\zeta, t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)\, d\zeta \Big\}$$
$$+ \cdots,$$

where we have written only the terms which contain $u, v_1, \cdots, v_k, w, b$.

Denote $\tilde{\psi} = -\sum_{\gamma=0}^{m} \lambda_\gamma q_\gamma - \sum_{i=1}^{n} \lambda_{m+i} p_i$; we see that

$$\sum_{\rho=0}^{m+n} \lambda_\rho F_\rho = \sum_{\gamma=0}^{m} \lambda_\gamma L_\gamma - \tilde{\psi}(t) f(t, t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)$$
$$- \int_s^T \tilde{\psi}(\zeta) \frac{\partial}{\partial \zeta} f(\zeta, t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b)\, d\zeta + \cdots$$
$$= -H(t, \sigma, x, y_1, \cdots, y_k, z, u, v_1, \cdots, v_k, w, b) + \cdots.$$

The function $\tilde{\psi}$ is the same as in the statement of the theorem and thus assertion (b) in the statement is obtained.

The conditions of transversality are obtained if we write explicitly $\sum_{\rho=0}^{m+n} \lambda_\rho G_\rho$ and use the fact that B is open.

The theorem is proved.

5. A system with variable lags. We shall consider now a system with variable lags of the form

$$\dot{x}(t) = f(t, x(t), x(t - \tau_1(t)), \cdots, x(t - \tau_k(t)), u(t), u(t - \tau_1(t)), \cdots, u(t - \tau_k(t)), b),$$

$$0 < \tau_i(t) \leqq \tau_{i+1}(t) - \alpha, \quad \alpha > 0, \quad \dot{\tau}_i < 1,$$

$$x_{t_0}(\sigma) = \phi(\sigma, b), \quad \sigma \in [-\tau_k(t_0), 0], \quad u_{t_0}(\sigma) = \eta(\sigma), \quad \sigma \in [-\tau_k(t_0), 0],$$

$$I_\gamma(u, b) = g_\gamma(b) + \int_{t_0}^{T} L_\gamma(t, x(t), x(t - \tau_1(t)), \cdots, x(t - \tau_k(t)), u(t), \cdots, u(t - \tau_k(t)), b)\, dt.$$

Let $\beta_i(t) = t - \tau_i(t)$, $\dot{\beta}_i = 1 - \dot{\tau}_i > 0$; hence the $\beta_i(t)$ are increasing. Denote γ_i the inverse of β_i. We have $\beta_j(\gamma_j(t)) \equiv t$ and $t > \beta_i(t) \geqq \beta_{i+1}(t) + \alpha$. Since the β_i are defined on $[t_0, T]$ it follows that γ_i is defined on $[\beta_i(t_0), \beta_i(T)]$. We shall suppose that $\beta_1(t_0) < \beta_k(T)$; then on $[\beta_1(t_0), \beta_k(T)]$ all γ_i are defined. It is easy to see that there exists $\alpha' > 0$ such that $\gamma_{i+1}(t) - \gamma_i(t) \geqq \alpha'$ for $t \in [\beta_i(t_0), \beta_{i+1}(T)]$, where the two functions γ_i, γ_{i+1} are defined. Let

$$\alpha' = \min_j \inf_{t \in [\beta_i(t_0), \beta_{i+1}(T)]} [\gamma_j(t + \alpha) - \gamma_j(t)];$$

we have $\alpha' > 0$ and $\gamma_{i+1}(t + \alpha) \geqq \gamma_i(t) + \alpha'$. Let

$$t \in [\beta_i(t_0), \beta_{i+1}(T)], \quad s_i = \gamma_i(t), \quad s_{i+1} = \gamma_{i+1}(t);$$

hence $t = \beta_i(s_i) = \beta_{i+1}(s_{i+1})$. We have $\beta_i(s_i) \geqq \beta_{i+1}(s_i) + \alpha$; hence $\beta_{i+1}(s_{i+1}) \geqq \beta_{i+1}(s_i) + \alpha$. It follows that $\gamma_{i+1}[\beta_{i+1}(s_{i+1})] \geqq \gamma_{i+1}[\beta_{i+1}(s_i) + \alpha] \geqq \gamma_{i+1}[\beta_{i+1}(s_i)] + \alpha'$; hence $s_{i+1} \geqq s_i + \alpha'$, i.e., $\gamma_{i+1}(t) \geqq \gamma_i(t) + \alpha'$.

Let now $(\tilde{u}, \tilde{b})$ be an optimal couple, $\tilde{x}$ the corresponding optimal solution. Denote

$$A(t) = \frac{\partial f}{\partial x}(t, \tilde{x}(t), \tilde{x}[\beta_1(t)], \cdots, \tilde{x}[\beta_k(t)], \tilde{u}(t), \cdots, \tilde{u}[\beta_k(t)], \tilde{b}),$$

$$B_j(t) = \frac{\partial f}{\partial y_j}(t, \tilde{x}(t), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}),$$

$$c_\gamma(t) = \frac{\partial L_\gamma}{\partial x}(t, \tilde{x}(t), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}),$$

$$d_{\gamma j}(t) = \frac{\partial L_\gamma}{\partial y_j}(t, \tilde{x}(t), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}).$$

THEOREM 2. *If* $(\tilde{u}, \tilde{b})$ *is optimal there exist multipliers* $\lambda_0, \lambda_1, \cdots, \lambda_{m+n}$ *with the following properties*:

(a) $\lambda_0^2 + \cdots + \lambda_{m+n}^2 \neq 0$, $\lambda_\gamma \geqq 0$ *for* $0 \leqq \gamma \leqq m'$, $\lambda_\gamma = 0$ *if* $I_\gamma(\tilde{u}, \tilde{b}) < 0$.

(b) *Let* $\tilde{\psi}$ *be the solution of the system*

$$\dot{\psi}(t) = -\psi(t)A(t) - \sum_j \psi(\gamma_j(t))B_j(\gamma_j(t))\dot{\gamma}_j(t) + \sum_{\alpha=1}^{m} \lambda_\alpha [c_\alpha(t) + \sum_j d_{\alpha j}(\gamma_j(t))\dot{\gamma}_j(t)]$$

defined by the condition $\psi(t) \equiv 0$ *for* $t > T$, $\psi_j(T) = -\lambda_{m+j}$, *and let*

$$H(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b) = \tilde{\psi}(t)f(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b) - \sum_\alpha \lambda_\alpha L_\alpha(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b).$$

Then, for all $u \in U$ *and for all* $t \in [t_0, T]$ *in which* $\tilde{u}$ *is continuous, the following inequality (maximum principle) holds*:

$$\begin{aligned} &H(t, \tilde{x}(t), \tilde{x}(\beta_1(t)), \cdots, \tilde{x}(\beta_k(t)), u, \tilde{u}(\beta_1(t)), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}) \\ &\quad + H(\gamma_1(t), \tilde{x}(\gamma_1(t)), \tilde{x}(t), \cdots, \tilde{x}(\beta_k(\gamma_1(t))), \tilde{u}(\gamma_1(t)), u, \cdots, \\ &\quad \tilde{u}(\beta_k(\gamma_1(t))), \tilde{b}) + \cdots + H(\gamma_k(t), \tilde{x}(\gamma_k(t)), \tilde{x}(\beta_1(\gamma_k(t))), \cdots, \tilde{x}(t), \\ &\quad \tilde{u}(\gamma_k(t)), \tilde{u}(\beta_1(\gamma_k(t))), \cdots, u, \tilde{b}) \\ &\leqq H(t, \tilde{x}(t), \tilde{x}(\beta_1(t), \cdots, \tilde{x}(\beta_k(t)), \tilde{u}(t), \tilde{u}(\beta_1(t)), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}) \\ &\quad + H(\gamma_1(t), \tilde{x}(\gamma_1(t)), \tilde{x}(t), \cdots, \tilde{x}(\beta_k(\gamma_1(t))), \tilde{u}(\gamma_1(t)), \tilde{u}(t), \cdots, \\ &\quad \tilde{u}(\beta_k(\gamma_1(t))), \tilde{b}) + \cdots + H(\gamma_k(t), \tilde{x}(\gamma_k(t)), \tilde{x}(\beta_1(\gamma_k(t))), \cdots, \tilde{x}(t), \\ &\quad \tilde{u}(\gamma_k(t)), \tilde{u}(\beta_1(\gamma_k(t))), \cdots, \tilde{u}(t), \tilde{b}). \end{aligned}$$

(c) *We have the transversality conditions*:

$$\sum_{\alpha=0}^{m} \lambda_\alpha \frac{\partial g_\alpha}{\partial b}(\tilde{b}) - \tilde{\psi}(t_0)\frac{\partial \phi(t_0, \tilde{b})}{\partial b} + \tilde{\psi}(T)\frac{\partial \omega(\tilde{b})}{\partial b} - \sum_j \int_{t_0}^{\gamma_j(t_0)} \left[\tilde{\psi}(s)B_j(s) + \sum_{\alpha=0}^{m} \lambda_\alpha d_{kj}(s)\right] \frac{\partial \phi(\beta_j(s), \tilde{b})}{\partial b}\, ds - \int_{t_0}^{T} \frac{\partial H(t, \tilde{x}(t), \cdots, \tilde{u}(\beta_k(t)), \tilde{b})}{\partial b}\, dt = 0.$$

We remark once more that in all these formulas functions f and L_α are supposed vanishing for arguments greater than T.

We shall prove this theorem in the same way as for Theorem 1.

Let functions q_α be defined by the systems

$$\dot{q}_\alpha(t) = -q_\alpha(t)A(t) - \sum_j q_\alpha(\gamma(_jt))B_j(\gamma_j(t))\dot{\gamma}_j(t) - c_\alpha(t) - \sum_j d_{\alpha j}(\gamma_j(t))\dot{\gamma}_j(t)$$

and the conditions $q_\alpha(t) \equiv 0$ for $t \geqq T$.

Let also functions p_i be defined by the system

$$\dot{p}(t) = -p(t)A(t) - \sum_j p(\gamma_j(t))B_j(\gamma_j(t))\dot{\gamma}_j(t)$$

and the conditions $p_i(t) \equiv 0$ for $t > T$, $p_{ij}(T) = \delta_{ij}$.

Define for $\alpha = 0, 1, \cdots, m$,

$$F_\alpha(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b) = L_\alpha(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b) - c_\alpha x - \sum_j d_{\alpha j} y_j + q_\alpha(t)[f(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b) - A(t)x - \sum_j B_j(t)y_j],$$

and for $i = 1, \cdots, n$,

$$F_{m+i}(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b) = p_i(t)[f(t, x, y_1, \cdots, y_k, u, v_1, \cdots, v_k, b) - A(t)x - \sum_j B_j(t)y_j].$$

It is clear that $\partial f_\rho/\partial x$ and $\partial f_\rho/\partial y_k$ identically vanish along the optimal solution considered.

Define further for $\alpha = 0, 1, \cdots, m$,

$$G_\alpha(b) = g_\alpha(b) + q_\alpha(t_0)\phi(t_0, b) + \sum_j \int_{t_0}^{\gamma_j(t_0)} [d_{\alpha j}(s) + q_\alpha(s)B_j(s)]\phi(\beta_j(s))\, ds,$$

and for $i = 1, \cdots, n$,

$$G_{m+i}(b) = -\omega^i(b) + p_i(t_0)\phi(t_0, b) + \sum_j \int_{t_0}^{\gamma_j(t_0)} p_i(s)B_j(s)\phi(\beta_j(s))\, ds.$$

Let

$$J_\rho(u, b) = G_\rho(b) + \int_{t_0}^{T} F_\rho(t, x(t), \cdots, x(\beta_k(t)), u(t), \cdots, u(\beta_k(t)), b)\, dt.$$

As in the case considered above we see that

$$J_\alpha(u, b) = I_\alpha(u, b) \quad \text{for} \quad \alpha = 0, 1, \cdots, m,$$

and

$$J_{m+i}(u, b) = x^i(T) - \omega^i(b).$$

We have indeed for $\alpha = 0, 1, \cdots, m$,

$$\int_{t_0}^{T} F_\alpha(t, x(t), \cdots, x(\beta_k(t)), u(t), \cdots, u(\beta_k(t)), b)\, dt$$

$$= \int_{t_0}^{T} L_\alpha(t, x(t), \cdots, u(\beta_k(t)), b)\, dt + \int_{t_0}^{T} \frac{d}{dt}\,(q_\alpha(t)x(t))\, dt$$

$$- \sum_j \int_{t_0}^{\gamma_j(t_0)} [d_{\alpha j}(\sigma) + q_\alpha(\sigma)B_j(\sigma)]x(\beta_j(\sigma))\, d\sigma$$

and

$$\int_{t_0}^{T} F_{m+i}(t, x(t), \cdots, u(\beta_k(t)), b)\, dt$$

$$= \int_{t_0}^{T} \frac{d}{dt}\,(p_i(t)x(t))\, dt - \sum_j \int_{t_0}^{\gamma_j(t_0)} p_i(\sigma)B_j(\sigma)x(\beta_j(\sigma))\, d\sigma.$$

We shall use again the theorem of Hestenes. Consider the set K of vectors with coordinates k^ρ given by

$$k^\rho = F_\rho(t, \tilde{x}(t), \cdots, \tilde{x}(\beta_k(t)), u, \tilde{u}(\beta_1(t)), \cdots, \tilde{u}(\beta_k(t)), \tilde{b})$$

$$- F_\rho(t, \tilde{x}(t), \cdots, \tilde{x}(\beta_k(t)), \tilde{u}(t), \tilde{u}(\beta_1(t)), \cdots, \tilde{u}(\beta_k(t)), \tilde{b})$$

$$+ F_\rho(\gamma_1(t), \tilde{x}(\gamma_1(t)), \tilde{x}(t), \cdots, \tilde{x}(\beta_k(\gamma_1(t))), \tilde{u}(\gamma_1(t)), u, \cdots,$$

$$\tilde{u}(\beta_k(\gamma_1(t))), \tilde{b}) - F_\rho(\gamma_1(t), \tilde{x}(\gamma_1(t)), \cdots, \tilde{u}(\gamma_1(t)), \tilde{u}(t), \cdots,$$

$$\tilde{u}(\beta_k(\gamma_1(t))), \tilde{b}) + \cdots + F_\rho(\gamma_k(t), \tilde{x}(\gamma_k(t)), \cdots, \tilde{x}(t), \tilde{u}(\gamma_k(t)), \cdots,$$

$$u, \tilde{b}) - F_\rho(\gamma_k(t), \tilde{x}(\gamma_k(t)), \cdots, \tilde{x}(t), \tilde{u}(\gamma_k(t)), \cdots, \tilde{u}(t), \tilde{b})$$

for all $u \in U$ and all t, points of continuity of $\tilde{u}$, to which we add the vectors with coordinates k^ρ given by $k^\rho = K^\rho h$, where

$$K^\rho = \frac{\partial G_\rho}{\partial b}\,(\tilde{b}) + \int_{t_0}^{T} \frac{\partial F_\rho}{\partial b}\,(t, \tilde{x}(t), \cdots, \tilde{u}(\beta_k(t)), \tilde{b})\, dt.$$

We have to prove as above that K is a derivative set for our functionals $J_\rho(u, b)$. As above let

$$k_j^\rho = F_\rho(t_j, \tilde{x}(t_j), \cdots, u_j, \tilde{u}(\beta_1(t_j)), \cdots, \tilde{u}(\beta_k(t_j)), \tilde{b})$$

$$- \cdots - F_\rho(\gamma_k(t_j), \tilde{x}(\gamma_k(t_j)), \cdots, \tilde{u}(t_j), \tilde{b}).$$

We shall have

$$f_\rho(0, \cdots, \epsilon_j, \cdots, 0) = \int_{t_j}^{t_j+\epsilon_j} + \int_{t_j+\epsilon_j}^{\gamma_1(t_j)} + \int_{\gamma_1(t_j)}^{\gamma_1(t_j+\epsilon_j)} + \int_{\gamma_1(t_j+\epsilon_j)}^{\gamma_2(t_j)} + \int_{\gamma_2(t_j)}^{\gamma_2(t_j+\epsilon_j)} + \cdots + \int_{\gamma_k(t_j)}^{\gamma_k(t_j+\epsilon_j)} + \int_{\gamma_k(t_j+\epsilon_j)}^{T},$$

and we see that $\partial f_\rho(0)/\partial\epsilon_j = k_j^\rho$. The proof ends as for Theorem 1. Remark that $\gamma_2(t_j) > \gamma_1(t_j + \epsilon_j)$ provided that ϵ_j is small enough since, in our conditions for the lags, $\gamma_2(t) - \gamma_1(t) \geqq \alpha' > 0$.

REFERENCES

[1] G. L. KHARATISHVILI, *The maximum principle in the theory of optimal processes with time lag*, Dokl. Akad. Nauk SSSR, 136 (1961), pp. 39–42.

[1a] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[2] A. FRIEDMAN, *Optimal control for hereditary processes*, Arch. Rational Mech. Anal., 15 (1964), pp. 396–414.

[3] I. A. OZIGANOVA, *On the theory of optimal control of systems with time lag*, Seminar on Differential Equations with Deviating Arguments, University of the Friendship of Peoples, Moscow, vol. II, 1963, pp. 136–145. See also: *On the theory of optimal control for problems with time lag*, thesis, University of the Friendship of Peoples, Moscow, 1966.

[4] MAGNUS R. HESTENES, *On variational theory and optimal control theory*, this Journal, 3 (1965), pp. 23–48.

[5] RODNEY D. DRIVER, *Existence and stability of solutions of a delay-differential system*, Arch. Rational Mech. Anal., 10 (1962), pp. 401–426.

[6] N. M. OĞUZTÖRELI, *Time Lag Control Processes*, Academic Press, New York, 1966.

[7] G. L. KHARATISHVILI, *A maximum principle in extremal problems with delays*, Proc. Conference on Mathematical Theory of Control, University of Southern California, 1967, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 26–34.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

ANALYSIS OF STATISTICAL LINEARIZATION OF NONLINEAR CONTROL SYSTEMS*

J. M. HOLTZMAN†

1. Introduction. The method of statistical linearization due to Booton [1] may be considered to be the random function counterpart to the method of equivalent linearization for periodic functions. Contraction mapping analyses have been found useful in connection with the latter method (see [2] and [3]). The purpose of the present report is to present a contraction mapping type analysis of statistical linearization. We adapt the techniques used by Sandberg [2] for a space of periodic functions square integrable over a period to a Banach space of second order random functions. We obtain conditions for convergence of the method of successive approximations starting with a statistical linearization approximation. A bound on the error between the actual and approximate solutions is given. Our approach is similar to that given in an interesting recent paper by Kolovskii dealing with Kazakov's statistical linearization method (see [4]); however, we dwell more on some aspects of the mathematical analysis. The analysis in the present paper yields the smallest contraction constant in a context explained in [2, pp. 916, 917].

2. Preliminaries and notation. Background material on second order stationary processes is given in Loève [5, §34] and Doob [6, Chap. XI]. It is assumed that there is a fixed probability space $(\Omega, \mathcal{A}, P)$. The required definitions and results are:

(i) We shall use the viewpoint that a random function x on $R = (-\infty, \infty)$ is a mapping of R to a space of random variables on the probability space [5, p. 498].

(ii) For second order random functions, the following function exists:

$$\Gamma_x(t, t') = E[x(t)\bar{x}(t')],$$

where the overbar denotes conjugation. Second order stationarity is defined here by $\Gamma_x(t, t') = \Gamma_x(t - t')$ for all $t, t' \in R$ and $E[x(t)] = \text{const.}$

(iii) Every second order stationary random function x which is continuous[1] at any t has the spectral representation

$$x(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}\, dy(\lambda), \quad E\,|\,dy(\lambda)\,|^2 = dF(\lambda)$$

* Received by the editors August 18, 1967, and in revised form November 22, 1967.

† Bell Telephone Laboratories Incorporated, Whippany, New Jersey 07981.

[1] Throughout this report, continuity of a random function refers to continuity in the quadratic mean.

and

$$\Gamma_x(t - t') = \int_{-\infty}^{\infty} e^{2\pi i(t-t')\lambda}\, dF(\lambda),$$

where F is bounded and nondecreasing (and y has orthogonal increments).

(iv) If x is a continuous second order stationary random function with spectral representation

$$x_a(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}\, dy_a(\lambda),$$

then a linear operation $x_b = Lx_a$ is defined by

$$x_b(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda} H(\lambda)\, dy_a(\lambda),$$

where H is measurable with respect to F_a and

$$\int_{-\infty}^{\infty} |H(\lambda)|^2\, dF_a(\lambda) < \infty$$

(see [6, p. 534]).

3. The method of statistical linearization. Consider the feedback system described by the equation

$$x = LN(r - x), \tag{3.1}$$

where r is assumed to be a real stationary second order continuous random function, L is a linear operation defined by its transfer function $H(\lambda)$ and N is a time-invariant, zero-memory nonlinear operator defined by an odd Lipschitzian real-valued function n of a real variable. The method of statistical linearization replaces (3.1) by

$$x_0 = LK_{eq}(r - x_0), \tag{3.2}$$

where K_{eq} is the constant that minimizes over all constants K the mean square difference between $n(r(t) - x_0(t))$ and $K(r(t) - x_0(t))$.[2] K_{eq} is found to be given by the following formula:

$$K_{eq} = \frac{E\,[(r(t) - x_0(t))n(r(t) - x_0(t))]}{E\,[r(t) - x_0(t)]^2}. \tag{3.3}$$

With this value of K_{eq}, we have

$$\begin{aligned} &E\,[n(r(t) - x_0(t)) - K_{eq}(r(t) - x_0(t))]^2 \\ &\quad = E\,[n(r(t) - x_0(t))]^2 - \frac{E\{[(r(t) - x_0(t))n(r(t) - x_0(t))]\}^2}{E\,[r(t) - x_0(t)]^2}. \end{aligned} \tag{3.4}$$

[2] Ignoring the fact that x_0 may depend on K.

Note that, as with equivalent linearization, K_{eq} does not actually represent a linear operator since K_{eq} depends on its argument. Nevertheless, the method is often not too difficult to apply (though usually requiring some graphical procedure). The usefulness of the method has been experimentally verified in some cases. Our objective is to investigate the method mathematically using a Banach space of random functions.

In the above, it is assumed that we are given a statistical linearization approximation whose accuracy is to be investigated, i.e., the existence of K_{eq} and of an x_0 satisfying (3.2) is assumed. However, our analysis actually shows convergence of the method of successive approximations starting with any x_0 (in a set to be defined) independent of that existence assumption.

4. Spaces of random functions. Though spaces of random functions are familiar to specialists in stochastic processes (see, e.g., [7]), they are not apparently widely used in control theory. We thus devote a few words here to a Banach space which is adequate for the present study but which may not be appropriate for many other analyses.

The following results are given by Dieudonné (see [8, pp. 126–129]):

Let A be any set and F a normed space. A mapping f of A into F is said to be bounded in F if $f(A)$ is bounded in F, i.e., if $\sup_{t\in A} \| f(t) \| < \infty$. The set $B_F(A)$ of all bounded mappings of A into F is a vector space and

$$\| f \| = \sup_{t\in A} \| f(t) \| \tag{4.1}$$

is a norm on this space. If F is a Banach space, $B_F(A)$ is a Banach space.

Now let E be a metric space. Denote by $C_F(E)$ the vector space of all continuous mappings of E into the normed space F and denote by $C_F^{\infty}(E)$ the set of all bounded continuous mappings of E into F. In general, $C_F^{\infty}(E) = C_F(E) \cap B_F(E)$. Consider $C_F^{\infty}(E)$ as a normed subspace of $B_F(E)$. The subspace $C_F^{\infty}(E)$ is closed in $B_F(E)$.

The above results make it easy to construct a Banach space of random functions. Let F be the Banach space of second order random variables (those with finite mean square values).[3] Let x map $R = (-\infty, \infty)$ into F. If x is bounded according to the above definition, x is a second order random function. Then $B_F(R)$ is a Banach space of second order random functions[4] with norm

$$\| x \| = \sup_{t\in R} E^{1/2} | x(t) |^2. \tag{4.2}$$

[3] See [5, pp. 455, 456].

[4] This Banach space does not contain all second order random functions but only those for which there is a uniform bound on the mean square values of the random variables $x(t)$.

$C_F^{\infty}(R)$ is now a closed subspace of second order continuous random functions and when considered as a metric space by itself, it is complete.

We shall work in the subset of $C_F^{\infty}(R)$ consisting of (strictly) stationary random functions. Appropriate conditions will be imposed on the operators L and N so that this subset is mapped into itself. Our object is to show the convergence of a sequence obtained by the method of successive approximations:

$$x_{i+1} = LN(r - x_i) \tag{4.3}$$

with x_0 being the statistical linearization approximation (actually, we shall first modify the functions in the successive approximations). Each x_i will be a real, stationary, continuous, second order random function. With stationarity, the norm of a random function in the Banach space becomes

$$\begin{aligned} \| x \| &= \sup_t E^{1/2} | x(t) |^2 \\ &= E^{1/2} | x(t) |^2 \quad \text{for any} \quad t \in R, \end{aligned} \tag{4.4}$$

and analysis reduces to that of random variables. That is, (quadratic mean) convergence of the random variables at any t implies convergence of the random functions, with convergence of the random functions being given a precise definition.

5. Successive approximations. As mentioned, we shall investigate the convergence of the method of successive approximations

$$x_{i+1} = M(x_i), \tag{5.1}$$

where M is to be defined and with x_0 being the statistical linearization approximation and r being a real, stationary, second order, continuous random function. The following four assumptions will be shown to ensure the convergence:

(i) $H(\lambda)$ is the Fourier transform of a real function h satisfying

$$\int_{-\infty}^{\infty} | h(t) | \, dt < \infty. \tag{5.2}$$

(ii) There are two real constants α and β satisfying $\frac{1}{2}(\alpha + \beta) = 1$ and such that

$$\alpha(u_1 - u_2) \leqq n(u_1) - n(u_2) \leqq \beta(u_1 - u_2) \tag{5.3}$$

for all real $u_1 \geqq u_2$.

(iii)

$$1 + H(\lambda) \neq 0, \quad \lambda \in (-\infty, \infty), \tag{5.4}$$

$$1 + K_{eq}H(\lambda) \neq 0, \quad \lambda \in (-\infty, \infty). \tag{5.5}$$

(iv)

$$\sup_{\lambda} \left| \frac{H(\lambda)}{1 + H(\lambda)} \right| (\beta - 1) = k_2 < 1.^{5} \tag{5.6}$$

Now define the operator $\tilde{N}$ as follows:

$$N(x) = x + \tilde{N}(x) \tag{5.7}$$

or

$$n(x(t)) = x(t) + \tilde{n}(x(t)). \tag{5.8}$$

Then,

$$x = L(r - x) + L\tilde{N}(r - x) \tag{5.9}$$

and

$$(1 + L)x = L\tilde{N}(r - x) + Lr. \tag{5.10}$$

With assumption (5.4), the operator $(1 + L)$ has a bounded inverse $(1 + L)^{-1}$ in the following sense. Let x_a be second order stationary and continuous and

$$x_b = (1 + L)x_a\,, \tag{5.11}$$

$$x_a(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}\, dy_a(\lambda). \tag{5.12}$$

Then

$$x_b(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}(1 + H(\lambda))\, dy_a(\lambda), \tag{5.13}$$

and if

$$x_b(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}\, dy_b(\lambda), \tag{5.14}$$

we have

$$x_a(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}(1 + H(\lambda))^{-1}\, dy_b(\lambda), \tag{5.15}$$

$$\| x_a \| \leqq \sup_{\lambda} | 1 + H(\lambda) |^{-1} \| x_b \|.^{6} \tag{5.16}$$

[5] If we had not used the normalization $\frac{1}{2}(\alpha + \beta) = 1$ for convenience, condition (5.4) would be replaced by $1 + \frac{1}{2}(\alpha + \beta)H(\lambda) \neq 0$, and condition (5.6) by

$$\sup_{\lambda} \left| \frac{H(\lambda)}{1 + \frac{1}{2}(\alpha + \beta)H(\lambda)} \right| \frac{1}{2}(\beta - \alpha) < 1.$$

[6] Since $H(\lambda)$ is a Fourier transform of an L_1 function, it is easily shown that assumption (5.4) is equivalent to $\inf_{\lambda} | 1 + H(\lambda)| > 0$.

Now we can write

$$
\begin{aligned}
x &= (1 + L)^{-1}L\tilde{N}(r - x) + (1 + L)^{-1}Lr \\
&\equiv M(x).
\end{aligned} \tag{5.17}
$$

We shall look for a fixed point of $M(x)$ near the statistical linearization approximation x_0. (The above manipulations exactly parallel those of Sandberg [2].)

We first wish to show that r being stationary, second order, and continuous implies that all the x_i of the method of successive approximations are also stationary, second order, and continuous. We first note that if x is stationary, second order and continuous, then Lx and $N(x)$ are also stationary, second order, and continuous. All of the x_i are stationary, second order, and continuous since each x_i is the result of a combination of L and N operating on r (see (3.2) and (5.1)):

$$
\begin{aligned}
x_0 &= (1 + K_{eq}L)^{-1}LK_{eq}r, \\
x_1 = M(x_0) &= (1 + L)^{-1}\, L\tilde{N}(r - (1 + K_{eq}L)^{-1}LK_{eq}r) \\
&\qquad + (1 + L)^{-1}Lr, \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots
\end{aligned} \tag{5.18}
$$

(We use assumption (5.5) to imply the existence of $(1 + K_{eq}L)^{-1}$.) Furthermore, we have, for all i, that $[\tilde{N}(r - x_i) - \tilde{N}(r - x_{i-1})]$ is stationary, second order, and continuous, since, as above, $[\tilde{N}(r - x_i) - \tilde{N}(r - x_{i-1})]$ is again a combination of L and N operating on r.

With the stationarity, second order and continuity properties assured, we can conveniently use the spectral representation.[7] If

$$
\tilde{n}(r(t) - x_i(t)) - \tilde{n}(r(t) - x_{i-1}(t)) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}\, dy_i(\lambda), \tag{5.19}
$$

$$
E\,[\tilde{n}(r(t) - x_i(t)) - \tilde{n}(r(t) - x_{i-1}(t))]^2 = \int_{-\infty}^{\infty} dF_i(\lambda), \tag{5.20}
$$

then

$$
x_{i+1}(t) - x_i(t) = \int_{-\infty}^{\infty} e^{2\pi i t\lambda}(1 + H(\lambda))^{-1}H(\lambda)\, dy_i(\lambda)\,^{8} \tag{5.21}
$$

[7] A stationary, second order random function is obviously second order stationary.

[8] $H(\lambda)(1 + H(\lambda))^{-1}$ is the Fourier transform of an L_1 function (see [9, pp. 31–32]).

and

$$
\begin{aligned}
&\| x_{i+1} - x_i \| \\
&\quad = E^{1/2}[x_{i+1}(t) - x_i(t)]^2 \\
&\quad \leqq \sup_{\lambda} \left| \frac{H(\lambda)}{1 + H(\lambda)} \right| \sqrt{\int_{-\infty}^{\infty} dF_i(\lambda)} \\
&\quad = \sup_{\lambda} \left| \frac{H(\lambda)}{1 + H(\lambda)} \right| E^{1/2}[\tilde{n}(r(t) - x_i(t)) - \tilde{n}(r(t) - x_{i-1}(t))]^2 \qquad (5.22) \\
&\quad \leqq \sup_{\lambda} \left| \frac{H(\lambda)}{1 + H(\lambda)} \right| \max\{(\beta - 1), (1 - \alpha)\} E^{1/2}[x_i(t) - x_{i-1}(t)]^2 \\
&\quad = \sup_{\lambda} \left| \frac{H(\lambda)}{1 + H(\lambda)} \right| (\beta - 1) \| x_i - x_{i-1} \|.
\end{aligned}
$$

Thus, using assumption (5.6), we have a Cauchy sequence and there is a limit x^* of the sequence x_i.[9] (For the details of this familiar argument see the proof of the contraction mapping fixed-point theorem in almost any book on functional analysis.) Furthermore,

$$
\begin{aligned}
E^{1/2}[x^*(t) - x_0(t)]^2 &\leqq \frac{E^{1/2}[x_1(t) - x_0(t)]^2}{1 - k_2} \\
&\leqq \sup_{\lambda} \left| \frac{H(\lambda)}{1 + H(\lambda)} \right| \qquad (5.23) \\
&\quad \cdot \frac{E^{1/2}[n(r(t) - x_0(t)) - K_{eq}(r(t) - x_0(t))]^2}{1 - k_2}.
\end{aligned}
$$

The last expectation is the minimum root mean square error incurred in the statistical linearization of the nonlinearity by itself (see (3.4)).

6. Discussion of results. The reason for the manipulations of §5 rather than directly determining conditions for $x_{i+1} = LN(r - x_i)$ to be a contracting sequence is that the latter condition would require the "open loop gain" to be less than unity, a most severe restriction. The condition of assumption (5.6) of §5 refers to the closed loop system and is much less restrictive. However, the conditions are still quite restrictive. For example, we have disallowed the common case of the transfer function possessing a pole at $\lambda = 0$. Further work to extend the results would be useful.[10]

[9] x^* is clearly q.m. continuous since $C_F^{\infty}(R)$ is closed. It is second order stationary since it is easily shown that the limit of second order stationary random functions is second order stationary.

[10] The "pole shifting" transformation used in some stability analyses might be used to get around this restriction (for $\alpha > 0$).

Note that we have only shown the convergence of the method of successive approximations. It is not clear whether we get the uniqueness associated with a global contraction. We have not shown that M (see (5.17)) is a global contraction in $C_F^\infty(R)$ since we have not considered its properties with the nonstationary random functions in $C_F^\infty(R)$. Note that the subset of stationary functions is not a linear space since the sum of two stationary functions need not be stationary. Thus, if x_1 and x_2 are stationary, $LN(x_1) - LN(x_2)$ need not be, and the spectral representation cannot be used as above. Conditions for LN (or M) to be a contraction in $C_F^\infty(R)$ could be stated in terms of the convolution kernel (impulse response function) $h(t)$ associated with $H(\lambda)$.[11] However, it then does not appear that a convenient relationship can be stated in terms of $H(\lambda)$ since we have an inequality in the wrong direction (with $h \in L_1$ and $H(\lambda)$ the Fourier transform of $h(t)$):

$$\sup_\lambda | H(\lambda) | \leqq \int_{-\infty}^{\infty} | h(t) | \, dt.$$

In view of our proof not (obviously) giving uniqueness, it is of interest to mention the results of [9]. With a system closely related to that used in the present paper, it is shown in [9] that there is a unique response (up to an equivalence) to an input z which satisfies the "weak finite power" condition

$$\limsup_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} | z(t) |^2 \, dt < \infty.$$

Analysis of the relationship between the sample function and mean square properties of these feedback systems would be of interest (see, e.g., [6, p. 518]).

It is clear from the above remarks that this report is far from being the last word on statistical linearization. Rather, it is hoped that it will stimulate further analysis of the method which is of such great practical utility.

[11] That is, to determine conditions for LN to be a contraction, let $N(x_1)(t) - N(x_2)(t) = z(t)$. Then (with conditions to justify the manipulations),

$$\| LN(x_1) - LN(x_2) \| = \sup_t E^{1/2} \left[\int_{-\infty}^{\infty} h(t-\tau) z(\tau) \, d\tau \int_{-\infty}^{\infty} h(t-s) z(s) \, ds \right]$$

$$\leqq \sup_{\tau,s} E^{1/2} | z(\tau) z(s) | \left[\int_{-\infty}^{\infty} | h(t) | \, dt \right]$$

$$\leqq \beta \left[\int_{-\infty}^{\infty} | h(t) | \, dt \right] \| x_1 - x_2 \|,$$

where β is a Lipschitz constant.

Acknowledgment. The comments of V. E. Beneš and W. L. Roach are greatly appreciated.

REFERENCES

[1] R. C. Booton, Jr., *The analysis of nonlinear control systems with random inputs*, Proc. Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, New York, 1953, pp. 369–391.

[2] I. W. Sandberg, *On the response of nonlinear control systems to periodic input signals*, Bell System Tech. J., 43 (1964), pp. 911–926.

[3] J. M. Holtzman, *Contraction maps and equivalent linearization*, Ibid., 46 (1967), pp. 2405–2435.

[4] M. Z. Kolovskii, *Estimating the accuracy of solutions obtained by the method of statistical linearization*, Automat. Remote Control, 27 (1967), pp. 1692–1701.

[5] M. Loève, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, New Jersey, 1963.

[6] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[7] A. V. Skorokhod, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Massachusetts, 1965.

[8] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[9] V. E. Beneš, *A nonlinear integral equation in the Marcinkiewicz space* $\mathfrak{M}_2$, J. Math. Phys., 44 (1965), pp. 24–35.

[10] A. A. Pervozvanskii, *Random Processes in Nonlinear Control Systems*[12], Academic Press, New York, 1965.

[12] A general reference (with bibliography).

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

SYSTEM IDENTIFICATION AND THE PRINCIPLE OF RANDOM CONTRACTION MAPPING*

K. G. OZA† AND E. I. JURY‡

Abstract. The problem of identifying a linear discrete system is considered where the input-output data is noise-corrupted. An iterative algorithm is suggested which converges in a statistical metric. This convergence is obtained through the principle of random contraction mapping.

1. Introduction. The problem of identifying an "unknown" system among a class of systems is very important for the adaptive control of the system. It is assumed that the unknown system is characterized by a difference equation relating the input time series $\{u(k)\}$ and the output time series $\{x(k)\}$ through an unknown parameter vector $\mathbf{a}$. The input $u(k)$ and the output $x(k)$ are additively corrupted by noise processes $n_1(k)$ and $n_2(k)$, respectively. This noise-corrupted data is supplied to the identifying computer which finds out the explicit error on the basis of this data and the estimate $\boldsymbol{\alpha}$ of the parameter vector $\mathbf{a}$. In order to obtain an asymptotically unbiased estimate of $\mathbf{a}$, an implicit error is defined and a positive definite quadratic form J is obtained from this error. An iterative algorithm is suggested which searches the root $\mathbf{a}$ of the equation $\nabla_{\mathbf{a}} J = 0$. The convergence of this algorithm is obtained using some results of time-series analysis and the principle of random contraction mapping. Finally, a modification of this algorithm is suggested for practical convenience and the modified algorithm is shown to be related to the stochastic approximation method.

The stochastic approximation method has recently been used by Sakrison [1] for the system identification problem. There the requirement of conditional independence in the usual stochastic approximation is replaced by an equivalent condition on the prediction error of the processes involved. In the present paper, the processes are required to be wide-sense stationary up to order four and a finite-time dependence is allowed. Another recent approach to the system identification problem is the instrumental variable method [2], [3]. Unlike this approach and similar to Sakrison's work, we require a knowledge of the correlation functions of the noise which is also

* Received by the editors May 16, 1967, and in revised form November 13, 1967. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant AF-AFOSR-292-66.

† School of Electrical Engineering, University of Oklahoma, Norman, Oklahoma 73069.

‡ Electronics Research Laboratory, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720.

required for the least squares method [4]–[6]. The algorithm proposed in this paper is useful for on-line identification and is computationally considerably simpler than either the instrumental variable or the least squares method and at least as simple as the stochastic approximation method.

2. Statement of the problem. Since the concept of a system about which we know nothing is as meaningless as the idea of "complete ignorance" which confuses the whole field of inductive reasoning, we assume that the system to be identified is described by a difference equation

$$x(k) + \sum_{i=1}^{n} \alpha_i x(k-i) = \sum_{j=0}^{m} \alpha_{j+n+1} u(k-j), \tag{1}$$

where $\{u(k)\}$ = input time-series to the system, $\{x(k)\}$ = output time-series to the system. We shall consider the time index k ranging over the set of nonnegative integers and assume that the necessary initial states are known. The method to be presented remains valid even if the initial states are unknown, provided the identification starts at the time $N = n$. The above system will be called $S(\boldsymbol{\alpha})$.

Let $S(\mathbf{a})$ be the given system where $\mathbf{a} = (a_1, \cdots, a_n, a_{n+1}, \cdots, a_{n+m+1})$ is an unknown vector. Thus the class $\mathbb{S}$ of the system $S(\mathbf{a})$ is identified with the $(n + m + 1)$-dimensional space and the system $S(\mathbf{a})$ is a point in this space. Since we are going to estimate this point on the basis of observations (realizations of random variables), the class $\mathbb{S}$ will really be the space of $(n + m + 1)$-dimensional random vectors. Assuming that all the random variables involved are second order, we can define a statistical metric d in this space by the following:

$$d(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = E\{\| \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}} \|\}, \tag{2}$$

where

$$\| \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}} \|^2 = \sum_{i=1}^{n+m+1} | \hat{\alpha}_i - \hat{\beta}_i |^2. \tag{3}$$

Let $v(k)$ and $y(k)$ be the noise-corrupted input and output data, respectively. Thus

$$v(k) = u(k) + n_1(k), \tag{4}$$

$$y(k) = x(k) + n_2(k). \tag{5}$$

Let the identifying computer assume that the given system is $S(\boldsymbol{\alpha})$ with the input $\{v(k)\}$. Hence the output of the system $S(\boldsymbol{\alpha})$ will be $\{\eta(k)\}$ given by the equation

$$\eta(k) + \sum_{i=1}^{n} \alpha_i \eta(k-i) = \sum_{j=0}^{m} \alpha_{j+n+1} v(k-j). \tag{6}$$

In the real-time identification procedure, the vector $\boldsymbol{\alpha}$ is adjusted sequentially as the new data $v(k)$ and $y(k)$ are obtained. Thus the problem is to deduce an algorithm generating a sequence $\{\boldsymbol{\alpha}(k)\}$ which converges to the point $\mathbf{a}$ in some sense.

The sequence $\{\boldsymbol{\alpha}(k)\}$ generated by the algorithm given in this paper converges to $\mathbf{a}$ in the metric (2).

3. Definitions and preliminaries. Now we shall define several terms, make some assumptions, give certain results without the algebraic derivations and finally reduce the problem to a form to which we can apply the principle of contraction mapping.

DEFINITION. The *explicit error* $\epsilon_e(k)$ is defined as the difference between the output data $y(k)$ and the output of the system $S(\boldsymbol{\alpha})$, i.e.,

$$\epsilon_e(k) = y(k) - \eta(k). \tag{7}$$

DEFINITION. The *identification error* $\mathbf{e}_i$ is defined as the difference between the vectors $\boldsymbol{\alpha}$ and $\mathbf{a}$, i.e.,

$$\mathbf{e}_i = \boldsymbol{\alpha} - \mathbf{a}. \tag{8}$$

Since the system $S(\boldsymbol{\alpha})$ uses its own past output to calculate the present output $\eta(k)$ and since it looks in the past as far back as n instants of time, we are motivated to define the virtual or implicit error $\epsilon_i(k)$ as the weighted sum of the explicit errors, $\epsilon_e(k-j)$, $j = 0, 1, \cdots, n$. The weights in this sum are chosen in order to get eventually a positive definite quadratic form J in the variable $\mathbf{e}_i$.

DEFINITION. The *implicit error* $\epsilon_i(k)$ is defined as

$$\epsilon_i(k) = \epsilon_e(k) + \sum_{i=1}^{n} \alpha_i \epsilon_e(k-i), \tag{9}$$

where $\alpha_1, \cdots, \alpha_n$ are the first n components of $\boldsymbol{\alpha}$.

ASSUMPTION 1. The random processes $\{u(k)\}$, $\{n_1(k)\}$ and $\{n_2(k)\}$ are mutually uncorrelated, zero-mean processes.

ASSUMPTION 2. $\{u(k)\}$, $\{n_1(k)\}$ and $\{n_2(k)\}$ are wide-sense stationary of order 4; that is, the covariance functions

$$\begin{aligned} R_u(j) &= E\{u(k)u(k+j)\}, \\ R_{n_1}(j) &= E\{n_1(k)n_1(k+j)\}, \\ R_{n_2}(j) &= E\{n_2(k)n_2(k+j)\} \end{aligned} \tag{10}$$

and the fourth cumulant functions

$$\begin{aligned} Q_u(j_1, j_2, j_3) &= E\{u(k)u(k+j_1)u(k+j_2)u(k+j_3)\} \\ &- R_u(j_1)R_u(j_2-j_3) - R_u(j_2)R_u(j_1-j_3) - R_u(j_3)R_u(j_1-j_2) \end{aligned} \tag{11}$$

and $Q_{n_1}(j_1, j_2, j_3)$ and $Q_{n_2}(j_1, j_2, j_3)$ (obtained by replacing u in the preceding expression by n_1 and n_2 respectively) are independent of k.

ASSUMPTION 3. The covariance functions and the fourth cumulant functions of $\{u(k)\}$, $\{n_1(k)\}$ and $\{n_2(k)\}$ satisfy the following conditions for $j_1 = 0, 1, \cdots$:

$$\text{(i)} \quad \lim_{N\to\infty} \frac{1}{N} \sum_{j=0}^{N-1} R_\xi^2(j) = 0 \quad \text{for} \quad \xi = u, n_1, n_2;$$

$$\text{(ii)} \quad \lim_{N\to\infty} \frac{1}{N} \sum_{j=0}^{N-j_1-1} Q_\xi(j_1, j, j_1 + j) = 0 \quad \text{for} \quad \xi = u, n_1, n_2.$$

Assumption 1 is made to facilitate many of the derivations that follow and Assumptions 2 and 3 are needed for the consistency of sample covariances and cross-covariances. None of these is too severe a restriction to be satisfied in practice.

ASSUMPTION 4. The two-sided Z-transform [7, pp. 160–163] of the covariance function, $R_u(j)$, of $u(k)$ is strictly positive on some part of the unit circle in the z-plane, i.e., $\Phi_{uu}(e^{j\omega}) \not\equiv 0$ for some ω-set of nonzero measure in $(-\pi, \pi)$.

Now we shall use some vector and matrix notations listed in Appendix A. In terms of these notations, it is shown in Appendix B that

$$E\{\epsilon_i^2(k)\} = e_i' R_q e_i + m_1(\alpha) + m_2(\alpha), \tag{12}$$

where $\mathbf{e}_i = \boldsymbol{\alpha} - \mathbf{a}$, and

$$m_1(\alpha) = E\left\{\left[\sum_{j=0}^{m} \alpha_{j+n+1}\, n_1(k-j)\right]^2\right\}, \tag{13}$$

$$m_2(\alpha) = E\left\{\left[\sum_{i=1}^{n} \alpha_i n_2(k-i) + n_2(k)\right]^2\right\}. \tag{14}$$

LEMMA 1. *Under Assumption* 4, R_q *is positive definite and consequently the function*

$$J(\mathbf{e}_i) = \mathbf{e}_i' R_q \mathbf{e}_i \tag{15}$$

$$= E\{\epsilon_i^2(k)\} - m_1(\alpha) - m_2(\alpha) \tag{16}$$

has a unique minimum at $\mathbf{e}_i = 0$ *or* $\boldsymbol{\alpha} = \mathbf{a}$.

The components of $q(k)$ are linearly independent with probability one under Assumption 4 (see [16]), and then Lemma 1 follows [8, p. 87]. It should be noted that we do not require the coefficient a_n to be nonzero and this in turn eliminates the need of the a priori knowledge about the order of the system $S(\mathbf{a})$.

The minimum of $J(\mathbf{e}_i)$ can be found by taking the gradient with respect

to $\boldsymbol{\alpha}$ and equating it to zero. Assuming that $\operatorname{grad}_{\boldsymbol{\alpha}} \epsilon_i^2(k)$ exists, we can take the gradient of (16), interchange the order of differentiation and expectation and obtain

$$M(\boldsymbol{\alpha}) \triangleq \nabla_{\boldsymbol{\alpha}} J(\mathbf{e}_i) = 2R_{\mathrm{q}}(\boldsymbol{\alpha} - \mathbf{a}) = \mathbf{0} \tag{17}$$

$$= E\{\operatorname{grad}_{\boldsymbol{\alpha}} \epsilon_i^2(k)\} - \operatorname{grad}_{\boldsymbol{\alpha}} m_1(\boldsymbol{\alpha}) - \operatorname{grad}_{\boldsymbol{\alpha}} m_2(\boldsymbol{\alpha}) = \mathbf{0}. \tag{18}$$

It can be shown that

$$\operatorname{grad}_{\boldsymbol{\alpha}} \epsilon_i^2(k) = 2\epsilon_i(k)\mathbf{p}(k) \tag{19}$$

and

$$\operatorname{grad}_{\boldsymbol{\alpha}} m_1(\boldsymbol{\alpha}) + \operatorname{grad}_{\boldsymbol{\alpha}} m_2(\boldsymbol{\alpha}) = 2R_{\mathrm{n}}\boldsymbol{\alpha} + 2\mathbf{r}_{\mathrm{n}} \tag{20}$$

(Appendix B). Using (19) and (20), (18) can be rewritten as

$$(R_{\mathrm{p}} - R_{\mathrm{n}})\boldsymbol{\alpha} + (\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{n}}) = \mathbf{0}. \tag{21}$$

The solution of this equation is clearly $\boldsymbol{\alpha} = \mathbf{a}$ and an iterative scheme to find this solution is suggested in the next section. It is worthwhile to note that the covariance matrix R_{p} and the covariance vector $\mathbf{r}_{\mathrm{p}}$ in (21) are unknown since they depend on the vector $\mathbf{a}$ and also on the probability distribution of the process $\{u(k)\}$.

Finally we assume that the autocorrelation functions of the processes $\{n_1(k)\}$ and $\{n_2(k)\}$ are known.

4. Contraction mapping, random transforms and algorithm. A mapping T of a complete metric space X into itself is said to be a *contraction* if there exists a number c such that $0 < c < 1$ and

$$d(Tx, Ty) \leqq c\, d(x, y) \tag{22}$$

for any two points $x, y \in X$, where d is the distance function on X [10, p. 43].

Defining

$$T_0(\boldsymbol{\alpha}) = \boldsymbol{\alpha} - \gamma[(R_{\mathrm{p}} - R_{\mathrm{n}})\boldsymbol{\alpha} + (\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{n}})] \tag{23}$$

and choosing γ such that

$$\| I - \gamma(R_{\mathrm{p}} - R_{\mathrm{n}}) \| \leqq c < 1, \tag{24}$$

it is clear that T_0 is a contraction mapping on the space of $\boldsymbol{\alpha}$. Since $R_{\mathrm{p}} - R_{\mathrm{n}} = R_{\mathrm{q}}$ is a positive definite symmetric matrix, its eigenvalues are strictly positive; let $0 < \lambda_1 < \cdots < \lambda_{n+m+1}$ be the eigenvalues of R_{q}. Then the condition

$$0 < |1 - \gamma\lambda_i| \leqq c < 1, \qquad i = 1, \cdots, n + m + 1, \tag{25}$$

is equivalent to the condition (24) where c is arbitrary.

If we denote the ith element of the vector $\mathbf{p}(k)$ by $p_i(k)$, $i = 1, 2, \cdots, n + m + 1$, then the matrix R_p has elements of the following form:

$$r_{ij}(t) = E\{p_i(k)p_j(k - t)\}. \tag{26}$$

From a sample of N observations on $\mathbf{p}(k)$, we can consistently estimate $r_{ij}(t)$ by

$$\rho_{ij}^N(t) = \frac{1}{N}\sum_{k=1}^{N} p_i(k)p_j(k - t), \tag{27}$$

which converges to $r_{ij}(t)$ in the mean of order two as $N \to \infty$ if

$$\lim_{N\to\infty} \frac{1}{N}\sum_{s=0}^{N-1} E\{p_i(t + s)p_i(t)p_j(s)p_j(0)\} = r_{ij}^2(t). \tag{28}$$

A sufficient condition for the convergence in (28) is that $p_i(t + s)p_j(s)$ and $p_i(t)p_j(0)$ be independent for $s >$ some integer s_0 [13], pp. [14, pp. 30–32].

Let $\mathfrak{R}_p(N)$ be the matrix $[\rho_{ij}^N(t)]$, where $i, j = 1, \cdots, n + m + 1$ and $t = 0, 1, \cdots, m$ or $t = 0, 1, \cdots, n$. Similarly, a consistent estimate $\varrho_p(N)$ can be defined for $\mathbf{r}_p$. Now we can define a sequence of random transforms

$$\mathbf{T}_N(\boldsymbol{\alpha}) = \boldsymbol{\alpha} - \gamma[(\mathfrak{R}_p(N) - R_n)\boldsymbol{\alpha} + (\varrho_p(N) - \mathbf{r}_n)], \tag{29}$$

which converges to $\mathbf{T}_0(\alpha)$ in the metric defined by (2) for any given value of $\boldsymbol{\alpha}$.

THEOREM. *The sequence of random vectors* $\{\hat{\boldsymbol{\alpha}}_N\}$ *generated by the algorithm*

$$\hat{\boldsymbol{\alpha}}_{N+1} = \mathbf{T}_N(\hat{\boldsymbol{\alpha}}_N) \tag{30}$$

with an arbitrary $\hat{\boldsymbol{\alpha}}_1$, *converges to* $\mathbf{a}$ *in the metric* (2) *as* $N \to \infty$.

The proof of this theorem is given in Appendix C. The convergence with probability one can also be proved following Hanš [11]. The matrix $\mathfrak{R}_p(N)$ and the vector $\varrho_p(N)$ satisfy the recursive relationships

$$\mathfrak{R}_p(N) = \mathfrak{R}_p(N - 1) + \frac{1}{N}[\mathbf{p}(N)\mathbf{p}(N)' - \mathfrak{R}_p(N - 1)] \tag{31}$$

and

$$\varrho_p(N) = \varrho_p(N - 1) + \frac{1}{N}[y(N)\mathbf{p}(N) - \varrho_p(N - 1)], \tag{32}$$

which show that a large storage capacity is not required for the algorithm (30).

The algorithm (30) calls for a value of γ constrained by relation (25). Since the eigenvalues of the matrix R_q depend on the vector $\mathbf{a}$ as well as on the auto-correlation function of the process $\{u(k)\}$, it is not practical to satisfy (25) in the present case. For actual computation, therefore, the

constant γ in (30) can be replaced by a sequence $\{\gamma_N\} = \{A/N\}$, where A is a positive real number. So we have the algorithm

$$\hat{\mathbf{a}}_{N+1} = \hat{\mathbf{a}}_N - \frac{A}{N}[(\mathfrak{R}_{\mathbf{p}}(N) - R_{\mathbf{n}})\hat{\mathbf{a}}_N + (\mathfrak{g}_{\mathbf{p}}(N) - \mathbf{r}_N)]. \tag{33}$$

In order to ensure the convergence of (33), we note that there exists an integer $N_0(A, \lambda_{n+m+1})$ such that for all $N \geqq N_0$ the following is true:

$$0 < \left(1 - \frac{A\lambda_{n+m+1}}{N}\right) \leqq \frac{\| T_0(\boldsymbol{\alpha}) - T_0(\mathbf{a})\|}{\| \boldsymbol{\alpha} - \mathbf{a} \|} \leqq \left(1 - \frac{A\lambda_1}{N}\right) \leqq c. \tag{34}$$

Thus T_0 is still a contraction mapping for $N \geqq N_0$ and the sequence of random transforms T_N with γ replaced by A/N still converges to T_0. Hence the algorithm (33) is convergent. Specifically, we have

$$d(\hat{\mathbf{a}}_{N+1}, \mathbf{a}) \leqq \prod_{k=N_0}^{N} \left(1 - \frac{A\lambda_1}{k}\right) d(\hat{\mathbf{a}}_{N_0}, \mathbf{a}) \tag{35}$$

and, since $A > 0$, $\lambda_1 > 0$ and

$$\sum_{N_0}^{\infty} \frac{1}{k} = \infty,$$

we conclude that $d(\hat{\mathbf{a}}_{N+1}, \mathbf{a})$ converges to zero as N tends to infinity.

Although the algorithm (33) looks similar to the usual stochastic approximation procedure, it is different from the latter as shown below.

Let us consider the random process $\mathbf{Y}(N, \boldsymbol{\alpha})$ defined by

$$\mathbf{Y}(N, \boldsymbol{\alpha}) = F(N)\boldsymbol{\alpha} + \mathbf{f}(N), \tag{36}$$

where

$$F(N) = \mathbf{p}(N)\mathbf{p}(N') - R_{\mathbf{n}}, \quad \mathbf{f}(N) = y(N)\mathbf{p}(N) - \mathbf{r}(N). \tag{37}$$

Then (21) is a regression equation for the process $\mathbf{Y}(N, \boldsymbol{\alpha})$. The usual stochastic approximation procedure dictates the algorithm

$$\hat{\mathbf{a}}_{N+1} = \hat{\mathbf{a}}_N - \gamma_N \mathbf{Y}(N, \hat{\mathbf{a}}_N) \tag{38}$$

for finding the root of the regression equation (21) and requires the conditional distribution of $\mathbf{Y}(N, \hat{\mathbf{a}}_N)$ given $\hat{\mathbf{a}}_1, \cdots, \hat{\mathbf{a}}_N$ to coincide with the distribution of $\mathbf{Y}(N, \hat{\mathbf{a}}_N)$ given $\hat{\mathbf{a}}_N$ only (see [12]). Recently, Sakrison [1] has proposed an algorithm like (37) to solve the system identification problem where the requirement of conditional independence of $\mathbf{Y}(N, \hat{\mathbf{a}}_N)$ is replaced by the condition that the minimum mean square prediction error of the processes involved must decrease at least as fast as $1/\tau^2$ for large values of the prediction time.

Now, in view of (36) and (37), the algorithm (33) can be rewritten

$$\hat{\alpha}_{N+1} = \hat{\alpha}_N - \gamma_N \left[\frac{1}{N} \sum_{k=1}^{N} F(k)\hat{\alpha}_N + \frac{1}{N} \sum_{k=1}^{N} \mathbf{f}(k) \right]. \tag{39}$$

Assumptions 2 and 3 are required for the mean square convergence of the averages appearing in (39). Since the process $\mathbf{Y}(N, \alpha)$ does not satisfy the requirement of conditional independence, we "average out" the process over the observation interval at each stage of the iteration. The recursive relationships (31) and (32) eliminate the need of storing all the data for taking such averages.

To further elucidate the difference between the present approach and the stochastic approximation approach taken in [1], we formally consider the continuous version of the algorithm (33). Let $\mathbf{Y}(t, \alpha)$ be the continuous random process corresponding to (36). Then one can pick $\hat{\alpha}(0)$ arbitrarily and generate the estimate $\hat{\alpha}(t)$ by the integro-differential equation

$$\frac{d\hat{\alpha}(t)}{dt} = -\gamma(t) \left[\frac{1}{t} \int_0^t \mathbf{Y}(\tau, \hat{\alpha}(t))\, d\tau \right], \tag{40}$$

where the function $\gamma(t)$ must satisfy

$$\int_0^\infty \gamma(t)\, dt = \infty, \qquad \int_0^\infty \gamma^2(t)\, dt < \infty. \tag{41}$$

One can compare (40) with the algorithm proposed in [1].

The rate of convergence of (33) will depend on A as well as on other factors, as it does in the usual stochastic approximation procedure. The computational results which show the convergence of (33) in a reasonable amount of computer time are given in [16], where the convergence of (33) with probability one is also established. In particular, the rate of mean square convergence is of the order of $1/N$ when $2\lambda_1 A > 1$. Thus, it is important to make a proper choice of the gain parameter A. This can be done by observing the behavior of the estimation procedure and increasing A if the convergence seems sluggish, or decreasing A if the behavior is highly oscillatory.

5. Conclusion. An algorithm based on the random contraction principle is given for the identification of a linear discrete-time system. Since it allows the finite-time dependence of the random processes involved, it could be utilized without any difficulty for a finite memory system. An analogous treatment can be carried out for linear continuous time systems and for certain nonlinear systems and also for specific time-varying systems. Since the algorithm presented looks similar to the stochastic approximation procedure, it is compared with the latter and the differences are pointed out.

Appendix A. A list of vector and matrix notations.

$'$ denotes the transpose.

$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n, \alpha_{n+1}, \cdots, \alpha_{n+m+1})'$,

$\mathbf{a} = (a_1, \cdots, a_n, a_{n+1}, \cdots, a_{n+m+1})'$,

$\mathbf{q}(k) = (x(k-1), \cdots, x(k-n), -u(k), \cdots, -u(k-m))'$,

$\mathbf{p}(k) = (y(k-1), \cdots, y(k-n), -v(k), \cdots, -v(k-m))'$,

$\mathbf{n}(k) = (n_2(k-1), \cdots, n_2(k-n), -n_1(k), \cdots, -n_1(k-m))'$,

$R_{\mathbf{q}}$ = covaraince matrix $E\{\mathbf{q}(k)\mathbf{q}(k)'\}$,

$R_{\mathbf{p}}$ = covariance matrix $E\{\mathbf{p}(k)\mathbf{p}(k)'\}$,

$R_{\mathbf{n}}$ = covariance matrix $E\{\mathbf{n}(k)\mathbf{n}(k)'\}$,

$\mathbf{r}_n$ = covariance vector $E\{n_2(k)\mathbf{n}(k)\}$,

$\mathbf{r}_{\mathbf{p}}$ = covariance vector $E\{y(k)\mathbf{p}(k)\}$,

$\mathcal{R}_{\mathbf{p}}(N)$ = sample covariance matrix $[\rho_{ij}^N(t)]$,

where

(i) for $i, j = 1, \cdots, n$,

$$\rho_{ij}^N(t) = \frac{1}{N}\sum_{k=1}^{N} y(k-i)y(k-j),$$

and

$$t = |\, i - j \,|;$$

(ii) for $i = 1, \cdots, n$, $j = n+1, \cdots, n+m+1$,

$$\rho_{ij}^N(t) = -\frac{1}{N}\sum_{k=1}^{N} y(k-i)v(k-j-n)$$

and

$$t = |\, i - j - n \,|;$$

(iii) for $i = n+1, \cdots, n+m+1$, $j = 1, \cdots, n$,

$$\rho_{ij}^N(t) = -\frac{1}{N}\sum_{k=1}^{N} v(k-i-n)y(k-j)$$

and

$$t = |\, i - j - n \,|;$$

(iv) for $i, j = n+1, \cdots, n+m+1$,

$$\rho_{ij}^N(t) = \frac{1}{N}\sum_{k=1}^{N} v(k-i-n)v(k-j-n)$$

and

$$t = |\, i - j \,|.$$

$\rho_{\mathbf{p}}(N)$ = sample covariance vector $[\rho_i^N(t)]$,

where

(i) for $i = 1, \cdots, n$,

$$\rho_i{}^N(t) = \frac{1}{N}\sum_{k=1}^{N} y(k-i)y(k)$$

and

$$t = i;$$

(ii) for $i = n+1, \cdots, n+m+1$,

$$\rho_i{}^N(t) = -\frac{1}{N}\sum_{k=1}^{N} v(k-i-n)y(k)$$

and

$$t = i - n.$$

Appendix B. From (4)–(7) and (9) and from the definition of the system $S(\mathbf{a})$, we have

$$\begin{aligned}\epsilon_i(k) &= \sum_{i=1}^{n}(\alpha_i - a_i)y(k-i) - \sum_{j=0}^{m}(\alpha_{j+n+1} - a_{j+n+1})v(k-j) \\ &\quad + n_2(k) + \sum_{i=1}^{n} a_i\, n_2(k-i) - \sum_{j=0}^{m} a_{j+n+1}\, n_1(k-j),\end{aligned} \tag{B1}$$

$$\begin{aligned}\epsilon_i(k) &= \sum_{i=1}^{n}(\alpha_i - a_i)x(k-i) - \sum_{j=0}^{m}(\alpha_{j+n+1} - a_{j+n+1})u(k-j) \\ &\quad + n_2(k) + \sum_{i=1}^{n} \alpha_i\, n_2(k-i) - \sum_{j=0}^{m} \alpha_{j+n+1}\, n_1(k-j),\end{aligned} \tag{B2}$$

$$\epsilon_i(k) = (\boldsymbol{\alpha} - \mathbf{a})'\mathbf{q}(k) + n_2(k) + \boldsymbol{\alpha}'\mathbf{n}(k). \tag{B3}$$

Therefore,

$$E\{\epsilon_i^2(k)\} = (\boldsymbol{\alpha} - \mathbf{a})'E\{\mathbf{q}(k)\mathbf{q}(k)'\}(\boldsymbol{\alpha} - \mathbf{a}) + m_1(\boldsymbol{\alpha}) + m_2(\boldsymbol{\alpha}). \tag{B4}$$

Expanding expressions (11) and (12), we get

$$\begin{aligned}m_1(\boldsymbol{\alpha}) + m_2(\boldsymbol{\alpha}) &= \sum_{j_1}\sum_{j_2} \alpha_{j_1+n+1}\alpha_{j_2+n+1}E\{n_1(k-j_1)n_1(k-j_2)\} \\ &\quad + \sum_{i_1}\sum_{i_2} \alpha_{i_1}\alpha_{i_2}E\{n_2(k-i_1)n_2(k-i_2)\} \\ &\quad + 2\sum_{i} \alpha_i E\{n_2(k)n_2(k-i)\} + E\{n_2^2(k)\} \\ &= \boldsymbol{\alpha}'R_{\mathbf{n}}\boldsymbol{\alpha} + 2\mathbf{r}_{\mathbf{n}}'\boldsymbol{\alpha} + R_{n_2}(0).\end{aligned} \tag{B5}$$

Hence,

$$\text{(B6)} \qquad \operatorname{grad}_{\boldsymbol{\alpha}} m_1(\boldsymbol{\alpha}) + \operatorname{grad}_{\boldsymbol{\alpha}} m_2(\boldsymbol{\alpha}) = 2R_{\mathrm{n}}\boldsymbol{\alpha} + 2r_{\mathrm{n}}.$$

Equation (B1) can be rewritten as

$$\text{(B7)} \qquad \epsilon_i(k) = (\boldsymbol{\alpha} - \mathbf{a})'\mathbf{p}(k) + n_2(k) + \mathbf{a}'\mathbf{n}(k).$$

Therefore,

$$\text{(B8)} \qquad \begin{aligned} \operatorname{grad}_{\boldsymbol{\alpha}} \epsilon_i^2(k) &= 2\epsilon_i(k)\cdot\nabla_{\boldsymbol{\alpha}}\epsilon_i(k) \\ &= 2\epsilon_i(k)\mathbf{p}(k). \end{aligned}$$

Now, using (B6) and (B8) in (18), we obtain

$$\text{(B9)} \qquad E\{\epsilon_i(k)\mathbf{p}(k)\} - R_{\mathrm{n}}\boldsymbol{\alpha} - \mathbf{r}_{\mathrm{n}} = \mathbf{0}.$$

Upon using the definition of $S(\mathbf{a})$ in (1) and carrying out some simplification, substitution of (B7) into (B9) yields

$$\text{(B10)} \qquad (R_{\mathrm{p}} - R_{\mathrm{n}})\boldsymbol{\alpha} + \mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{n}} = 0,$$

which is (21).

Appendix C. The theorem is proved by use of the following lemmas.

LEMMA C1. *If condition* (24) (*equivalently condition* (25)) *is satisfied, then the random transform* $\mathbf{T}_0(\hat{\boldsymbol{\alpha}})$ *is a contraction mapping with respect to metric* (2).

Proof. For any two random vectors $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$, we have

$$\text{(C1)} \qquad d(\mathbf{T}_0(\hat{\boldsymbol{\alpha}}_1), \mathbf{T}_0(\hat{\boldsymbol{\alpha}}_2)) = E\{\|[I - \gamma(R_{\mathrm{p}} - R_{\mathrm{n}})](\hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\alpha}}_2)\|\}.$$

Since for any constant matrix A and any random vector $\hat{\boldsymbol{\alpha}}$,

$$\| A\hat{\boldsymbol{\alpha}} \| \leqq \| A \| \cdot \| \hat{\boldsymbol{\alpha}} \| \quad \text{a.s.},$$

the order-preservation of expectation (Loève [15, p. 119]) implies

$$E\{\| A\hat{\boldsymbol{\alpha}} \|\} \leqq E\{\| A \| \cdot \| \hat{\boldsymbol{\alpha}} \|\} = \| A \| E\{\| \hat{\boldsymbol{\alpha}} \|\}$$

because $\| A \|$ is not a random variable. Application of this fact to (C1) yields

$$\text{(C2)} \qquad \begin{aligned} d(\mathbf{T}_0(\hat{\boldsymbol{\alpha}}_1), \mathbf{T}_0(\hat{\boldsymbol{\alpha}}_2)) &\leqq \| I - \gamma(R_{\mathrm{p}} - R_{\mathrm{n}})\| E\{\| \hat{\boldsymbol{\alpha}}_1 - \hat{\boldsymbol{\alpha}}_2 \|\} \\ &\leqq c\, d(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2). \end{aligned}$$

Equivalence of (24) and (25) follows from the fact that, for any matrix A, $\| A \|^2$ is the largest eigenvalue of $A'A$ and, if A is symmetric, the largest eigenvalue of $A'A$ is the square of the largest eigenvalue of A. Applying

this fact to the symmetric matrix $[I - \gamma(R_p - R_n)]$, the desired result is obtained.

LEMMA C2. *For a given random vector* $\hat{\mathbf{a}}$, $\mathbf{T}_N(\hat{\mathbf{a}})$ *converges to* $\mathbf{T}_0(\hat{\mathbf{a}})$ *in the metric* (2).

Proof. It follows from the triangle inequality that

$$\begin{aligned} d(\mathbf{T}_N(\hat{\mathbf{a}}), \mathbf{T}_0(\hat{\mathbf{a}})) &= E\{\| \gamma[(\mathcal{R}_p(N) - R_p)\boldsymbol{\alpha} - (\varrho_p(N) - \mathbf{r}_p)]\|\} \\ &\leqq \gamma[E\{\|[\mathcal{R}_p(N) - R_p]\boldsymbol{\alpha}\|\} + E\{\|\varrho_p(N) - \mathbf{r}_p\|\}]. \end{aligned} \tag{C3}$$

Now it can be shown by using the Schwarz and triangle inequalities that

$$E\{\|(\mathcal{R}_p(n) - R_p)\alpha\|\} \leqq \left(\sum_{i,j=1}^{n+n+1} E\{|\rho_{ij}^N(t) - r_{ij}(t)|^2\}\right)^{1/2} (E\{\|\alpha\|^2\})^{1/2}.$$

If Assumptions 2 and 3 are satisfied and if n, the order of the system $S(\mathbf{a})$, is finite, the condition (28) is satisfied and each term in the last sum goes to zero as $N \to \infty$. Similarly,

$$\begin{aligned} E\{\|\varrho_p(N) - \mathbf{r}_p(N)\|\} &\leqq \sqrt{E\{\|\varrho_p(N) - \mathbf{r}_p\|^2\}} \\ &= \sqrt{E\left\{\sum_{i=1}^{n+n+1} |\rho_i^N(t) - r_i(t)|^2\right\}} \\ &\leqq \sum_{i=1}^{n+n+1} \sqrt{E\{|\rho_i^N(t) - r_i(t)|^2\}} \end{aligned}$$

goes to zero as $N \to \infty$.

LEMMA C3. *For an arbitrary* $\epsilon > 0$, *there exists an integer* N_1 *such that*

$$d(\hat{\boldsymbol{\alpha}}_N, \mathbf{a}) \leqq \epsilon \quad \textit{for all} \quad N \geqq N_1, \tag{C4}$$

where $\{\hat{\boldsymbol{\alpha}}_N\}$ *is the sequence generated by the algorithm* (30).

Proof. It follows from Lemma 3 that there exists an integer N_2 such that, for every $\hat{\boldsymbol{\alpha}}$,

$$d(\mathbf{T}_N(\hat{\boldsymbol{\alpha}}), \mathbf{T}_0(\hat{\boldsymbol{\alpha}})) \leqq \epsilon^2 \quad \text{for all} \quad N \geqq N_2. \tag{C5}$$

Let ϵ be less than $(1 - c)$.

If there exists a positive integer $N_3 \geqq N_2$ such that

$$d(\hat{\boldsymbol{\alpha}}_{N_3}, \mathbf{a}) \leqq \epsilon, \tag{C6}$$

then

$$\begin{aligned} d(\hat{\boldsymbol{\alpha}}_{N_3+1}, \mathbf{a}) &\leqq d(\mathbf{T}_{N_3}(\hat{\boldsymbol{\alpha}}_{N_3}), \mathbf{T}_0(\hat{\boldsymbol{\alpha}}_{N_3})) + d(\mathbf{T}_0(\hat{\boldsymbol{\alpha}}_{N_3}), \mathbf{a}) \\ &\leqq \epsilon^2 + c\, d(\hat{\boldsymbol{\alpha}}_{N_3}, \mathbf{a}) \\ &\leqq \epsilon(\epsilon + c) < \epsilon. \end{aligned} \tag{C7}$$

By repeating the argument,

$$d(\hat{\mathbf{a}}_N, \mathbf{a}) \leqq \epsilon \quad \text{for all} \quad N \geqq N_3. \tag{C8}$$

Assume that there does not exist any N_3 such that (C6) is satisfied. Then for every integer $N \geqq N_2$,

$$d(\hat{\mathbf{a}}_N, \mathbf{a}) > \epsilon. \tag{C9}$$

In particular, for some integer $\bar{N} > N_2$, we have

$$\begin{aligned} E\{\| \hat{\mathbf{a}}_{\bar{N}+1} - \mathbf{a} \|\} &= E\{\| \mathbf{T}_{\bar{N}}(\hat{\mathbf{a}}_{\bar{N}}) - \mathbf{a} \|\} \\ &\leqq E\{\| \mathbf{T}_{\bar{N}}(\hat{\mathbf{a}}_{\bar{N}}) - \mathbf{T}_0(\hat{\mathbf{a}}_{\bar{N}})\|\} + E\{\| \mathbf{T}_0(\hat{\mathbf{a}}_{\bar{N}}) - \mathbf{a} \|\} \\ &\leqq \epsilon^2 + cE\{\| \hat{\mathbf{a}}_{\bar{N}} - \mathbf{a} \|\} \qquad \text{(from (C5) and Lemma C1)} \\ &\leqq (\epsilon + c)E\{\| \hat{\mathbf{a}}_{\bar{N}} - \mathbf{a} \|\} \qquad \text{(by hypothesis (C6)).} \end{aligned}$$

Since $0 < \epsilon + c < 1$, there exists an integer i such that

$$E\{\| \hat{\mathbf{a}}_{\bar{N}+i} - \mathbf{a} \|\} \leqq (\epsilon + c)^i E\{\| \hat{\mathbf{a}}_{\bar{N}} - \mathbf{a} \|\} \leqq \epsilon.$$

This is a contradiction with the hypothesis. Hence, there must exist at least one integer $N_1 \geqq N_2$ such that

$$d(\hat{\mathbf{a}}_{N_1}, \mathbf{a}) < \epsilon.$$

But then (C8) holds with $N_3 = N_1$.

REFERENCES

[1] D. J. Sakrison, *Use of stochastic approximation to solve the system identification problem*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 563–567.

[2] P. Joseph, J. Lewis and J. Tou, *Plant identification in the presence of disturbances and application to digital adaptive systems*, AIEE. Trans., Part 2, Application and Industry, 80(1961), pp. 18–24.

[3] K. Y. Wong and E. L. Polak, *Identification of linear discrete time systems using the instrumental variable method*, to appear.

[4] R. E. Kalman, *Design of a self-optimizing control system*, Trans. ASME Ser. D. J. Basic Engrg., 80 (1958), pp. 468–478.

[5] M. J. Levin, *Estimation of a system pulse transfer function in the presence of noise*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 229–235.

[6] K. Steiglitz and L. E. McBride, *A technique for the identification of linear systems*, Ibid., AC-10 (1965), pp. 461–464.

[7] E. I. Jury, *Theory and Application of the Z-transform*, John Wiley, New York, 1964.

[8] C. R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley, New York, 1965.

[9] A. Kolmogoroff, *Sur l'interpolation et extrapolation des suites stationnaires*, C. R. Acad. Sci. Paris, 208 (1939), pp. 2043–2045.

[10] A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, vol. 1, Graylock Press, Rochester, 1957.

[11] O. Hanš, *Random fixed point theorems*, Trans. First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Czechoslovak Academy of Sciences, 1957, pp. 105–125.
[12] J. R. Blum, *Multidimensional stochastic approximation methods*, Ann. Math. Statist., 25 (1954), pp. 737–744.
[13] E. Parzen, *An approach to time series analysis*, Ibid., 32 (1961), pp. 951–959.
[14] E. J. Hannan, *Time Series Analysis*, Methuen, London, 1960.
[15] M. Loève, *Probability Theory*, Van Nostrand, New York, 1963.
[16] Kandarp Oza, *Identification problem and random contraction mappings*, Doctoral dissertation, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 1967.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

FURTHER EXTENSIONS OF FIBONACCIAN SEARCH TO NONLINEAR PROGRAMMING PROBLEMS*

PATRICK D. KROLAK†

1. Introduction. In two earlier papers [1], [2], the author described a possible extension of Fibonaccian search to several integer variables and showed that this extension was optimal when compared to a limited class of searches. In this paper, a class of extensions of Fibonaccian search is described. These extensions have the property that under some conditions they are able to accelerate the rate of convergence over previous methods, while in the worst conditions they require no more functional evaluations than the previously described search. In some sense these new extensions are like the set of optimal policies in a game. The second player may hold the first player to the expenditure of a given number of functional evaluations in order to locate the maximum, but if the second player does not choose an optimal strategy, then some of player one's optimal strategies may cost fewer functional evaluations.

2. Theory. This article concerns itself with methods of finding the maximum of a strictly unimodal function over a hyperrectangular lattice of points. In an earlier paper [2] the following terms were defined: vertical ridgeline, strictly unimodal function, connected path; the author will use these terms in the same sense as in the original paper.

Let $L_{m_1, \cdots, m_n}$ be an n-dimensional hyperrectangular array of points such that:

(i) a function $G(i_1, \cdots, i_n)$ is defined at every point;

(ii) $L_{m_1, \cdots, m_n}$ will have points defined for arguments $0 < i_1 \leqq F_{(m_k+1)} - 1$, $m_k \geqq 2$, for all $k = 1, \cdots, n$. We shall call such a lattice a *maximum lattice* of dimension n. Define $VR_{j,l}$ to be the connected subset of all lattice points belonging to $L_{m_1, \cdots, m_n}$ and having their jth coordinate equal to l. Now $VR_{j,l} \in L_{m_1, \cdots, m_{j-1}, m_{j+1}, \cdots, m_n}$ and it is also a ridgeline. In the previous paper we showed how to search such a subset in the required number of functional evaluations. We shall refer to such a set of points as a *hyperplane.*

LEMMA 1. *If $L_{m_1, \cdots, m_n}$ is a maximum lattice and $G(i_1, \cdots, i_n)$ is a strictly unimodal funciion, then we can search $L_{m_1, \cdots, m_n}$ by comparing the maximum values that G takes on over $VR_{1,l}$ and $VR_{1,k}$, where $l \neq k$ and discarding those points of the lattice which are disconnected from the maximum of*

* Received by the editors June 9, 1967, and in revised form December 5, 1967.

† Applied Mathematics Department, Washington University, St. Louis, Missouri. Now at Business Division, Southern Illinois University, Edwardsville, Illinois 62025.

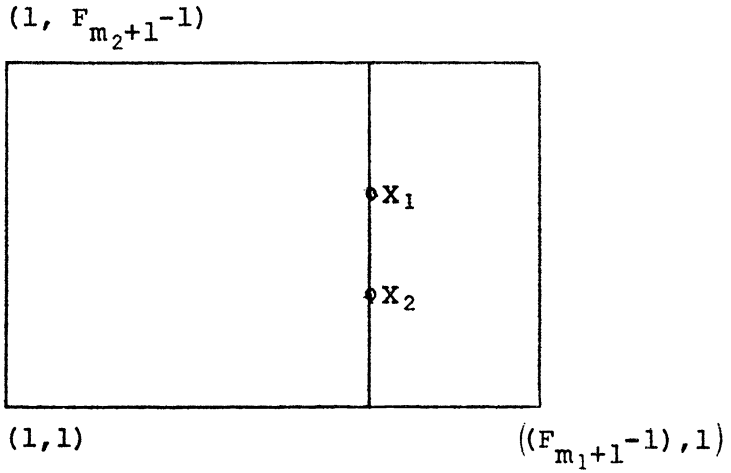


FIG. 1

these two sets when the points of the hyperplane not containing this point are removed from the lattice (see [2]).

We shall call such a search a *parallel hyperplane search.*

LEMMA 2. *If $L_{m_1, \cdots, m_n}$ is a maximum lattice, and $G(i_1, \cdots, i_n)$ is a strictly unimodal function, then $L_{m_1, \cdots, m_n}$ can be searched in exactly $\prod_{i=1}^{n} m_i$ functional evaluations using parallel hyperplanes.*

This search technique will be referred to as the *minimal parallel hyperplane search* (MPHS) method; this search is developed in detail in [1] and [2].

THEOREM. *If $L_{m_1, \cdots, m_n}$ is a maximum lattice and $G(i_1, \cdots, i_n)$ is a strictly unimodal function and if the maximum value of G over $VR_{1,l}$ is located at $X^* = (l, i_2^*, \cdots, i_n^*)$, then we can search $L_{m_1, \cdots, m_n}$ by comparing the maximum of $VR_{2,k}$, where $k \neq i_2^*$, and the maximum of $VR_{1,l}$, and discarding all points disconnected from the greater value when the subset of points associated with the lesser value is removed from the lattice.*

Proof. Suppose $\max(VR_{1,l}) > \max(VR_{2,k})$; then $\max(VR_{2,i_2^*}) \geqq \max(VR_{1,l})$ due to the fact that $X^* \in VR_{2,i_2^*}$ implies that $\max(VR_{2,i_2^*}) > \max(VR_{2,k})$. Now since VR_{2,i^*} and $VR_{2,k}$ are parallel hyperplanes, $VR_{2,i_2^*} \cap VR_{2,k} = 0$. Further, by the definition of a strictly unimodal function, if we remove $VR_{2,k}$ from $L_{m_1, \cdots, m_n}$, a path must still exist between $\max(VR_{2,i_2^*})$ and $\max(L_{m_1, \cdots, m_n})$, but VR_{2,i_2^*} is connected and $X^* \in VR_{2,i_2^*}$. Therefore a path exists between $\max(VR_{1,l})$ and $\max(L_{m_1, \cdots, m_n})$.

We shall call such a search procedure a *perpendicular hyperplane search.* It is important to note sets $VR_{1,l}$ and $VR_{2,k}$ have a nonempty intersection.

Consider the following perpendicular hyperplane search of a two-dimensional maximum lattice L_{m_1,m_2} over which is defined a strictly unimodal function. For this special case VR_1 and VR_2 correspond to vertical and horizontal lines. Assume we are limited to the expenditure of at most $m_1 \times m_2$ functional evaluations. First, find the maximum of $VR_{1,F_{m_1}}$ using the usual Fibonaccian search over an integer variable, where F_{m_1} is the m_1th Fibonaccian number. In doing so, we must evaluate (F_{m_1}, F_{m_2}) and (F_{m_1}, F_{m_2-1}) which we shall call points X_1 and X_2, respectively (see Fig. 1). Now the location of the maximum value of this line $(X_{\max})$ can be in any one of 5 regions:

(i) $X_{\max} > X_1$,
(ii) $X_{\max} = X_1$,
(iii) $X_2 < X_{\max} < X_1$,
(iv) $X_{\max} = X_2$,
(v) $X_{\max} < X_2$.

If either condition (i), (iii) or (iv) occurs, search $VR_{2,F_{m_2}}$; if not, search $VR_{2,F_{m_2-1}}$. Now suppose it happens that we are to search $VR_{2,F_{m_2}}$; in order to do this we must evaluate the point X_1; but since we have already done this we can search $VR_{2,F_{m_2}}$ in $m_1 - 1$ functional evaluations instead of the usual m_1.

Compare $\max(VR_{2,F_{m_2}})$ and $\max(VR_{1,F_{m_1}})$. Suppose $\max(VR_{1,F_{m_1}}) > \max(VR_{2,F_{m_2}})$; that is, the maximum value of these two lines is on the vertical line. At this point we could discard all those points not connected to $\max(VR_1)$ and repeat the above procedure or we could discard the points and use the parallel hyperplane procedure since the points remaining form a maximum lattice and the ridgeline is correctly placed to initiate the search.

Why should we want to use such a search procedure as we have just described? Consider the following possibility that could occur at the time of the first discard: the maximum of the two lines occurred on the vertical line with $X_{\max} > X_1$. Until now we have used $m_1 + (m_2 - 1)$ functional evaluations but we have only discarded $m_1 + 1$ of these. The lattice we have left is a member of L_{m_1,m_2-2} which can be searched in $m_1 \times (m_2 - 2)$ functional evaluations by MPHS. The $m_2 - 2$ points we have already evaluated and not discarded make up the "first line". Hence we need only $(m_1 - 1) \times (m_2 - 2)$ more functional evaluations to complete the search. Thus, for this case we could have searched the lattice in a total of $m_1(m_2 - 2) + m_1 + 1 = m_1 \cdot m_2 - m_1 + 1$ functional evaluations, or $m_1 - 1$ fewer than if we had used the MPHS, which would have taken $m_1 \times m_2$ functional evaluations. Now there are 10 possible cases to consider

TABLE 1

Maximum located on line	Region	Functional evaluations used and discarded	Functional evaluations needed to complete search	Net gain
1	1	$(m_1 + 1)$	$m_1(m_2 - 2)$	$m_1 - 1$
	2	m_1	$m_1(m_2 - 1)$	0
	3	m_1	$m_1(m_2 - 1)$	0
	4	m_1	$m_1(m_2 - 1)$	0
	5	$(m_1 + 1)$	$m_1(m_2 - 2)$	$m_1 - 1$
2	1	$(m_2 + 1)$	$m_2(m_1 - 2)$	$m_2 - 1$
	2	m_2	$m_2(m_1 - 1)$	0
	3	m_2	$m_2(m_1 - 1)$	0
	4	m_2	$m_2(m_1 - 1)$	0
	5	$(m_2 + 1)$	$m_2(m_1 - 2)$	$m_2 - 1$

and the analysis is similar to the above in every case. The results appear in Table 1. Note in every case the method of perpendicular hyperplanes requires no more than MPHS and in 4 cases requires less. Since this situation repeats itself at every discard the perpendicular method requires at most $m_1 \cdot m_2$ functional evaluations and, in actual practice, usually far fewer.

The above method describes a search of a two-dimensional lattice, i.e., a function in two integer variables. Now we shall call a search which holds all variables fixed except one a *one-variable pattern search*, and a search which varies k variables and holds all the other variables fixed, a *k-variable pattern search*.

The MPHS is a one-variable pattern search and the search just described is a two-variable pattern search. It is clear that we can search a $2N$-dimensional lattice with an associated unimodal function $G(x_1, \cdots, x_{2N})$ by defining a new set of functions:

$$f_N(x_1, x_2) = \max_{x_3, x_4} f_{N-1}(x_1, x_2, x_3, x_4),$$

$$\vdots \qquad\qquad \vdots$$

$$f_1(x_1, \cdots, x_{2N}) = \max_{x_{2N-1}, x_{2N}} G(x_1, x_2, \cdots, x_{2N-1}, x_{2N}),$$

and using the two-variable pattern to maximize the above function.

It is evident that we can construct a k-variable pattern search similar to the two-variable one by constructing vertical ridgelines in a k-dimensional lattice which are perpendicular to each other and whose intersection does not include both their maxima. If we use care in selecting our second ridgeline we can realize a possible saving similar to the two-variable pattern, or at most, use the same number of functional evaluations as the MPHS.

The procedure for the k-variable pattern search is as follows:

1. Search $V_{k,F_{m_k}}$ using a $(k-1)$-variable pattern search:
   (a) store location and value of the maximum;
   (b) store complete information on the $(k-2)$-dimensional ridgelines and record them in the order of discard;
   (c) call $V_{k,F_{m_k}}$ hyperplane I.
2. Next determine the set to be labeled hyperplane II. Examine the first two discarded $(k-2)$-dimensional ridgelines of hyperplane I. Either these two ridgelines are perpendicular or they are parallel. If they are perpendicular, define hyperplane II to be that hyperplane perpendicular to hyperplane I whose intersection is the first discarded $(k-2)$-dimensional ridgeline. If the ridgelines are parallel and if they are the first two hyperplanes of an MPHS, then hyperplane II is the hyperplane perpendicular to hyperplane I whose intersection is that ridgeline of the first discarded pair, and which is closest to the maximum of hyperplane I; if not, then define hyperplane II as before.
3. Construct hyperplane II as in step 1 using the information in the intersection set for the first $(k-2)$-dimensional ridgeline.
4. Compare maxima of the two hyperplanes and discard in the same manner as in step 2.
5. Relabel the surviving hyperplane.
6. Either you have eliminated all possibilities, in which case you are through, or go to step 2.

The above procedure allows a greater possible savings in functional evaluations as k increases, but as k increases so must the amount of information we must record. The lower bound on the number of functional evaluations needed to search a k-dimensional hyperlattice cube is of the order $m^k/2^{k-1}$ for large m.

3. Results. A computer program was written in FORTRAN IV for an IBM 7044 for the one-, two- and three-variable patterns. It solved integer nonlinear programming problems of up to nine variables. The results in general agreed with the theory that a three-variable search will require fewer functional evaluations than a one- or two-variable search. However, for the unconstrained case, the number of functional evaluations still grows too fast to solve large problems.

A slight modification of the code may have a use, however, for solving certain nonlinear problems which have very tight constraints, i.e., a small feasible region and a space occupation matrix. The occupation matrix is a 0-1 matrix associated with the inequalities which make up the constraint set. If there are m inequalities and n variables, then define the occupation matrix to be that $m \times n$ matrix whose elements $m_{ij} = 1$ if the jth variable is involved in the ith inequality, and $m_{ij} = 0$ if not.

In order to use the modification, it will be necessary to recast the constrained optimization problem into a new form. A new objective function is defined (which may not be strictly unimodal even if the original problem was strictly unimodal), which will be a vector instead of a scalar.

Let the original problem be:

$$\max z = f(\mathbf{x}),$$
$$g_1(\mathbf{x}) \leqq b_1,$$
$$\vdots$$
$$g_m(\mathbf{x}) \leqq b_m,$$

where $\mathbf{L} \leqq \mathbf{x} \leqq \mathbf{U}$ and $\mathbf{x}$ is an integer n-vector.

Now define the new problem to be:

$$\max \hat{z} = F(i, \xi),$$

where $\mathbf{L} \leqq \mathbf{x} \leqq \mathbf{U}$ and $\mathbf{x}$ is an integer n-vector, and we maximize in the lexicographic sense. Here $F(i, \xi)$ is defined to be a two-component vector with the first component i equal to the first constraint not satisfied by the vector $\mathbf{x}$ and the second component ξ equal to $b_i - g_i(\mathbf{x})$. Of course, if $\mathbf{x}$ satisfies all m constraints, $i = m + 1$ and $\xi = f(\mathbf{x})$. The usual definition of lexicographic comparison is here employed, i.e., $F(i_1, \xi_1) > F(i_2, \xi_2)$ implies either $i_1 > i_2$ or $i_1 = i_2$ and $\xi_1 > \xi_2$. For this new problem formulation it pays to reorder the constraints. The reason is that if a k-variable pattern search is used, then after one sets the first k variables a check can be made on the first constraints that involve only the first k variables to see if a feasible point is possible. If no feasible point is possible, then we have evaluated $F(i, \xi)$ for all $\mathbf{x}$ having those first k components, and we can move to another point in our k-variable pattern search. If a feasible point is possible for those fixed values of the first k variables, then we drop down to the second set of k variables and continue until we can assign a vector value to the point in the first level search. Note the efficiency of this procedure in that we do not have to set all variables before we arrive at a vector value in the first level. This may reduce the amount of necessary work by many orders of magnitude.

This approach is extremely simple but it worked well on ten problems which were in part taken from the literature of continuous nonlinear functions, and of which the following is an example.

Example.

$$\max z = x_1^2 + x_2^{x_1} + x_3x_4x_5x_6 + x_7x_8^{x_9},$$
$$(x_7 - 2)^2 + (x_8 - 2)^2 = 0,$$
$$\sum_{i=4}^{9} x_i \leqq 10,$$

$$x_4^2 + 4x_5x_6 \leqq 9,$$

$$x_1 + x_2 \leqq 4,$$

$$2x_1^2 + x_2^2 \leqq 15,$$

$$1 \leqq x_i \leqq 4, \qquad i = 1, \cdots, 9,$$

x_i is an integer.

The problem was solved in 190 functional evaluations using the modified three-variable pattern search. If the one-, two- or three-variable pattern search without the vector functional had been used it would have required approximately 20,000, 12,000 and 10,000 functional evaluations, respectively. Since the lattice has 262,144 points on it, the possibility of finding the optimum by Monte Carlo methods in 190 functional evaluations would seem remote. For the 10 problems attempted these results are typical and indicate that the modified search does not require work which grows exponentially with the number of variables. Thus for problems meeting requirements mentioned earlier the modified k-variable pattern search would seem to be far superior to the k-variable pattern search.

A few comments are in order before proceeding. It is a trivial result to show that if $F(i, \xi)$ is a strictly unimodal function and if the original problem has a feasible solution, then the maximum of $\hat{z}$ will be the optimal solution to the original problem. However, the author has not found the conditions under which $F(i, \xi)$ will be strictly unimodal. An efficient algorithm which relabels the variables and reorders the constraints in the manner described above, i.e., places the occupation matrix into row echelon form, has been coded. The algorithm is based on the work of Jewell [3].

The program does not always find a global maximum if the function is nonunimodal, but will find a local maximum since it must evaluate all of the feasible neighbors surrounding the local maximum. The program takes about 1–10 minutes for nine-variable problems on an IBM 7044. It would appear that computing times would depend heavily on the type of constraints encountered, but it seems reasonable to be able to solve 20-30-variable problems. Further suggestions for the improvement of this algorithm are to be found in a recently completed study [4].

4. Conclusion. These results represent a first attempt at solving a nonlinear integer programming problem with a moderate number of constraints and, as such, is of interest. They may also lead to better search techniques. The FORTRAN program and the associated flow charts are available.

Acknowledgment. The author wishes to acknowledge the help of his advisor, Professor Leon Cooper. He also wishes to thank the Applied Mathematics section of Monsanto Company for supplying some of the computer time and programming assistance.

REFERENCES

[1] P. KROLAK AND L. COOPER, *An extension of Fibonaccian search to several variables*, Comm. ACM, 6 (1963), pp. 639–641.

[2] P. KROLAK, *A property of the Krolak-Cooper extension of Fibonaccian search*, SIAM Rev., 8 (1966), pp. 510–517.

[3] J. G. JEWELL, *Ordering systems of equations*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 55–71.

[4] P. KROLAK, *The bounded variable algorithm for solving integer linear programming problems*, Tech. rep. COO-1493-18, Sever Institute of Applied Science, Washington University, St. Louis, Missouri, 1967.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

FINITE STATE CONTINUOUS TIME MARKOV DECISION PROCESSES WITH A FINITE PLANNING HORIZON*

BRUCE L. MILLER†

Abstract. The system we consider may be in one of n states at any point in time and its probability law is a Markov process which depends on the policy (control) chosen. The return to the system over a given planning horizon is the integral (over that horizon) of a return rate which depends on both the policy and the sample path of the process. Our objective is to find a policy which maximizes the expected return over the given planning horizon. A necessary and sufficient condition for optimality is obtained, and a constructive proof is given that there is a piecewise constant policy which is optimal. A bound on the number of switches (points where the piecewise constant policy jumps) is obtained for the case where there are two states.

1. Definition of a policy. The system we consider may be in one of n states labeled $1, 2, \cdots, n$, at any point in time. The system operates from time zero to time T, where $T < \infty$. When the system is in state i, an action a is chosen from a finite set A_i of possible actions and a return rate $r(i, a)$ is received which depends only on the current state and action taken. The evolution of the system from state to state is described by a probability law, to be given later, which depends on vectors whose components are $q(j \mid i, a)$, $j = 1, 2, \cdots, n$. These components have the property that $0 \leqq q(j \mid i, a) < \infty$, $j \neq i$, and $\sum_{j=1}^{n} q(j \mid i, a) = 0$. The component $q(j \mid i, a)$, $j \neq i$, can be thought of as the transition rate from state i to state j, i.e., the probability that the system will be in state j at time $t + \Delta t$, $0 \leqq \Delta t < \delta$, $\delta > 0$, given the system is in state i at time t and action $a \in A_i$ is always used in the interval $[t, t + \delta)$ when the system is in state i, is $q(j \mid i, a)\Delta t + o(\Delta t)$.

Let $F = \mathbf{X}_{i=1}^{n} A_i$. A policy π is a function of time on $[0, T]$ into F. Using policy π means that if the system is in state i at time t, the action chosen is $\pi_i(t)$, the ith component of $\pi(t)$. In all cases we require that π be a measurable function where measurable is understood throughout to mean Lebesgue measurable. For any $f \in F$, we let $r(f)$ be the $n \times 1$ column vector whose ith element is $r(i, f_i)$ and $Q(f)$ be the $n \times n$ Markov infinitestimal generator matrix whose (i, j) element is $q(j \mid i, f_i)$.

2. The Markov process. Let S be the state space $\{1, 2, \cdots, n\}$ and Ω be the set of all step functions on $[0, T]$ into S. We write $\mathfrak{F}$ for the σ-algebra of sets in the space Ω generated by the sets $\{\omega: \omega(t) = i\}$, where $\omega \in \Omega$, for

* Received by the editors June 23, 1967, and in revised form October 27, 1967.

† Logistics Department, The RAND Corporation, 1700 Main Street, Santa Monica, California 90406. This work is part of the author's doctoral dissertation in the Operations Research program at Stanford and was supported by the National Science Foundation under Grant GP 3739.

all $t \in [0, T]$ and all i, $1 \leqq i \leqq n$. When the measurable policy π is used and k is the initial state of the system, we define the Markov process of our system by the probability triple $(\Omega, \mathfrak{F}, P^k)$, where P^k is a probability measure such that $\omega(0) = k$, $\omega \in \Omega$, and the probability transition matrix function of the process is the unique absolutely continuous (in t for fixed s) Markov transition matrix function satisfying the condition that for almost all s in $(0, T)$,

$$P(s, t) = I + Q(\pi(s))(t - s) + o(t - s),^1 \tag{1}$$

where $t > s$.

Two questions must be answered in order to justify this definition. The first is whether, given a measurable matrix function $Q(\pi(\cdot))$, there is a unique absolutely continuous Markov transition matrix function which satisfies (1) almost everywhere. The second is whether there is a probability measure P^k such that the process $(\Omega, \mathfrak{F}, P^k)$ has the given absolutely continuous Markov transition matrix function, the point being whether Ω is a large enough class of sample functions.

The first question is answered in [9, Theorem 2.12] by showing that the solutions to the differential equations of the form

$$\frac{d}{dt} P(s, t) = P(s, t)Q(\pi(t)), \tag{2}$$

with the initial condition $P(s, s) = I$ and s assuming all values in $[0, T]$ determine an absolutely continuous Markov transition matrix function satisfying (1) almost everywhere. Uniqueness comes from the requirement that the Markov transition matrix function be absolutely continuous. The second question is answered in the affirmative by using the result of Dynkin [5, p. 160]. Dynkin's result requires only that the given Markov transition matrix function be continuous.

3. The objective function. When $\omega(\cdot)$, $\omega \in \Omega$, is a sample path of our system and the policy π is used, the return to the system is defined to be

$$\int_0^T r(\omega(t), \pi_{\omega(t)}(t))\, dt.$$

Our objective is to choose a measurable policy π which maximizes the expected value of this integral for any initial condition of the system.

It is desirable to interchange the integral and the expectation and we can

[1] The (i, j) element of the matrix $P(s, t)$ is defined as $P^k\{\omega: \omega(s) = i, \omega(t) = j\}/P^k\{\omega: \omega(s) = i\}$, the conditional probability the system is in state j at time t given the system is in state i at time s. This conditional probability is undefined if $P^k\{\omega: \omega(s) = i\} = 0$.

do this from Fubini's theorem if $R(\omega, t) = r(\omega(t), \pi_{\omega(t)}(t))$ is measurable with respect to $(\mathfrak{F} \times M)$, where the M are the Lebesgue measurable sets in $[0, T]$.

We first establish that the stochastic process $(\Omega, \mathfrak{F}, P^k)$, $k \in S$, is measurable with respect to $(\mathfrak{F} \times M)$. Since all of the sample functions of our stochastic process $(\Omega, \mathfrak{F}, P^k)$ are step functions, it follows easily from the definition (Chung [3, p. 143]) that each is separable with respect to any dense set. This implies (Chung [3, Theorem 1, p. 143]) that the process $(\Omega, \mathfrak{F}, P^k)$ is separable with respect to any dense set, hence it is well-separable. Since each sample function has only a finite number of discontinuities, our stochastic process is continuous almost everywhere almost surely. From Doob [4, Theorem 2.5, p. 60] these facts imply the stochastic process $(\Omega, \mathfrak{F}, P^k)$ is measurable with respect to $(\mathfrak{F} \times M)$.

The function $R(\cdot, \cdot)$ is measurable with respect to $(\mathfrak{F} \times M)$, since for each state i and action $a \in A_i$ the set of points such that the system is in state i and action a is used is $\{(\omega, t): \omega(t) = i\} \cap \{(\omega, t): \pi_i(t) = a\}$, which is measurable with respect to $(\mathfrak{F} \times M)$.

Hence we seek the policy π among the class of measurable policies which maximizes the vector

$$v = \int_0^T P(t) r(\pi(t))\, dt \tag{3}$$

in all coordinates (each coordinate corresponding to a different initial state) and this policy is called the optimal policy. Here we have suppressed s (which equals zero) in the notation for the transition matrix.

4. Necessary and sufficient conditions for optimality.

THEOREM 1. *A necessary and sufficient condition for a measurable policy $\pi(\cdot)$ to be optimal is that for almost all $t \in [0, T]$,*

$$r(f) + Q(f)\psi(t) \tag{4}$$

is maximized over the set F by $\pi(t)$, where the column vector $\psi(t)$ is the unique absolutely continuous solution to

$$-\frac{d}{dt}\psi = r(\pi(t)) + Q(\pi(t))\psi, \qquad \psi(T) = 0, \qquad 0 \leqq t \leqq T. \tag{5}$$

Proof. Let π' be any measurable policy. In the following we distinguish between P, Q, v, and r for π and π' by writing, respectively, P, Q, v, r, and P', Q', v', r'. We now establish

$$v - v' = \int_0^T P'[r + Q\psi - r' - Q'\psi]\, dt. \tag{6}$$

To see this we note first that since $P(0) = P'(0) = I$ and $\psi(T) = 0$, we have

$$[P(T) - P'(T)]\psi(T) - [P(0) - P'(0)]\psi(0) = 0.$$

Since $P'(\cdot)$, $P(\cdot)$ and $\psi(\cdot)$ are absolutely continuous the function $[(P(\cdot) - P'(\cdot))\psi(\cdot)]$ is absolutely continuous (Graves [6, Theorem 25, p. 203]) and therefore this function equals the integral of its derivative. Hence,

$$0 = \int_0^T \frac{d}{dt}[(P - P')\psi]\, dt = \int_0^T \left[\frac{d}{dt}(P - P')\psi + (P - P')\frac{d}{dt}\psi\right] dt.$$

By substituting (2) and (5) into this formula we get

$$\int_0^T [(PQ - P'Q')\psi + (P - P')(-(r + Q\psi))]\, dt = \int_0^T [(P' - P)r + P'(Q - Q')\psi]\, dt.$$

Using this fact it follows that

$$v - v' = \int_0^T [Pr - P'r']\, dt = \int_0^T P'[r + Q\psi - r' - Q'\psi]\, dt,$$

which proves (6). Now if π maximizes (4) almost everywhere, the integrand of (6) is nonnegative a.e. so that π is optimal. The necessity also follows from (6). Let π' be a policy which maximizes (4) almost everywhere and assume that π does not. The existence of such a policy π' is not an issue since $\psi(\cdot)$ in (4) depends on $\pi(\cdot)$ which is fixed, but π' is not necessarily measurable since, for example, if two elements of F both maximize (4) over the same set of positive measure, then each element might be chosen on a nonmeasurable subset.

In order to exhibit a measurable policy π' maximizing (4) almost everywhere, it is convenient to enumerate the elements of F as $f(1), f(2), \cdots, f(N)$. For a fixed $f \in F$, $r(f) + Q(f)\psi(t)$ is a continuous and hence measurable function of time so that the sets $T_i' = \{t: f(i)$ maximizes $r(f) + Q(f)\psi(t)$ over $f \in F\}$ are measurable. If we define the mutually exclusive sets T_i by $T_i = T_i' \backslash \bigcup_{j=1}^{i-1} T_i'$, then the policy π' defined by $\pi'(t) = \{f(i): t \in T_i\}$ is both measurable and maximizes (4) almost everywhere. For this policy π' the integrand of (6) is nonpositive and strictly negative on a set of positive measure since $P' \geqq 0$ and the diagonal elements are strictly positive. The strict positivity of the diagonal elements is established in [9, Lemma 2.5] using a nonprobabilistic argument based on the solution of the differential equations (2). It is a well-known result of analysis that if a nonpositive

function on a set is negative on a subset of positive measure, then the integral over the set is negative. Therefore the return using π' is higher than the return from using π which establishes the necessity and completes the proof.

The condition that (4) be maximized a.e. is equivalent to

$$-\frac{d}{dt}\psi = \max_{f \in F} \{r(f) + Q(f)\psi(t)\}, \tag{7}$$

$\psi(T) = 0$. Bellman considers the finite horizon Markov decision problem in [1, Chap. 11]. He defines his system by (7) instead of deriving (7) as an optimality condition.

The sufficiency of maximizing (4) almost everywhere follows as a special case of the 2-person continuous time Markov games results of Zachrisson [11, Theorem 11]. Zachrisson also proved the existence of an admissible (measurable) strategy maximizing (4) everywhere. However, the sufficiency of maximizing (4) everywhere does not follow from the results of Mangasarian [7] since no component of the state differential equations or the objective function is concave in $(P, Q(\pi(\cdot)))$ or $(P, r(\pi(\cdot)))$, respectively.

The necessity part of Theorem 1 follows from [10, Theorem 8], where the functions $\psi(\cdot)$ are the auxiliary functions of the maximum principle.

The solution of (5) can be determined explicitly by elementary methods after noting that (2) and (5) are adjoint systems of differential equations.

THEOREM 2. *For any measurable policy* π,

$$\psi(s) = \int_s^T P(s, t)r(\pi(t))\, dt.$$

From this theorem we have the interpretation of $\psi_i(t)$ as the expected return that will be obtained on the interval $[t, T]$ when policy π is used and the system is in state i at time t.

5. A piecewise constant policy is optimal. We now come to a main result of this paper: a piecewise constant policy is optimal. This result holds for other types of control problems. Pontryagin, Boltyanskii, Gamkrelidze and Mishchenko [10, Chap. 3] prove that a piecewise constant policy is optimal for the linear time-optimal problem. There the differential equations associated with the state variables are linear in x, the state variable vector, in u, the policy, and in (x, u) jointly. In this case, the auxiliary functions are independent of the policy u and are analytic functions. The proof in [10, Theorem 9, p. 117] is based on the auxiliary functions being analytic.

In our problem the differential equations associated with the state variables are linear in P and in $Q(\pi(\cdot))$, but not in $(P, Q(\pi(\cdot))$ jointly. All that can be said about the auxiliary functions is that they are absolutely

continuous, as the solution of a system of differential equations. A useful property possessed by an analytic function but not an absolutely continuous function is that it can have only a finite number of zeros on a finite interval.

Our proof culminates with Theorem 6. Lemma 3 is used in the proofs of Lemmas 4 and 5. We begin with some well-known results for linear differential equations with constant coefficients. Let the vector function $v(\cdot)$ be defined by the following differential equations and terminal conditions:

$$\begin{aligned} v(T) &= c, \\ -\frac{d}{dt}v(t) &= r + Qv(t), \end{aligned} \tag{8}$$

where

$$\begin{aligned} v &\text{ is an } n \times 1 \text{ vector}, \\ Q &\text{ is an } n \times n \text{ matrix}, \\ r &\text{ is an } n \times 1 \text{ vector}, \end{aligned}$$

and $t \in [0, T]$. Then

$$v(t) = ce^{Q(T-t)} + rf(Q, T-t) \quad \text{for} \quad 0 \leqq t \leqq T, \tag{9}$$

where

$$f(Q, t) = t + \frac{t^2}{2!}Q + \frac{t^3}{3!}Q^2 + \cdots$$

satisfies (8) everywhere. Equation (9) can be confirmed by direct differentiation.

We can write $v(t)$ in terms of the derivatives at $t = T$ by direct substitution into (9) so that we have

$$v(t) = v(T) + \sum_{m=1}^{\infty} v^{(m)}(T)\,\frac{(t-T)^m}{m!}. \tag{10}$$

LEMMA 3. *Let $v(\cdot)$ and $v'(\cdot)$ be two $n \times 1$ vector functions defined by the following differential equations and terminal conditions*:

$$\begin{aligned} v(T) &= c, & v'(T) &= c, \\ -\frac{d}{dt}v(t) &= r + Qv(t), & -\frac{d}{dt}v'(t) &= r' + Q'v'(t), \end{aligned}$$

where

$$\begin{aligned} c &\text{ is an } n \times 1 \text{ vector}, \\ Q \text{ and } Q' &\text{ are } n \times n \text{ matrices}, \\ r \text{ and } r' &\text{ are } n \times 1 \text{ vectors}, \end{aligned}$$

and $t \in [0, T]$. *Then if* $v^{(m)}(T) = v'^{(m)}(T)$ *for* $m = 1, 2, \cdots, n + 1$, $v^{(i)}(T) = v'^{(i)}(T)$ *for all positive integers* i *and* $v(t) = v'(t)$ *for* $t \in [0, T]$.

Proof. The n-dimensional vectors, $v^{(i)}(T)$, $i = 1, 2, \cdots, n + 1$, must be linearly dependent. This implies that for some integer j, $2 \leqq j \leqq n + 1$,

$$v^{(j)}(T) = \sum_{k=1}^{j-1} d_k v^{(k)}(T) = v'^{(j)}(T) = \sum_{k=1}^{j-1} d_k v'^{(k)}(T).$$

We show by induction that for all $i \geqq j$,

$$\begin{aligned} v^{(i)}(T) &= \sum_{k=i-j+1}^{i-1} d_{k+j-i} v^{(k)}(T) \\ &= v'^{(i)}(T) \\ &= \sum_{k=i-j+1}^{i-1} d_{k+j-i} v'^{(k)}(T). \end{aligned}$$

Our equations hold for $i = j$ from the above equations. Now we assume that they hold for $i = j, j + 1, \cdots, m - 1$, and show that they hold for $i = m$:

$$\begin{aligned} v^{(m)}(T) &= -Qv^{(m-1)}(T) \\ &= -\sum_{k=m-j}^{m-2} d_{k+j-m+1} Q v^{(k)}(T) \\ &= -\sum_{k=m-j+1}^{m-1} d_{k+j-m} v^{(k)}(T). \end{aligned}$$

In the same way $v'^{(m)}(T) = \sum_{k=m-j+1}^{m-1} d_{k+j-m} v'^{(k)}(T)$. Since $v^{(i)}(T) = v'^{(i)}(T)$ for $i = 1, 2, \cdots, j - 1$, by the hypothesis of the lemma, and for $i = j, j + 1, \cdots, m - 1$, by the induction hypothesis, $v^{(m)}(T) = v'^{(m)}(T)$. From representation (10) this result implies $v'(t) = v(t)$ for all $t \in [0, T]$.

A method for choosing a policy will now be described. It will be shown by Theorem 6 that there is an optimal piecewise constant policy based on this method of choosing a policy.

The intuitive idea of this choice rule is that for each point in time we pick the set of actions which maximizes the first derivative of the vector function ψ and thereby satisfies the optimality conditions at this point. If there is a tie we break the tie by considering the second derivative, etc. Lemma 3 is important because it says we need consider only the first $n + 1$ derivatives.

Given an $n \times 1$ vector function $v(t)$ we can define the $n + 1$ sets:

$$F_1(t) = \{f : f \in F_0(t) = F, \quad f \quad \text{maximizes} \quad v^{(1)}(t, f)\},$$

$$F_2(t) = \{f: f \in F_1(t), \quad f \text{ maximizes } -v^{(2)}(t, f)\},$$

$$\vdots$$

$$F_{n+1}(t) = \{f: f \in F_n(t), \quad f \text{ maximizes } (-1)^n v^{(n+1)}(t, f)\},$$

where

$$\begin{aligned} v^{(1)}(t, f) &= r(f) + Q(f)v(t), \\ v^{(j)}(t, f) &= Q(f)v^{(j-1)}(t) \qquad \text{for } 2 \leqq j \leqq n+1, \\ v^{(j-1)}(t) &= v^{(j-1)}(t, f) \qquad \text{for any } f \in F_{j-1}(t). \end{aligned}$$

In order to ensure uniqueness for our selection procedure it is necessary to enumerate the finite set F as

$$f(1), f(2), \cdots, f(N). \tag{11}$$

We say that f satisfies the selection procedure based on v at time t if f is the element in $F_{n+1}(t)$ with the lowest index according to the enumeration (11). Usually we shall be concerned with the case $v = \psi$, where ψ corresponds to some policy π and is defined by (5).

LEMMA 4. *Consider the arbitrary measurable policy π defined on the interval $(t', T]$. The corresponding vector function ψ is therefore defined on $[t', T]$ (the interval is closed since $\psi(\cdot)$ is continuous). Let f^* be the vector of actions picked by the selection procedure based on $\psi(t')$ and set π equal to f^* on some interval $[t'', t']$, $t'' < t'$. The vector function ψ is also now defined on $[t'', t')$. Then π satisfies the selection procedure based on ψ for $t \in [t' - \epsilon, t']$ for some ϵ, where $0 < \epsilon < t' - t''$.*

Proof. The first half of the proof consists of showing that if $f' \in F \backslash F_{n+1}(t')$, then $f' \notin F_1(t)$ for any $t \in (t' - \epsilon(f'), t')$, $0 < \epsilon(f') < t' - t''$, and hence is not chosen by the selection procedure on that interval.

Since a constant policy is used for $t'' < t < t'$, the vector function $\psi(\cdot)$ is of the form (9), where $T = t'$, $c = \psi(t')$, $Q = Q(f^*)$ and $r = r(f^*)$, and therefore is infinitely differentiable so that for any l, any $f \in F$, $t'' \leqq t \leqq t'$, we can write the Taylor's expansion of $r(f) + Q(f)\psi(t)$:

$$\begin{aligned} r(f) + Q(f)\psi(t) &= r(f) + Q(f)\psi(t') \\ &+ \sum_{k=1}^{l} Q(f)\psi^{(k)}(t') \frac{(t - t')^k}{k!} + Q(f)\psi^{(l+1)}(t_f) \frac{(t - t')^{l+1}}{(l+1)!}, \end{aligned} \tag{12}$$

where $t \leqq t_f \leqq t'$. Let l be the largest integer such that $f' \in F_l(t')$ (where $0 \leqq l \leqq n$, since $f' \in F \backslash F_{n+1}(t)$). If $l > 0$ the value of

$$r(f) + Q(f)\psi(t') + \sum_{k=1}^{l-1} Q(f)\psi^{(k)}(t') \frac{(t - t')^k}{k!}$$

is equal for both f' and f^* since both are elements of the sets $F_0(t')$, $F_1(t'), \cdots, F_l(t')$. Since $f' \notin F_{l+1}(t')$ and $f^* \in F_{l+1}(t')$, the value of $(-1)^l Q(f)\psi^{(l)}(t')$ is strictly greater in some coordinate for $f = f^*$ than $f = f'$ and we let this vector difference be $\delta f'$. When $l = 0$ the value of $r(f) + Q(f)\psi(t')$ is strictly greater in some coordinate for $f = f^*$ than $f = f'$ and we let $\delta f'$ be this vector difference. The vector function $\psi^{(l+1)}(\cdot)$ is uniformly bounded in t, $t'' \leqq t \leqq t'$, for fixed l so that there is an $\epsilon(f') > 0$ such that

$$\frac{\delta f'}{(-1)^l} \frac{(t - t')^l}{l!}$$

is strictly greater for all $t, t' - \epsilon(f') \leqq t \leqq t'$, in some coordinate than

$$\{Q(f')\psi^{(l+1)}(t_{f'}) - Q(f^*)\psi^{(l+1)}(t_{f^*})\} \frac{(t - t')^{l+1}}{(l+1)!},$$

where $t \leqq t_{f'}, t_{f^*} \leqq t'$. This implies, using the representation (12), that $r(f^*) + Q(f^*)\psi(t)$ is greater in some coordinate than $r(f') + Q(f')\psi(t)$ for all $t \in (t' - \epsilon(f'), t')$, which proves $f' \notin F_1(t)$ for $t \in (t' - \epsilon(f'), t')$. The ϵ of our lemma is

$$\min_{f' \in F \setminus F_{n+1}(t')} \{\epsilon(f')\},$$

which is strictly positive since F is finite.

We now consider $f' \in F_{n+1}(t')$, $f' \neq f^*$, and the differential equations

$$-\frac{d}{dt}\psi(t) = r(f^*) + Q(f^*)\psi(t) \quad \text{for} \quad t'' \leqq t \leqq t',$$

$$\psi'(t') = \psi(t'),$$

$$-\frac{d}{dt}\psi'(t) = r(f') + Q(f')\psi'(t) \quad \text{for} \quad t'' \leqq t \leqq t'.$$

Since both f' and $f^* \in F_{n+1}(t')$, the first $n + 1$ derivatives of ψ' and ψ are equal at t' and Lemma 3 applies so that $\psi'(t) = \psi(t)$ for $t'' \leqq t \leqq t'$.

This implies for all $t \in (t'', t')$ that $r(f^*) + Q(f^*)\psi(t) = \psi^{(1)}(t) = \psi'^{(1)}(t) = r(f') + Q(f')\psi'(t) = r(f') + Q(f')\psi(t)$, and for $1 \leqq k \leqq n$, $-Q(f^*)\psi^{(k)}(t) = \psi^{(k+1)}(t) = \psi'^{(k+1)}(t) = -Q(f')\psi'^{(k)}(t) = -Q(f')\psi^{(k)}(t)$. But this means the set $F_{n+1}(t)$ is constant for $t \in (t' - \epsilon, t')$. Therefore, f^* satisfies the selection procedure based on ψ for $t \in (t' - \epsilon, t')$, which proves the lemma.

In order to establish Lemma 5 it is necessary to consider a revised selection procedure. The revised selection procedure is used only to prove Lemma 5 and the original selection procedure will continue to be called the selection procedure.

Given an $n \times 1$ vector function $v(t)$ we can define the $n + 1$ sets:

$$\mathbf{F}_1(t) = \{f: f \in F, f \text{ maximizes } \mathbf{v}^{(1)}(t, f)\},$$

$$\mathbf{F}_2(t) = \{f: f \in \mathbf{F}_1(t), f \text{ maximizes } \mathbf{v}^{(2)}(t, f)\},$$

$$\vdots$$

$$\mathbf{F}_{n+1}(t) = \{f: f \in \mathbf{F}_n(t), f \text{ maximizes } \mathbf{v}^{(n+1)}(t, f)\},$$

where $\mathbf{v}^{(1)}(t, f) = r(f) + Q(f)v(t)$, $\mathbf{v}^{(j)}(t, f) = Q(f)\mathbf{v}^{(j-1)}(t)$, $2 \leqq j \leqq n + 1$, and $\mathbf{v}^{(j-1)}(t) = \mathbf{v}^{(j-1)}(t, f)$ for any $f \in \mathbf{F}_{j-1}(t)$. We enumerate the set F as before by (11). We say that f satisfies the revised selection procedure based on v at time t if f is the element in $\mathbf{F}_{n+1}(t)$ with the lowest index according to the enumeration (11).

LEMMA 5. *Let π be a policy which satisfies the selection procedure (original) based on its corresponding vector function ψ for $t \in (t', T]$, $t' < T$. Then π must be constant on the interval $(t', t' + \epsilon)$ for some $\epsilon > 0$.*

Proof. We shall establish Lemma 5 by exhibiting the $f \in F$ such that $\pi(t) = f$ on the interval $(t', t' + \epsilon)$ for some $\epsilon > 0$. Since $\psi(\cdot)$ is continuous it is defined also at t', and we let f^* be the unique element of F chosen by the revised selection procedure based on $\psi(t')$. Now consider the vector function $v(\cdot)$ defined by the differential equations

$$v(t') = \psi(t'),$$

$$-\frac{d}{dt}v(t) = r(f^*) + Q(f^*)v(t)$$

for $t' \leqq t \leqq T$.

We now show f^* satisfies both the revised and original selection procedure based on v for $t \in (t', t' + \epsilon)$ and some $\epsilon > 0$ (even though f^* does not necessarily satisfy the selection procedure (original) at t'). Using the same argument as in Lemma 4, if $f \notin \mathbf{F}_{n+1}(t')$, then $f \notin \mathbf{F}_1(t)$ for $t \in (t', t' + \epsilon)$ and some $\epsilon > 0$, and if $f \in \mathbf{F}_{n+1}(t')$, then $f \in \mathbf{F}_{n+1}(t)$ for $t \in (t', t' + \epsilon)$ so that f^* satisfies the revised selection procedure on this interval. Also for $t \in (t', t' + \epsilon)$ either $f \notin \mathbf{F}_1(t)$ or $f \in \mathbf{F}_{n+1}(t)$ so that $\mathbf{F}_1(t) = \mathbf{F}_2(t) = \cdots = \mathbf{F}_{n+1}(t)$. We now prove by induction on i that $\mathbf{F}_i(t) = F_i(t)$ for $1 \leqq i \leqq n + 1$ and $t \in (t', t' + \epsilon)$. $\mathbf{F}_1(t) = F_1(t)$ since both sets are defined identically. Now we assume $\mathbf{F}_i(t) = F_i(t)$ for $i = 1, 2, \cdots, l - 1$, and show it holds for $i = l$. The relation $v^{(1)}(\cdot) = \mathbf{v}^{(1)}(\cdot)$ on $(t', t' + \epsilon)$ and the induction hypothesis imply $v^{(l-1)}(\cdot) = \mathbf{v}^{(l-1)}(\cdot)$ on $(t', t' + \epsilon)$. Since $\mathbf{F}_l(t) = \mathbf{F}_{l-1}(t)$, the value of $\{Q(f)\mathbf{v}^{(l-1)}(t)\}$ must be the same for all $f \in \mathbf{F}_{l-1}(t)$ which implies the value of $\{(-1)^{l-1}Q(f)\mathbf{v}^{(l-1)}(t)\}$ must be the same for all $f \in \mathbf{F}_{l-1}(t)$. Therefore, since $F_{l-1}(t) = \mathbf{F}_{l-1}(t)$ and $v^{(l-1)}(t) = \mathbf{v}^{(l-1)}(t)$, $\{(-1)^{l-1}Q(f)v^{(l-1)}(t)\}$ must have the same value for all

$f \in F_{l-1}(t)$ so that $F_l(t) = F_{l-1}(t) = \mathbf{F}_{l-1}(t) = \mathbf{F}_l(t)$. Hence we can conclude that $F_{n+1}(t) = \mathbf{F}_{n+1}(t)$ for $t \in (t', t' + \epsilon)$ and since f^* satisfies the revised selection procedure on that interval, it satisfies the selection procedure (original) on that interval. We shall have shown that $\pi(t) = f^*$ on $(t', t' + \epsilon)$ if we can prove that $v(\cdot) = \psi(\cdot)$ on this interval because the selection procedure is unique.

Since $f^* \in \mathbf{F}_1(t)$ for $t \in (t', t' + \epsilon)$, the differential equations defining $v(\cdot)$ can be written

$$v(t') = \psi(t'),$$

$$-\frac{d}{dt}v(t) = \max_{f \in F}(r(f) + Q(f)v(t))$$

for $t \in (t', t' + \epsilon)$. But by hypothesis π satisfies the selection procedure (original) on (t', T) so that $\pi(t) \in F_1(t)$, $F_1(t)$ being based on $\psi(\cdot)$, and $\psi(\cdot)$ also satisfies the above differential equations. The uniqueness of the solution of these differential equations (see [1, Theorem 1, p. 321]) implies $\psi = v$ on $(t', t' + \epsilon)$ and completes the proof.

THEOREM 6. *There is a piecewise constant (from the left) policy π defined on $[0, T]$ which maximizes (4) everywhere. This policy is optimal.*

Proof. The proof is by contruction. Consider the following algorithm which goes through the steps (13)–(17) consecutively:

(13) Initialization; set $t' = T$ and $\psi(T) = 0$.

(14) Use the selection procedure based on $\psi(t')$ to determine $\pi(t')$.

(15) Obtain $\psi(t)$ for $0 \leqq t \leqq t'$ by solving the differential equation

$$-\frac{d}{dt}\psi(t) = r(\pi(t')) + Q(\pi(t'))\psi(t),$$

using the previous value of $\psi(t')$ as the terminal condition.

(16) Set $t'' = \inf\{t: \pi(t')$ satisfies the selection procedure on the interval (t, t') based on the vector function $\psi(t)\}$.

(17) If $t'' \leqq 0$ terminate; if $t'' > 0$ go to step (14) with $t' = t''$.

Because of condition (16) the policy π satisfies the selection procedure corresponding to π everywhere and hence always lies in the set $F_1(t)$ for all t. This condition is equivalent to maximizing (4) for all t so that (Theorem 1) π is optimal where it is defined. It remains to be shown that there is a finite number of switches when this algorithm is used, that the algorithm goes through steps (13) – (17) a finite number of times. From step (16) and Lemma 4 we note that the points t_i', corresponding to the value of t' in the algorithm at the ith iteration, are strictly decreasing. Suppose the algorithm is not finite. Let $t^* = \inf\{t_i\}$. Then the policy

$\pi(\cdot)$ defined by the algorithm on the half open interval $(t^*, T]$ satisfies the hypotheses of Lemma 5. Thus there is an $\epsilon > 0$ such that $\pi(\cdot)$ is constant on $(t^*, t^* + \epsilon)$, which contradicts the fact that infinitely many t_i lie in $(t^*, t^* + \epsilon)$. Therefore the algorithm must terminate in finitely many steps, which completes the proof.

It should be pointed out that the implied algorithm is not practical because of the difficulty in carrying out step (16). In [9, Chap. 4] an algorithm is presented which uses a fixed interval for the length of time the newly obtained vector of actions $\pi(t')$ holds instead of the length of time determined by step (16). It is shown that by making these intervals small enough one can obtain a policy which yields an expected return within ϵ of the optimal policy. This result was obtained independently by Martin-Lof [8] in a recent paper.

6. A bound on the number of switches when there are two states and both communicate. In Theorem 6 it was shown that there is an optimal policy for the finite horizon problem which has only a finite number of switches. In the case where there are two communicating states we show (Corollary 12) that the number of switches of the policy obtained using the algorithm of Theorem 6 is bounded by the number of elements of F. Whether such a result holds for the case $n > 2$ is an open question.

The hypothesis that both states communicate is ambiguous since two states may communicate for one element of F but not for another. Here we say both states communicate if there is one element f in F such that both states communicate. This is enough to ensure that there is a y such that the equation

$$(18)\quad \max_{a \in A_1} (r(1, a) + q(2 \mid 1, a)y) = \max_{a \in A_2} (r(2, a) - q(1 \mid 2, a)y)$$

holds since the right (left) side of (18) goes to $+\infty$ as y goes to $-\infty$ $(+\infty)$ and is a continuous monotone decreasing (increasing) function of y. In [9, Chap. 5] the constant y is shown to be the expected advantage from being in state 2 than in being in state 1 when the horizon is infinite and the discount factor goes to zero.

From Theorem 2 we have the interpretation of $\psi_2(t) - \psi_1(t)$ as the expected advantage from being in state 2 than from being in state 1 at time t in the finite horizon problem. We shall show that $\psi_2(\cdot) - \psi_1(\cdot)$ moves monotonically toward the steady state value y as t goes from T to 0. From the monotoneity of $\psi_2(\cdot) - \psi_1(\cdot)$ we can establish our desired result that if a decision vector f is optimal at time t_1 and not optimal at time t_2, $t_2 > t_1$, then f is not optimal for any time t, $t > t_2$. In summary, we assume that $n = 2$, both states communicate and y refers to the solution

of (18). We let ψ_1 and ψ_2 be the auxiliary functions associated with the optimal policy π obtained using the algorithm of Theorem 6.

LEMMA 7. *For all* $t \in [0, T]$,

$$\left(\frac{d}{dt}\psi_2(t) - \frac{d}{dt}\psi_1(t)\right)\begin{cases}\leqq 0 & \text{if } \psi_2(t) - \psi_1(t) \leqq y,\\ \geqq 0 & \text{if } \psi_2(t) - \psi_1(t) \geqq y.\end{cases}$$

Proof. The lemma is proved for the case $y \geqq \psi_2(t) - \psi_1(t)$ only, since both proofs are the same. Since π satisfies the optimality condition everywhere, (5) is satisfied everywhere (the right-hand side is continuous) and

$$-\frac{d}{dt}\psi_2(t) = \max_{a \in A_2}\left(r(2, a) - q(1 \mid 2, a)(\psi_2(t) - \psi_1(t))\right)$$

everywhere. By hypothesis $y \geqq \psi_2(t) - \psi_1(t)$ so that

$$\begin{aligned}-\frac{d}{dt}\psi_2(t) &\geqq \max_{a \in A_2}\left(r(2, a) - q(1 \mid 2, a)y\right)\\ &= \max_{a \in A_1}\left(r(1, a) + q(2 \mid 1, a)y\right)\\ &\geqq \max_{a \in A_1}\left(r(1, a) + q(2 \mid 1, a)(\psi_2(t) - \psi_1(t))\right)\\ &= -\frac{d}{dt}\psi_1(t),\end{aligned}$$

which completes the proof.

LEMMA 8. *If* f *has a continuous first derivative which is nonpositive whenever* $f(t) < 0$ *for* $t \in (a, b)$, *then* $f(b) \geqq 0$ *implies* $f(r) \geqq 0$ *for* $t \in [a, b]$.

Proof. Assume the contrary and apply the mean value theorem.

LEMMA 9. *For* $t \in [0, T]$,

$$\psi_2(t) - \psi_1(t) + y \geqq 0 \quad \text{if} \quad y \geqq 0,$$

$$\psi_2(t) - \psi_1(t) - y \geqq 0 \quad \text{if} \quad y \leqq 0.$$

Proof. The function $\psi_2(t) - \psi_1(t) + y$ of t satisfies the hypothesis of Lemma 8 using Lemma 7 and is therefore nonnegative. The same thing applies to the function $\psi_2(t) - \psi_1(t) - y$.

LEMMA 10. *The function* $\psi_2(t) - \psi_1(t)$ *is monotone on* $[0, T]$.

Proof. If $y \geqq 0$, then from Lemma 9, $\psi_2(t) - \psi_1(t) \leqq y$ for all $t \in [0, T]$ which implies $(d/dt)\psi_2(t) - (d/dt)\psi_1(t) \leqq 0$ using Lemma 7. In the same way $\psi_2(t) - \psi_1(t)$ is monotone when $y \leqq 0$.

THEOREM 11. *Let* $\pi(i)$ *be the value of* π *obtained using the algorithm defined in the proof of Theorem* 6 *at iteration* i. *Then if* $i > 1$, $\pi(i) \neq \pi(m)$ *for all* $m < i$.

Proof. For the algorithm of Theorem 6 to switch from $\pi(i-1)$ to $\pi(i)$ at some time t_i we must have for some state j and some derivative k, $1 \leqq k \leqq n+1$, $(-1)^{k-1}\psi_j^{(k)}(t_i, f)$ strictly greater for $f = \pi(i)$ than $f = \pi(i-1)$ and equal for all derivatives l, $0 \leqq l < k$. To be specific, let $j = 2$ so that we have the strict inequalities

$$r(2, \pi_2(i)) - q(1 \mid 2, \pi_2(i))(\psi_2(t_i -) - \psi_1(t_i -)) > r(2, \pi_2(i-1)) - q(1 \mid 2, \pi_2(i-1))(\psi_2(t_i -) - \psi_1(t_i -))$$

and

$$r(2, \pi_2(i)) - q(1 \mid 2, \pi_2(i))(\psi_2(t_i +) - \psi_1(t_i +)) < r(2, \pi_2(i-1)) - q(1 \mid 2, \pi_2(i-1))(\psi_2(t_i +) - \psi_1(t_i +)).$$

The function $\psi_2(t) - \psi_1(t)$ is monotone by Lemma 10. This fact and the two previous inequalities imply that

$$r(2, \pi_2(i-1)) - q(1 \mid 2, \pi_2(i-1))(\psi_2(t) - \psi_1(t)) > r(2, \pi_2(i)) - q(1 \mid 2, \pi_2(i))(\psi_2(t) - \psi_1(t))$$

for all $t \in (t_i, T]$. Hence $\pi_2(i) \neq \pi_2(m)$ for all $m < i$.

COROLLARY 12. *An upper bound on the number of switches is F, the range of the function π.*

It is of interest that no such result holds in the discrete time Markov decision problem. Brown [2, p. 1282] gives an example of a two-state finite horizon problem where the optimal policy changes back and forth between two elements of F.

7. Acknowledgment. It is a pleasure to thank my adviser, Professor Arthur F. Veinott, Jr., for his close examination of my work which resulted in the strengthening of the problem formulation and the revision of many of the proofs. I also wish to thank the referee for helpful comments on §2 and §3.

REFERENCES

[1] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[2] B. BROWN, *On the iterative method of dynamic programming on a finite space discrete time Markov process*, Ann. Math. Statist., 36 (1965), pp. 1279–1285.

[3] K. L. CHUNG, *Markov Chains with Stationary Transition Probabilities*, 2nd ed., Springer, Berlin, 1967.

[4] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[5] E. B. DYNKIN, *Theory of Markov Processes*, Prentice-Hall, Englewood Cliffs, New Jersey, 1961.

[6] L. GRAVES, *The Theory of Functions of Real Variables*, McGraw-Hill, New York, 1956.

[7] O. L. MANGASARIAN, *Sufficient conditions for the optimal control of nonlinear systems*, this Journal, 4 (1966), pp. 139–152.

[8] A. MARTIN-LOF, *Controlled Markov chains with periodic transition probabilities*, Operations Research, 15 (1967), pp. 872–881.

[9] B. L. MILLER, *Finite state continuous time Markov decision processes with applications to a class of optimization problems in queueing theory*, Doctoral dissertation, Department of Operations Research, Stanford University, Stanford, California, 1967.

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

[11] E. ZACHRISSON, *Markov games*, Advances In Game Theory, M. Dresher, L. S. Shapley and A. W. Tucker, eds., Princeton University Press, Princeton, 1964, pp. 211–254.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

A CLASS OF NONSTANDARD OPTIMAL CONTROL PROBLEMS WITH APPLICATION TO NUCLEAR REACTOR ECONOMICS*

PAUL NELSON, JR.† AND GALE YOUNG‡

1. Introduction. The problem to be studied herein is that of minimizing the functional

$$J(x_0, u) = Cx_0 + \int_0^T k(t)u(t)\,dt \tag{1}$$

subject to the constraint

$$0 \leqq u(t) \leqq \min\{x(t) - r(t), h(t)\}, \qquad 0 \leqq t \leqq T, \tag{2}$$

where the function x is defined on the nontrivial interval $[0, T]$ by the initial value problem

$$x'(t) = g(t) + b(t)u(t), \qquad x(0) = x_0, \tag{3}$$

the functions k, r, h, g and b are given on $[0, T]$, C is a nonnegative constant, and x_0 is constrained to lie in the given interval I, $I = [q_1, q_2]$ or $I = [q_1, \infty)$ for some nonnegative real numbers q_1, q_2. The present interest in such problems arose in connection with a study of economic models of nuclear reactor power systems, and the primary objective of the present work is to formulate a computationally effective algorithm which yields a solution of the problem under conditions on the data which are satisfied in most such models. It seems likely that the results presented will have application to quite general economic models, but our examples will be restricted to models drawn from reactor economics. The algorithm we present accomplishes the above objective, and furthermore, gives a clear picture of the structure of the solution and yields the unique solution to the problem under only slightly specialized hypotheses. The problem is essentially reduced to that of solving a collection of initial value problems for a certain first order nonlinear ordinary differential equation. Problems of the latter type are, of course, quite easily solved computationally.

Throughout this work we shall make the following assumptions, in addition to those described above, except where otherwise noted:

(H1) The functions b and k are real-valued, bounded and measurable on $[0, T]$ with b positive, k nonnegative and k/b nonincreasing.

* Received by the editors September 11, 1967, and in revised form January 2, 1968. This research was supported by the United States Atomic Energy Commission under contract with the Union Carbide Corporation.

† Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830. Presently on leave of absence at Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87106.

‡ Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830.

(H2) The function h has values in the extended nonnegative real numbers and is measurable on $[0, T]$.

(H3) The function r is real-valued, piecewise constant and everywhere continuous from the right in $[0, T]$.

(H4) The function g is Lebesgue integrable on $[0, T]$ and is such that $g^{-1}[(-\infty, 0)]$ has finitely many topological components (i.e., maximal connected subsets).

In the hypotheses and elsewhere, measurable means Lebesgue measurable. The requirements of measurability and integrability in these hypotheses are technical in nature. The assumptions regarding sign and monotonicity in (H1) are reasonable in economic problems and are essential for our results. The hypothesis that r is continuous from the right is a normalizing assumption which is not really essential to the results, but is merely convenient to avoid complicating the notation. The requirements that r be piecewise constant and $g^{-1}[(-\infty, 0)]$ have a finite number of components can probably be relaxed somewhat, but the present assumptions seem adequate for problems of practical interest. The hypothesis (H4) is satisfied, for example, if g is piecewise continuous and its graph crosses the axis only finitely many times (i.e., $g^{-1}(0)$ has a finite number of components).

The problem of minimizing (1) subject to the constraints (2) and (3) has the appearance of an optimal control problem with state variable x, control variable u, control parameter x_0 and the rather unusual feature that the control constraints (2) depend on the state variable. A pair (x_0, u) such that $x_0 \in I$, u is a nonnegative measurable function on $[0, T]$, and the constraints (2) are satisfied with x given by the *generalized state equation*

$$(3') \qquad x(t) = x_0 + \int_0^t g(\tau)\, d\tau + \int_0^t b(\tau)u(\tau)\, d\tau, \qquad 0 \leqq t \leqq T,$$

will be called an *admissible control*, and an admissible control which minimizes the functional J in the class of admissible controls is an *optimal control*.

We note that the problem of minimizing (1) subject to (2) and (3) is transformed into an equivalent problem of the same type if x is replaced by $x - G$, G an absolutely continuous function. For the transformed problem the state variable is $y = x - G$, g is replaced by $g - G'$, r by $r + G$, and the remaining data are unchanged. It is obvious that (H1) and (H2) are satisfied by the transformed problem if they are satisfied by the original problem. This shows that the results obtained under the hypotheses (H1)–(H4) actually can be applied to seemingly more general problems of the form (1)–(3) in which (H1) and (H2) are satisfied, and r can be decomposed into the sum of a piecewise constant function and an absolutely continuous function G such that $g + G'$ satisfies (H4). This observation will be quite useful in the applications.

Optimal control problems with constraints which depend on the state variable have been studied by several authors. Cesari [4] gives an existence theorem which is sufficiently general to apply to the present problem. Berkovitz [2] extended the maximum principle to problems with constraints containing both state and control variables, with essentially the restriction that each constraint either contain the control variable explicitly or be defined by a strict inequality. He later [3] generalized these results to include problems with one constraint containing only the state variable and defined by a weak inequality. Hestenes [6, Chap. 7] gives a form of the maximum principle for problems which he calls the *general control problem of Mayer*, his treatment being essentially limited to the same restrictions as the first mentioned article of Berkovitz. Guinn [5] has studied the possibility of extending Hestenes' results to include problems with an arbitrary number of state space constraints. None of these extensions of the maximum principle applies to the present problem because equality holds on both sides of the constraint (2) at any point of the form $(t, x, u) = (t, r(t), 0)$, and arcs through such "singular" points are excluded from consideration in all of the works cited above. Particularly, in regard to Hestenes' work, there is no *program* [6, p. 303] associated with any such point. One of our main results (Theorem 6) is to the effect that, under certain conditions, the optimal control necessarily contains such singular points.

The problem can be given an equivalent formulation without explicit mention of the variable x by using (3′) to replace x in the inequalities (2). For the special case $h = +\infty$ and I a singleton set $\{x_0\}$, this leads to a continuous linear programming problem as studied by Tyndall [11] and by Levinson [7]. Tyndall's existence theorem applies only to the special case in which b and r are constant and

$$x_0 + \int_0^t g(\tau)\, d\tau + r(t)$$

is nonnegative for $t \in [0, T]$. The existence result of Levinson applies to the special case described above under the hypotheses (H1)–(H3), with (H4) replaced by the requirement that g be merely Lebesgue integrable on $[0, T]$. Thus Levinson's theory gives a more general existence theorem for the continuous linear programming case than results from the construction we describe. Tyndall [12] has recently extended his results in such a manner as to avoid the restriction to constant b and r.

Bellman, Fleming and Widder [1] consider a problem with constraints of the type (2) with $r \equiv 0$, $h \equiv +\infty$, general integral cost functional (1) and autonomous nonlinear state equation (3). Their results on the structure of the solution are perhaps nearer in spirit to those of this article than are the results of any other reference.

A maximum principle would not be particularly interesting for problems of the exact type we consider, since the solution is described by our algorithm in considerably more detail than would be afforded by the first order necessary conditions of such a principle. However, an extension of the maximum principle to include problems with constraints of the type (2) might be of interest for more general state equations. In particular, we note that the construction we describe yields an optimal control which is a "bang-bang" control in the sense that equality holds on one side of the inequality (2) at all times. It would be interesting to know whether this result holds for general linear state equations.

In §2 we find necessary and sufficient conditions for existence of admissible controls. Section 3 contains a description and proof of effectiveness of the algorithm for the special case $k/b \leqq C$, and §4 contains the extension to the general case. Section 5 is given over to a proof that the optimal control constructed in §3 and §4 is the unique optimal control if k/b is actually decreasing rather than merely nonincreasing. Finally, §6 and §7 contain examples from nuclear reactor economics.

2. Existence of admissible controls. The main objective of this section is to give a necessary and sufficient condition on the data of the problem for existence of admissible controls (Theorem 1). First we give a few preliminary definitions and prove a lemma which will be used frequently.

If S_1, S_2 are measurable real-valued functions with the same domain, then $S_1 \leqq S_2$ means $S_1(t) \leqq S_2(t)$ for almost all t in the common domain of the S_i, and $S_1 < S_2$ means $S_1(t) < S_2(t)$ for almost all such t. The meaning of the symbol $\geqq$ as a relation between measurable functions is now clear. If a is a constant, then a will also denote the function whose value is everywhere a and all restrictions of this function. Addition of real-valued functions with a common domain is defined as usual. The class of Lebesgue summable functions on the measurable subset A of the real line will be denoted by $\mathfrak{L}^1(A)$, and $L^1(A)$ is the set of equivalence classes of elements of $\mathfrak{L}^1(A)$ under the equivalence relation of equality almost everywhere.

The real-valued Lebesgue measurable function f defined on the interval $[0, t_0]$ will be called a *solution* on $[0, t_0]$ of the nonlinear Volterra integral equation

$$w(t) = S(t) + \int_0^t b(\tau) \cdot \min \{w(\tau) - r(\tau), h(\tau)\}\, d\tau \tag{4}$$

in the unknown w, where b is as described in (H1) and S is a given measurable function on $[0, t_0]$, if (4) holds for *every* $t \in [0, t_0]$ when $w = f$. The integral in (4) is to be taken as a Lebesgue integral, and in particular this

integral exists for every $t \in [0, t_0]$ if $w = f$ and f is a solution of (4) on $[0, t_0]$. A similar definition of solution will be taken for integral equations of the type (4) over more general intervals.

LEMMA 1. *If* $S \in \mathfrak{L}^1([0, T])$, *then the integral equation* (4) *has a unique solution* $\mathbf{W}S$ *in* $\mathfrak{L}^1([0, T])$. *Furthermore, if* $S_1, S_2 \in \mathfrak{L}^1([0, T])$ *and* $S_1 \leqq S_2$ (*respectively* $S_1 < S_2$), *then* $\mathbf{W}S_1 \leqq \mathbf{W}S_2$ (*respectively* $\mathbf{W}S_1 < \mathbf{W}S_2$).

Proof. For given real a, measurable and essentially bounded B on $[0, T]$, and functions $S, w \in \mathfrak{L}^1([0, T])$, we denote by $\mathbf{K}(S)w$ the function whose value at $t \in [0, T]$ is given by the right side of (4). It is obvious that $\mathbf{K}(S)$ is an operator from $\mathfrak{L}^1([0, T])$ into itself and induces, in a natural manner, operators on $\mathfrak{L}^1([0, \tau])$ and $L^1([0, \tau])$ into themselves for $\tau \leqq T$, which latter operators will also be called $\mathbf{K}(S)$. Since ϕ: $(x, t) \to \min \{x - r(t), h(t)\}$ satisfies a Lipschitz condition in x with unit Lipschitz constant uniformly in $t \in [0, T]$, it is true that $\mathbf{K}(S)$ is a contractive mapping of $L^1([0, \tau])$ into itself for $\tau < 1/b_0$ where $b_0 = \operatorname{ess\,sup} \{b(t) \mid 0 \leqq t \leqq T\}$ and we use the usual L^1-norm. It follows easily from the contractive mapping theorem and the requirement that the equality (4) hold for all t in $[0, \tau]$ that (4) has exactly one solution in $\mathfrak{L}^1([0, \tau])$. By a standard argument the solution can be successively extended uniquely over intervals of length τ until it covers all of $[0, T]$.

The contractive mapping theorem also implies that the sequence $\{K^n(S)w\}$ converges to $\mathbf{W}S$ relative to the $\mathfrak{L}^1([0, \tau])$-seminorm as $n \to \infty$ for arbitrary $w \in \mathfrak{L}^1([0, \tau])$, where $\mathbf{K}^n(S)$ is $\mathbf{K}(S)$ iterated n times. Obviously $S_1 \leqq S_2$ and $w_1 \leqq w_2$ imply $\mathbf{K}(S_1)w_1 \leqq \mathbf{K}(S_2)w_2$. A simple inductive proof then shows that $S_1 \leqq S_2$ implies $\mathbf{K}^n(S_1)w \leqq \mathbf{K}^n(S_2)w$ for arbitrary $w \in \mathfrak{L}^1([0, T])$ and natural number n. But the quasi-order $\leqq$ is preserved under limits with respect to the $\mathfrak{L}^1$-seminorm, and therefore $\mathbf{W}S_1(t) \leqq \mathbf{W}S_2(t)$ for almost all t in $[0, \tau]$. On $[\tau, 2\tau]$ the solution $\mathbf{W}S$ of (4) is also a solution of the integral equation

$$w(t) = \tilde{S}(t) + \int_\tau^t b(s) \cdot \min \{w(s) - r(s), h(s)\}\, ds$$

in the unknown w, where

$$\tilde{S}(t) = S(t) + \int_0^\tau b(s) \cdot \min \{\mathbf{W}S(s) - r(s), h(s)\}\, ds.$$

Then obviously $S_1 \leqq S_2$ implies $\tilde{S}_1 \leqq \tilde{S}_2$ and, repeating the argument given above for $[0, \tau]$, we find $S_1 \leqq S_2$ implies $\mathbf{W}S_1(t) \leqq \mathbf{W}S_2(t)$ for almost all t in $[\tau, 2\tau]$. This result also may be successively extended over intervals of length τ until all of $[0, T]$ is covered.

Now suppose $S_1 < S_2$, ϕ is nondecreasing in x for fixed t, and let w_i

$= \mathbf{W}S_i$. Then the result just proved shows that

$$\int_0^t b(\tau)\cdot\min\{w_1(\tau) - r(\tau), h(\tau)\}\,d\tau \leqq \int_0^t b(\tau)\cdot\min\{w_2(\tau) - r(\tau), h(\tau)\}\,d\tau$$

for all t in $[0, T]$. Combining this with $S_1 < S_2$ and the fact that the w_i are solutions of (4), we see that $w_1 < w_2$. This completes the proof of Lemma 1.

Remark. Results similar to Lemma 1 hold for the lower limit of integration in (4) replaced by arbitrary $t_0 \in [0, T]$. In particular, if S_1, $S_2 \in \mathfrak{L}^1([0, T])$, $S_1 \leqq S_2$ on $[t_0, T]$ (respectively $[0, t_0]$) and w_1, w_2 are the corresponding solutions of (4) with the lower limit of integration t_0 (respectively upper limit t_0 and lower limit t), then $w_1 \leqq w_2$ on $[t_0, T]$ (respectively $[0, t_0]$) with, furthermore, strict inequality for almost all such t if strict inequality holds for the S_i. These results will be used later in this article.

THEOREM 1. *If $I = [q_1, q_2]$ for nonnegative real numbers q_1, q_2 such that $q_1 \leqq q_2$ (respectively $I = [q_1, \infty)$ for some nonnegative real number q_1), then admissible controls exist if and only if the solution w of the integral equation*

$$w(t) = q_2 + \int_0^t g(\tau)\,d\tau + \int_0^t b(\tau)\cdot\min\{w(\tau) - r(\tau), h(\tau)\}\,d\tau \tag{5}$$

satisfies $w \geqq r$ (respectively satisfies $w \geqq r$ for all sufficiently large q_2).

Proof. Sufficiency of the condition described for existence of admissible controls is obvious. For if w as described exists and $U_0(t) = \min\{w(t) - r(t), h(t)\}$, then (q_2, U_0) is an admissible control, since continuity of w and piecewise continuity of r imply $w(t) \geqq r(t)$ for all $t \in [0, T]$. Now suppose (a, u) is an admissible control with associated state variable x. Then $x = \mathbf{W}(a + \tilde{S})$, where

$$\tilde{S}(t) = \int_0^t g(\tau)\,d\tau + \int_0^t b(\tau)\,[u(\tau) - \min\{x(\tau) - r(\tau), h(\tau)\}]d\tau.$$

The integrand in the second integral is nonpositive by (2), and Lemma 1 then shows that $\mathbf{W}(a_0 + S_0) \geqq \mathbf{W}(a + \tilde{S}) \geqq r$ for all $a_0 \geqq a$, where $S_0(t) = \int_0^t g(\tau)\,d\tau$. This completes the proof of Theorem 1.

Remark. If the definition of admissible control is modified to require the inequality (2) to hold only for almost all t in $[0, T]$, then the statement and proof of Theorem 1 remain valid if (H3) and (H4) are replaced by the weaker requirement that g belong to $\mathfrak{L}^1([0, T])$ and r be measurable on $[0, T]$.

3. The case $C \geqq (k/b)$. In this section we describe and prove the effectiveness of an algorithm for constructing an optimal control under the additional hypothesis $C \geqq (k/b)$. Under this condition the parameter x_0 is essentially removed from the problem, since x_0 should obviously be selected as small as possible in consonance with the constraints (2) and (3). The remainder of the construction is guided by the principle of concentrating the nonzero values of $u(t)$ at values of t as large as possible. This idea originally arose intuitively from economic models.

From (H3) and (H4) we can partition the interval $[0, T]$ by $0 = a_1 \leqq z_1 \leqq a_2 \leqq \cdots \leqq z_N \leqq a_{N+1} \leqq z_{N+1} = T$ for some natural number N, where, for $1 \leqq n \leqq N$, $a_n < a_{n+1}$, r is constant on $[a_n, a_{n+1})$, $g(t) \geqq 0$ for almost all $t \in [a_n, z_n]$, and $g(t) < 0$ almost everywhere in $t \in [z_n, a_{n+1}]$. These conditions uniquely determine the a_n, z_n. Indeed $z_n = \sup \{t \in [a_n, T] \mid r(t) = r(a_n)$ and $g(\tau) \geqq 0$ for almost all $\tau \in [a_n, t]\}$, $a_{n+1} = \sup \{t \in [z_n, T] \mid r(t-) = r(a_n)$ and $g(\tau) < 0$ for almost all $\tau \in [z_n, t]\}$. Note that $a_{n+1} = z_n$ if and only if $r(z_n) \neq r(a_n)$. Existence of finite N as above is essentially due to the fact that each z_n is either a point of discontinuity of r or a left endpoint of a component of $g^{-1}[(-\infty, 0)]$.

Let

$$R_1 = \max \{r(0), q_1\},$$

and extend the definition to $n = 2, \cdots, N + 1$ by

$$R_{n+1} = \max \left\{ r(a_n), r(a_{n+1}), R_n + \int_{a_n}^{a_{n+1}} g(\tau)\, d\tau \right\}.$$

For $1 \leqq n \leqq N$ and $t \in [a_n, a_{n+1}]$ we define

$$v_n(t) = R_n + \int_{a_n}^{t} g(\tau)\, d\tau. \tag{6}$$

Let $P_{N+1} = R_{N+1}$, and assume P_{n+1} has been defined for some $n \geqq 1$ with the property $P_{n+1} \geqq R_{n+1}$. Let w_n be the solution w of the integral equation

$$\begin{aligned} w(t) = P_{n+1} &+ \int_{a_{n+1}}^{t} g(\tau)\, d\tau \\ &+ \int_{a_{n+1}}^{t} b(\tau) \cdot \min \{w(\tau) - r(\tau), h(\tau)\}\, d\tau. \end{aligned} \tag{7}$$

If $w_n(t) > v_n(t)$ for all $t \in [a_n, a_{n+1}]$, let $P_n = w_n(a_n) > v_n(a_n) = R_n$. If $w_n(t) = v_n(t)$ for some $t \in [a_n, a_{n+1}]$, let $P_n = v_n(a_n) = R_n$. This defines $P_n \geqq R_n$ in all possible cases, since the situation $w_n < v_n$ on $[a_n, a_{n+1}]$ is impossible because $w_n(a_{n+1}) = P_{n+1} \geqq R_{n+1} \geqq v_n(a_{n+1})$. In the first case we define

$$X_0(t) = w_n(t), \quad U_0(t) = \min \{w_n(t) - r(t), h(t)\}$$

for $t \in [a_n, a_{n+1})$. In the second alternative let

$$\tau_n = \sup \{t \in [a_n, a_{n+1}) \mid v_n(t) = w_n(t)\}$$

and define

$$X_0(t) = v_n(t), \quad U_0(t) = 0 \quad \text{for } t \in [a_n, \tau_n),$$

$$X_0(t) = w_n(t), \quad U_0(t) = \min\{w_n(t) - r(t), h(t)\} \quad \text{for } t \in [\tau_n, a_{n+1}).$$

Finally, to complete the definition of X_0 and U_0 on $[0, T]$, let $U_0(t) = 0$, $X_0(t) = v_{N+1}(t)$ for $t \in [a_{N+1}, T]$.

THEOREM 2. *Suppose $C \geqq k/b$ and admissible controls exist. Then $(X_0(0), U_0)$ is an optimal control with associated state variable X_0, where X_0, U_0 are defined above.*

It is convenient to give two lemmas before proceeding with the proof of this theorem. Note, for later use, that the lemmas do *not* require the assumption $C \geqq k/b$.

LEMMA 2. *$w_n(t) > r(a_n)$ for all t in $[z_n, a_{n+1})$, $n = 1, \cdots, N$.*

Proof. Suppose $w_n(\tau) \leqq r(a_n)$ for some $\tau \in [z_n, a_{n+1})$. Then (7) gives

$$w_n(t) - r(a_n) \leqq \int_\tau^t g(s)\, ds + \int_\tau^t b(s)[w_n(s) - r(a_n)]\, ds$$

for all $t \in [\tau, a_{n+1}]$, and application of the generalized Gronwall's lemma [10] yields

$$w_n(a_{n+1}) \leqq r(a_n) + \int_\tau^{a_{n+1}} g(s) \exp\left(\int_s^{a_{n+1}} b(y)\, dy\right) ds < r(a_n).$$

But $w_n(a_{n+1}) = P_{n+1} \geqq r(a_n)$ by virtue of $P_{n+1} \geqq R_{n+1}$ and the definition of the sequence $\{R_n\}$. This contradiction establishes the lemma.

LEMMA 3. *If (x_0, u) is an admissible control with corresponding state variable x, then $x(t) \geqq X_0(t)$ for all t in $[0, T]$.*

Proof. First note that the sequence of inequalities $x(a_n) \geqq R_n$, $n = 1, \cdots, N + 1$ is easily established by induction. In particular, the inequality $x(a_{N+1}) \geqq R_{N+1} = P_{N+1}$, with (2), (3) and the definition of X_0 on $[a_{N+1}, T]$, implies that $x \geqq X_0$ on $[a_{N+1}, T]$. Now we make the inductive assumption that $x \geqq X_0$ on $[a_{n+1}, T]$ for some natural number n, $1 \leqq n \leqq N$. If $P_n = R_n$, then the previously established inequality $x(a_n) \geqq R_n$, with (2), (3) and the fact that $X_0 = v_n$ on $[a_n, \tau_n)$, shows that $x \geqq X_0$ on $[a_n, \tau_n)$. Thus in order to prove $x \geqq X_0$ on $[a_n, a_{n+1})$ it remains only to prove that the situation $x(\tau) < X_0(\tau) = w_n(\tau)$ cannot prevail for any $\tau \in [a_n, a_{n+1})$ such that $X_0 = w_n$ on (τ, a_{n+1}). If such τ does exist, then, for $t \in [\tau, a_{n+1})$, x and X_0 satisfy the respective equations

$$x(t) = x(\tau) + \int_\tau^t g(s)\, ds$$

$$+ \int_\tau^t b(s)\cdot[u(s) - \min\{x(s) - r(s), h(s)\}]\,ds$$

$$+ \int_\tau^t b(s)\cdot\min\{x(s) - r(s), h(s)\}\,ds,$$

$$X_0(t) = X_0(\tau) + \int_\tau^t g(s)\,ds + \int_\tau^t b(s)\cdot\min\{X_0(s) - r(s), h(s)\}\,ds.$$

The integrand of the second integral in the first of these equations is nonpositive by (2). But (the remark following) Lemma 1 now implies $x(a_{n+1}) < X_0(a_{n+1})$, which contradicts the inductive assumption. This completes the proof of Lemma 3.

Proof of Theorem 2. It is an easy consequence of their definition that X_0 and U_0 satisfy the generalized state equation (3′). Furthermore, they also satisfy the constraint (2), provided only that $X_0 \geqq r$. Therefore in order to show $(X_0(0), U_0)$ is an admissible control it is only necessary to prove $X_0(0) \in I$ and $X_0 \geqq r$.

From the definition of the $\{P_n\}$ and X_0 we have $X_0(0) = P_1 \geqq R_1 \geqq q_1$. If $I = [q_1, \infty)$ this already shows $X_0(0) \in I$. Thus we assume $I = [q_1, q_2]$ for some finite q_2. But there exists an admissible control with associated state variable x. The state variable x must satisfy $x(0) \leqq q_2$, and Lemma 3 shows that $X_0(0) \leqq x(0)$, therefore $X_0(0) \leqq q_2$. We have now shown $X_0(0) \in I$ in all possible cases.

Note that v_n is nondecreasing on $[a_n, z_n]$, decreasing on $[z_n, a_{n+1}]$ and therefore, since $R_n \geqq r(a_n)$, there exists $h_n \in [z_n, a_{n+1}]$ such that $v_n - r(a_n)$ is nonnegative on $[a_n, h_n]$ and negative on $(h_n, a_{n+1}]$. For intervals $[a_n, a_{n+1}]$ on which $w_n > v_n$, this fact, with Lemma 2 and continuity of w_n, shows that $X_0 \mid [a_n, a_{n+1}) = w_n$ is greater than $r \mid [a_n, a_{n+1})$. Now suppose the graphs of v_n and w_n cross over $[a_n, a_{n+1}]$, and let

$$\tau_n = \sup\{t \in [a_n, a_{n+1}] \mid v_n(t) = w_n(t)\}$$

as above. Then $v_n(\tau_n) = w_n(\tau_n)$ by continuity, and furthermore the common value of v_n and w_n at τ_n is nonnegative because v_n is nonnegative on $[a_n, z_n]$ and w_n is nonnegative on $[z_n, a_{n+1}]$. Therefore, $\tau_n \leqq h_n$, and $X_0 \mid [a_n, \tau_n] = v_n \mid [a_n, \tau_n]$ is greater than or equal to $r(a_n)$. Now (the remark following) Lemma 1 shows that $w_n \geqq v_n$ on $[\tau_n, h_n]$, but $v_n \geqq r$ on $[\tau_n, h_n]$, therefore $w_n \geqq r$ on $[\tau_n, h_n]$. But Lemma 2 and the fact that $z_n \leqq h_n$ imply $w_n \geqq r$ on $[h_n, a_{n+1}]$. The last two sentences show that $X_0 \mid [\tau_n, a_{n+1}) = w_n \mid [\tau_n, a_{n+1})$ is greater than r, and thus we have shown $X_0 - r$ is nonnegative on $[a_n, a_{n+1})$ in all possible cases. Since $X_0 - r$ is trivially nonnegative on $[a_{N+1}, z_{N+1}]$, we have $X_0 \geqq r$ on all of $[0, T]$, and $(X_0(0), U_0)$ is an admissible control.

Let (x_0, u) be an arbitrary admissible control. We want to show

$$Cx_0 + \int_0^T k(t)u(t)\,dt \geqq CX_0(0) + \int_0^T k(t)U_0(t)\,dt$$

in order to prove $(X_0(0), U_0)$ is an optimal control. Lemma 3 shows that $X_0(0) \leqq x_0$. Let t_0 be defined by

$$t_0 = \sup\left\{t \in [0, T] \,|\, X_0(0) + \int_0^t b(\tau)U_0(\tau)\,d\tau \leqq x_0\right\}. \tag{8}$$

If $t_0 = T$, then

$$\begin{aligned} CX_0(0) + \int_0^T k(t)U_0(t) &\leqq C\left[X_0(0) + \int_0^T b(t)U_0(t)\,dt\right] \\ &\leqq Cx_0 \\ &\leqq Cx_0 + \int_0^T k(t)u(t)\,dt. \end{aligned} \tag{9}$$

Therefore we may assume $t_0 < T$, in which case

$$X_0(0) + \int_0^{t_0} b(t)U_0(t)\,dt = x_0, \tag{10}$$

and Lemma 3 shows that

$$\int_0^t b(\tau)u(\tau)\,d\tau \geqq \int_{t_0}^t b(\tau)U_0(\tau)\,d\tau, \qquad t_0 \leqq t \leqq T. \tag{11}$$

If $\bar{U}_0$ is defined by $\bar{U}_0(t) = U_0(t)$ for $t_0 < t \leqq T$, $\bar{U}_0(t) = 0$ for $0 \leqq t \leqq t_0$, then the last inequality may be written as

$$\int_0^t b(\tau)[u(\tau) - \bar{U}_0(\tau)]\,d\tau \geqq 0, \qquad 0 \leqq t \leqq T. \tag{12}$$

Now

$$\begin{aligned} \int_0^T k(t)u(t)\,dt - \int_{t_0}^T k(t)U_0(t)\,dt &= \int_0^T k(t)[u(t) - \bar{U}_0(t)]\,dt \\ &= \int_0^T \frac{k(t)}{b(t)}\, b(t)[u(t) - \bar{U}_0(t)]\,dt \\ &= \frac{k(0)}{b(0)} \int_0^{t_1} b(t)[u(t) - \bar{U}_0(t)]\,dt \\ &\geqq 0, \end{aligned} \tag{13}$$

where the last equality in this chain comes from the second mean value theorem for Lebesgue integrals [9, p. 134], redefining $(k/b)(T) = 0$ if

necessary, t_1 being some point of $[0, T]$, and the inequality comes from (12). But (10) yields

$$(14) \qquad \begin{aligned} Cx_0 &= CX_0(0) + \int_0^{t_0} Cb(t)U_0(t)\,dt \\ &\geqq CX_0(0) + \int_0^{t_0} k(t)U_0(t)\,dt, \end{aligned}$$

and the desired result follows from the last two inequalities. This completes the proof of Theorem 2.

4. The case $C \ngeqq k/b$. In this section we construct an optimal control for the case in which the inequality $C \geqq k/b$ does not hold by suitably modifying the functions X_0, U_0 defined in the preceding section. Note that X_0 and U_0 are well-defined regardless of the validity of the inequality $C \geqq k/b$.

Since $C < k(t)/b(t)$ for some $t \in [0, T]$, the number $T_0 = \sup\{t \mid C < k(t)/b(t)\}$ is well-defined. Let y_1 be given by

$$y_1(t) = X_0(T_0) + \int_{T_0}^{t} g(\tau)\,d\tau, \qquad 0 \leqq t \leqq T_0,$$

and if I is bounded above let y_2 be defined by

$$y_2(t) = q_2 + \int_0^t g(\tau)\,d\tau, \qquad 0 \leqq t \leqq T_0.$$

The remark following Lemma 1 (with $b = 0$) shows that $y_1 \geqq X_0$ on $[0, T_0]$, and consequently $q_1 \leqq P_1 = X_0(0) \leqq y_1(0)$. Therefore, $y_1(0) \in I$ if I is unbounded above, and for $I = [q_1, q_2]$ either $y_1(0) \in I$ or $y_1(0) > q_2$. In the event $y_1(0) \in I$ let $X(t) = X_0(t)$, $U(t) = U_0(t)$ for t in $(T_0, T]$ and $X(t) = y_1(t)$, $U(t) = 0$ for $t \in [0, T_0]$. If $y_1(0) \notin I$, then y_2 is well-defined, $y_2 < y_1$ by Lemma 1, and consequently $T_1 = \sup\{t \in [0, T_0] \mid y_2(t) = X_0(t)\}$ exists. In this case we define $X(t) = X_0(t)$, $U(t) = U_0(t)$ for $t \in (T_1, T]$, $X(t) = y_2(t)$, $U(t) = 0$ for $t \in [0, T_1]$.

THEOREM 3. *If $C < k(t)/b(t)$ for some $t \in [0, T]$, then $(X(0), U)$ is an optimal control with associated state variable X, where X and U are as defined above.*

Proof. The definition of X shows that $X(0) \in I$. Furthermore, Lemma 1 implies $X \geqq X_0$ and the argument that X_0 is greater than r as given in the third paragraph of the proof of Theorem 2 is still valid. Therefore, $(X(0), U)$ is an admissible control.

Let (x_0, u) be an arbitrary admissible control. The proof that $J(x_0, u) \geqq J(X(0), U)$ will be broken into two cases, according as $x_0 > X(0)$ or $x_0 \leqq X(0)$. First consider the case $x_0 > X(0)$, which can only occur if

$y_1(0) \in I$. The procedure for this case is similar to that in the proof of Theorem 2. We refer to equations in that proof with the understanding that always X_0, U_0 are to be replaced by X, U, respectively. Let t_0 be defined by (8). If $t_0 = T$, then, on using $U(t) = 0$ for $0 \leqq t \leqq T_0$ and $C \geqq k(t)/b(t)$ for $T_0 \leqq t \leqq T$, we obtain the inequality (9), and this shows $J(x_0, u) \geqq J(X(0), U)$. If $t_0 < T$, then (10) holds, and this implies $t_0 > T_0$ since $U = 0$ on $[0, T_0]$. Because $X = X_0$ on $(T_0, T]$, Lemma 3 again yields (11), equation (12) holds with $\bar{U}(t) = U(t)$ for $t_0 < t \leqq T$, $\bar{U}(t) = 0$ for $0 \leqq t \leqq t_0$, (13) follows exactly as before with $\bar{U}_0$ replaced by $\bar{U}$, and (14) follows from (10), $U = 0$ on $[0, T_0]$ and $Cb \geqq k$ on $[T_0, T]$. But (13) and (14) easily yield $J(x_0, u) \geqq J(X(0), U)$.

Now suppose $x_0 = x(0) \leqq X(0)$. Let $T_* = T_0$ or T_1, according as $y_1(0)$ is in I or not. Lemma 3 shows that $X(T_*) = X_0(T_*) \leqq x(T_*)$, and consequently, $t_1 = \sup \{t \in [0, T_*] \mid x(t) \leqq X(t)\}$ is well-defined. Furthermore, $X(t_1) = x(t_1)$ by continuity, and this implies

$$X(0) - x(0) = \int_0^{t_1} b(t) \cdot \{u(t) - U(t)\}\, dt. \tag{15}$$

But $C \leqq k/b$ on $[0, t_1]$ by $t_1 \leqq T_0$ and the fact that k/b is nonincreasing; therefore,

$$C[X(0) - x(0)] \leqq \int_0^{t_1} k(t)\{u(t) - U(t)\}\, dt. \tag{16}$$

We also have the inequality

$$\int_{t_1}^{t} b(\tau)\{u(\tau) - U(\tau)\}\, d\tau \geqq 0, \qquad t_1 \leqq t \leqq T. \tag{17}$$

For $t_1 \leqq t \leqq T_*$, (17) comes from $U = 0$ on $[t_1, T_*]$, and for $T_* \leqq t \leqq T$, it follows from (15), Lemma 3 and $X = X_0$ on $[T_*, T]$. But, again applying the second mean value theorem for Lebesgue integrals, (17) gives

$$\begin{aligned} &\int_{t_1}^{T} k(\tau)\{u(\tau) - U(\tau)\}\, d\tau \\ &\qquad = \frac{k(t_1)}{b(t_1)} \int_{t_1}^{t_2} b(\tau)\{u(\tau) - U(\tau)\}\, d\tau \geqq 0, \end{aligned} \tag{18}$$

where t_2 is some point in $[t_1, T]$. Combining (16) and (18) we have $J(x_0, u) \geqq J(X(0), U)$. This completes the proof of Theorem 3.

5. Uniqueness theorems. In this section we show that the optimal control constructed in §3 and §4 is the only optimal control provided k/b is actually decreasing rather than merely nonincreasing. The following lemma is the key result underlying the proof of uniqueness.

LEMMA 4. *Let f be decreasing and g integrable on the nontrivial interval $[t_1, t_2]$, and further suppose*:

(i) $\int_{t_1}^{t} g(\tau)\, d\tau \geqq 0$ *for all t in $[t_1, t_2]$*;

(ii) *there exists $t_3 \in [t_1, t_2]$ such that g is nonnegative almost everywhere in $[t_1, t_3]$ and is positive on a subset of $[t_1, t_3]$ having positive measure.*

Then $\int_{t_1}^{t_2} f(\tau)g(\tau)\, d\tau > 0$.

Proof. We may assume, without loss of generality, that g is positive on a nonnull subset of $[t_1, t_1 + \epsilon]$ for every $\epsilon > 0$. The second mean value theorem for integrals then gives

$$\begin{aligned}\int_{t_1}^{t_2} f(\tau)g(\tau)\, d\tau &= f(t_1)\int_{t_1}^{\tau_1} g(\tau)\, d\tau + f(t_3)\int_{\tau_1}^{t_3} g(\tau)\, d\tau \\ &\quad + f(t_3)\int_{t_3}^{\tau_2} g(\tau)\, d\tau + f(t_2)\int_{\tau_2}^{t_3} g(\tau)\, d\tau \\ &= f(t_2)\int_{t_1}^{t_3} g(\tau)\, d\tau + [f(t_3) - f(t_2)]\int_{t_1}^{\tau_2} g(\tau)\, dt \\ &\quad + [f(t_1) - f(t_3)]\int_{t_1}^{\tau_1} g(\tau)\, d\tau,\end{aligned}$$

where $t_1 < \tau_1 < t_3 \leqq \tau_2 \leqq t_2$, the strict inequalities coming from the fact that g is nonnegative on $[t_1, t_3]$. But the first two terms in this expression are nonnegative; and the third is positive because g is positive on a nonnull subset of $[t_1, \tau_1]$. This completes the proof.

THEOREM 4. *If $C \geqq k/b$ and k/b is decreasing on $[0, T]$, then the optimal control $(X_0(0), U_0)$ constructed in §3 is the unique optimal control, at least up to control variables u which are almost everywhere equal to U_0.*

Proof. The notation of §3, and especially that of the proof of Theorem 2, will be used extensively in the present proof. Let (x_0, u) be an admissible control. Lemma 3 shows that $x_0 \geqq X_0(0)$. The inequalities (13) and (14) again hold, and furthermore in the present problem strict inequality holds in (14) if $t_0 > 0$. But $t_0 > 0$ is equivalent to $X_0(0) < x_0$. This proves the theorem for the case $X_0(0) \neq x_0$, so we assume henceforth that $X_0(0) = x_0$ and u differs from U_0 on some set of positive measure.

Let m be the smallest natural number n such that u differs from U_0 on a nonnull subset of $[a_n, a_{n+1}]$, where $a_{N+2} = T$. If $w_m > v_m$ on $[a_m, a_{m+1}]$, let $\tau_m = a_m$, and otherwise let τ_m be as defined in §3. We claim that actually $\tau_m > a_m$ and $u(t) > U_0(t) = 0$ for all t in some nonnull subset of $[a_m, \tau_m]$.

For if this is not so, then $u = U_0$ on $[0, \tau_m)$, $x(\tau_m) = X_0(\tau_m)$, where x is the state variable associated with (x_0, u), and (the remark following) Lemma 1 shows that $x(t) \leqq X_0(t)$ for $t \in [\tau_m, a_m]$. But Lemma 3 implies $x(t) \geqq X_0(t)$ for $t \in [\tau_m, a_m]$. Therefore $x(t) = X_0(t)$ for all $t \in [\tau_m, a_m]$, which implies $u(t) = U_0(t)$ for almost all t in $[\tau_m, a_m]$ by absolute continuity of x and X_0. But this is not possible by the definition of m. Therefore, $u(t) > U_0(t)$ for all t in some nonnull subset of $[a_m, \tau_m]$, and this obviously necessitates $\tau_m > a_m$. Invoking Lemma 4, with $g = b \cdot (u - U_0)$ and (i) being satisfied by Lemma 3, we find

$$\int_{a_m}^{T} k(t)[u(t) - U_0(t)]\, dt = \int_{a_m}^{T} \frac{k(t)}{b(t)}\, b(t)[u(t) - U_0(t)]\, dt > 0.$$

Since $\int_0^{a_m} k(t)\ [u(t) - U_0(t)]\, dt = 0$, and $x_0 = X_0(0)$, this shows that $J(x_0, u) > J(X_0(0), U_0)$, and the proof of the theorem is complete.

THEOREM 5. *If $C > 0$, $C < k(t)/b(t)$ for some $t \in [0, T]$ and k/b is decreasing on $[0, T]$, then the optimal control $(X(0), U)$ constructed in* §4 *is the unique optimal control, at least up to control variables u which are almost everywhere equal to U.*

Proof. The notation of §3 and §4 will be used freely. Let (x_0, u) be an arbitrary admissible control with associated state variable x. For the case $x_0 > X(0)$ we proceed precisely as in the second paragraph of the proof of Theorem 3. However, in the present case, if $t_0 = T$ the first inequality in the chain (9) (with, always, X_0, U_0 replaced by X, U, respectively) is strict if U_0 is nonzero on a nonnull set, and the second inequality in this chain is strict otherwise. If $t_0 < T$, then the inequality (13) again holds (with $\bar{U}_0$ replaced by $\bar{U}$ as defined in the proof of Theorem 4) and strict inequality holds in (14). This shows $J(x_0, u) > J(X(0), U)$ in all possible cases for which $x_0 > X(0)$. In the event $x_0 < X(0)$ we again have the inequalities (16) and (18), but now strict inequality prevails in (16), and therefore again $J(x_0, u) > J(X(0), U)$.

The only remaining case is $x_0 = X(0)$, for which case we may assume u differs from U on a nonnull set. Then, arguing essentially as in the second paragraph of the proof of Theorem 4, we conclude that either $u > 0$ on some nonnull subset of $[0, T_*]$, or there exists a natural number m such that $a_m \geqq T_*$, $u = U$ almost everywhere on $[0, a_m]$, and $u(t) > 0$ for t in some nonnull subset of $[a_m, \tau_m]$. In the first case we let $a = 0$, in the second $a = a_m$. In either event Lemma 4, together with Lemma 3 and the assumption $X(0) = x_0$, implies

$$\int_a^T k(t)[u(t) - U(t)]\, dt = \int_a^T \frac{k(t)}{b(t)}\, b(t)[u(t) - U(t)]\, dt > 0.$$

But $u(t) = U(t)$ for almost all t such that $0 \leqq t \leqq a$, and $x_0 = X(0)$; thus this inequality implies $J(x_0, u) > J(X(0), U)$. This completes the proof of Theorem 5.

The following result is of interest because it shows that, under certain conditions, the only optimal control contains a singular point as described in §1, and therefore present versions of the maximum principle are not useful in dealing with such problems.

THEOREM 6. *Suppose $C > 0$ and k/b is decreasing on $[0, T]$. If $C > k/b$ let $T_* = 0$; otherwise let T_* be as defined in the proof of Theorem 3. If $q_1 + \int_0^t g(\tau)\, d\tau \leqq r(t)$ for some $t \in [T_*, T]$ and $C > k/b$ (respectively, $C \leqq k(t)/b(t)$ for some $t \in [0, T]$), then there exists a point $t \in [T_*, T]$ such that $X_0(t) = r(t)$ (respectively $X(t) = r(t)$).*

Proof. We shall merely outline the proof. Let u denote the optimal control variable, x the associated state variable, and suppose $x(t) > r(t)$ for all $t \in [0, T]$. If $x(0) > q_1$, then Gronwall's lemma shows $(x(0) - \epsilon, u)$ is an admissible control for sufficiently small positive ϵ, and obviously $J[x(0) - \epsilon, u] < J[x(0), u]$, contradicting optimality of $(x(0), u)$. Therefore $x(0) = q_1$. But $q_1 + \int_0^t g(\tau)\, d\tau \leqq r(t)$ for some $t \geqq T_*$, and consequently $u > 0$ on a nonnull subset of $[T_*, T]$. Let $\epsilon > 0$ be such that $x(t) - r(t) \geqq \epsilon$ for all $t \in [0, T]$, and $\bar{t} \in [T_*, T]$ such that $\epsilon \geqq \int_{\bar{t}}^T b(\tau)u(\tau)\, d\tau > 0$, and define $\tilde{u}(t) = u(t)$ for $t \in [0, \bar{t}]$, $\tilde{u}(t) = 0$ for $t \in (\bar{t}, T]$. Then $(x(0), \tilde{u})$ is an admissible control, and $J(x(0), \tilde{u}) < J(x(0), u)$, which contradicts optimality of $(x(0), u)$, and completes the proof of the theorem.

6. The one-reactor problem. In this section we apply the theory developed previously to the following model of a nuclear reactor power system. A specified power level $P(t)$, defined for times t satisfying $0 \leqq t \leqq T$, is to be supplied by nuclear reactors of a given type. The function P is, of course, nonnegative. A single reactor of the type considered is characterized by its power output p, fuel inventory m, operating cost per unit time K, and *net* fuel breeding rate b. The parameter b may have either sign, but the case $b \leqq 0$ is shown below to be trivial. The other reactor parameters are assumed to be positive. The amount of fuel available from natural sources at no cost is q, also assumed positive, and no more fuel is available at any price. One can consider this assumption an approximation to the situation in which there are two grades of ore from which fuel is obtained, one grade being of extremely high quality and the other of very poor quality. If money is discounted continuously at discount rate $\gamma \geqq 0$, and $n(t)$ denotes the

number of reactors in operation at time t, then the present net worth of the cost of operation over the planning period $[0, T]$ is

$$K_T(n) = K \int_0^T e^{-\gamma t} n(t)\, dt. \tag{19}$$

The requirement that the specified power level be achieved means n must satisfy

$$n(t) \geqq \frac{P(t)}{p}, \qquad 0 \leqq t \leqq T, \tag{20}$$

and the inequality

$$b \int_0^t n(\tau)\, d\tau - mn(t) + q \geqq 0, \qquad 0 \leqq t \leqq T, \tag{21}$$

must be satisfied in order that the system at no time use more fuel than is actually available.

Henceforth we suppose $P \in \mathfrak{L}'([0, T])$. Let $u(t) = n(t) - [P(t)/p]$. Then n minimizes (19) if and only if u minimizes

$$J_T(u) = K \int_0^T e^{-\gamma t} u(t)\, dt. \tag{22}$$

The constraints (20) and (21) are equivalent to

$$0 \leqq u(t) \leqq S(t) - \frac{P(t)}{p}, \qquad 0 \leqq t \leqq T, \tag{23}$$

where $S(t)$ is the total number of reactor loads of fuel available at time t, and is defined by

$$S'(t) = \frac{b}{m}\left\{ u(t) + \frac{P(t)}{p} \right\}, \qquad S(0) = \frac{q}{m}. \tag{24}$$

The problem defined by (22)–(24) is of the type defined by (1)–(3). We shall henceforth adopt the terminology used earlier for the abstract problem.

If $b \leqq 0$, then there exist functions u satisfying (23) with S given by (24) if and only if

$$\frac{q}{m} + \frac{b}{p} \int_0^t P(\tau)\, d\tau \geqq \frac{P(t)}{p}, \qquad 0 \leqq t \leqq T. \tag{25}$$

Furthermore, if this inequality does hold, then $u = 0$ satisfies (23) and obviously minimizes J_T in the subclass of $\mathfrak{L}^1([0, T])$ whose elements u satisfy (23). This completely solves the problem for the case $b \leqq 0$. The interpretation of these results is that, for a burner reactor, the best one can do is to just meet the minimal power demand. If no fuel shortage develops

under a policy of just supplying the power demand, then this is the optimal control. If a fuel shortage does develop under such a policy, then a shortage also results from any other admissible control, and the power demand cannot be supplied under the prescribed conditions.

However, if $b > 0$, then one has a breeder reactor, and there is the possibility of breeding sufficient fuel at early times to avoid a later shortage. For this case the data of the problem defined by (22)–(24) satisfy (H1) and (H2), and the results of §2 show that admissible controls exist (with (23) required to hold only almost everywhere in t) if and only if

$$\frac{pq}{m} e^{bt/m} \geqq P(t) \tag{26}$$

for almost all t in $[0, T]$.

Suppose $b > 0$ and $P = r_0 + P_0$, where r satisfies (H3) and P_0 is an absolutely continuous function. Then, defining $x = S - [P_0/p]$, we find the above problem is equivalent to that of minimizing (22), subject to

$$0 \leqq u(t) \leqq x(t) - \frac{r_0(t)}{p}, \tag{23'}$$

where

$$x'(t) = \frac{1}{p}\left\{\frac{b}{m} P(t) - P_0'(t)\right\} + \frac{b}{m} u(t), \qquad x(0) = \frac{q}{m} - \frac{P_0(0)}{p}. \tag{24'}$$

The data of the problem (22), (23′), (24′) satisfy (H1)–(H4) if $(b/m)P - P_0'$ satisfies (H4), and the results of §§3–5 are directly applicable to the problem in this event. Since $I = \{(q/m) - P(0)/p\}$ is a singleton set, the problem remains unchanged if Cx_0 is added to (22) and J_T is regarded as a function of $x_0 \in I$ and u. Thus, taking C sufficiently large, it is clear that the solution to this problem is described by the results of §3, and furthermore this is the unique solution if $\gamma > 0$.

A special case of some interest is that in which $b > 0$, P is absolutely continuous, and $(b/m)P - P'$ is nondecreasing. In this event $P = P_0$, and the optimal control u_0 is given by

$$u_0(t) = \begin{cases} x_0(t) & \text{if } t_1 \leqq t \leqq t_2, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$A = \{t \in [0, T] \mid (b/m)P(t) - P'(t) \geqq 0\},$$

$$t_2 = \begin{cases} \inf A & \text{if } A \text{ is nonempty,} \\ T & \text{otherwise,} \end{cases}$$

$$y_0(t) = \frac{q}{m} - \frac{P(t)}{p} + \frac{b}{mp}\int_0^t P(\tau)\, d\tau, \qquad 0 \leqq t \leqq T,$$

$$R = \max\{0, y_0(t_2)\},$$

$$x_0(t) = \left[R + \frac{P(t_2)}{p}\right] e^{b(t-t_2)/m} - \frac{P(t)}{p}, \qquad 0 \leqq t \leqq T,$$

$$B = \{t \in [0, t_1] \mid x_0(t) = y_0(t)\},$$

$$t_1 = \begin{cases} \sup B & \text{if } B \text{ is nonempty,} \\ 0 & \text{otherwise.} \end{cases}$$

Note that if $R = y_0(t_2) \geqq 0$, then $t_1 = t_2$, and u_0 is everywhere 0. Thus the interesting case is $R = 0$, which corresponds to $y_0(t_2) \leqq 0$. In this situation $x_0(t_2) = 0 \geqq y_0(t_2)$, and if (26) holds, then

$$x_0(0) = \frac{P(t_2)}{p} e^{-bt_2/m} - \frac{P(0)}{p} \leqq \frac{q}{m} - \frac{P(0)}{p} = y_0(0),$$

and therefore B is nonempty, $t_1 = \sup B$. The number of reactors n_0 and fuel stockpile S_0 corresponding to the optimal control for the case $R = 0$ are illustrated qualitatively in Fig. 1. The same control, with u_0 extended beyond t_2 as zero, is optimal for any $T \geqq \inf A$.

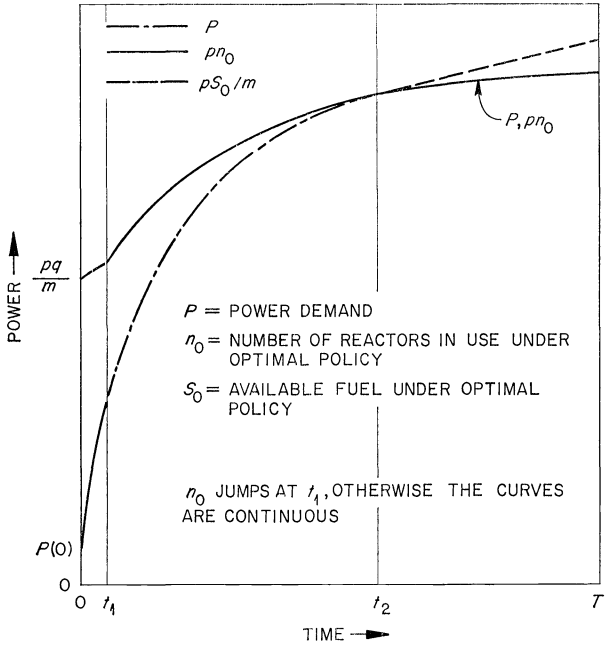


FIG. 1. *Optimal policy for the one-reactor problem, $b > 0$, P absolutely continuous, (b/m) $P - P'$ nondecreasing*

We note in passing that a solution of the special case $\gamma = 0$ of the problem considered in this section has been given previously [8]. The previously described solution is, in general, different from that given by the present theory, as may be seen from the special case just considered. This shows that the uniqueness theorems of §5 cannot be generalized to include situations in which (k/b) is merely nonincreasing.

7. The two-reactor problem. A specified amount of power $P(t)$ is to be supplied for times t satisfying $0 \leqq t \leqq T$, where T is a given positive number, by a reactor system composed of two types of reactors. The two types of reactors are assumed to use the same fissionable material for fuel. Each reactor type is specified by its power output p_i, fuel inventory m_i, *net* fuel breeding rate b_i, and operating cost per unit time k_i. To avoid trivialities we assume at least one p_i, say p_1, is positive. The parameters m_1, m_2, k_1, k_2 are positive, p_2 is nonnegative, and the b_i have arbitrary sign. We assume there is available from natural sources an amount q_0 of fuel at unit cost $C > 0$, and that there is no more fuel available from such sources at any cost. We let $n_i(t)$ denote the number of reactors of type i in operation at time t. It is required that the system operate so as to supply exactly the required amount of power,

$$p_1 n_1(t) + p_2 n_2(t) = P(t), \qquad 0 \leqq t \leqq T, \tag{27}$$

and the n_i must be nonnegative

$$0 \leqq n_1(t), \qquad 0 \leqq n_2(t), \qquad 0 \leqq t \leqq T. \tag{28}$$

We assume zero discount rate, so the functional to be minimized is

$$\bar{K}_T(q, n_1, n_2) = Cq + k_1 \int_0^T n_1(t)\, dt + k_2 \int_0^T n_2(t)\, dt, \tag{29}$$

where $q \in [0, q_0]$ is the total amount of fuel purchased from natural sources, and the constraint

$$m_1 n_1(t) + m_2 n_2(t) \leqq q + \int_0^t [b_1 n_1(\tau) + b_2 n_2(\tau)]\, d\tau, \quad 0 \leqq t \leqq T, \tag{30}$$

must be satisfied. This constraint simply says that no more fuel can be used at any time than is actually available at that time. The fuel purchased externally is assumed to be available initially, which is permissible since there is no advantage to delaying the purchase.

We use (27) to solve for n_1 in terms of n_2. If we use this result in (28)–(30), and write u for n_2, then the problem is transformed into the equivalent problem of finding (q, u) which minimizes

$$K(q, u) = Cq + \kappa \int_0^T u(t)\, dt \tag{31}$$

subject to the constraints:

$$0 \leqq u(t) \leqq \frac{P(t)}{p_2}, \tag{32}$$

$$\mu u(t) \leqq S(t) - \frac{m_1}{p_1} P(t), \tag{33}$$

where $S(t)$ is the fuel available at time t and is defined by the initial value problem

$$S'(t) = \beta u(t) + \frac{b_1}{p_1} P(t), \qquad S(0) = q \in [0, q_0]. \tag{34}$$

Here the constants κ, μ and β are defined by

$$\kappa = k_2 - k_1\left(\frac{p_2}{p_1}\right),$$

$$\mu = m_2 - m_1\left(\frac{p_2}{p_1}\right),$$

$$\beta = b_2 - b_1\left(\frac{p_2}{p_1}\right).$$

If $p_2 = 0$ then the right-hand inequality in (32) is to be interpreted as no constraint.

We shall assume in the following that $\kappa \geqq 0$, which can always be accomplished by interchanging the subscripts if necessary. We shall also assume that $\mu > 0$, which does involve some loss of generality. This assumption is essentially that one of the reactors has a strictly larger fuel inventory per unit power than the other and that this reactor does not have a smaller operating cost per unit power. Our theory does not apply to the case $\mu \leqq 0$, and it seems probable that this case is essentially more difficult than that of positive μ.

Consider the case $\beta \leqq 0$. Then there exist functions u satisfying (32)–(34) if and only if $u = 0$ satisfies these conditions, and in this event $u \equiv 0$ obviously minimizes (31) in the class of such functions. This is the trivial case in which reactor type 1 is uniformly better than reactor type 2.

We suppose henceforth that $\beta > 0$. Then the remark following Theorem 1 shows that admissible controls exist if and only if the function S_0, defined implicitly by the integral relation

$$S_0(t) = q_0 + \frac{b_1}{p_1}\int_0^t P(\tau)\,d\tau + \beta\int_0^t \min\left\{\frac{S_0(t)}{\mu} - \frac{m_1}{\mu p_1} P(t), \frac{P(t)}{p_2}\right\} d\tau,$$

satisfies $S_0(t) \geqq (m_1/p_1)P(t)$.

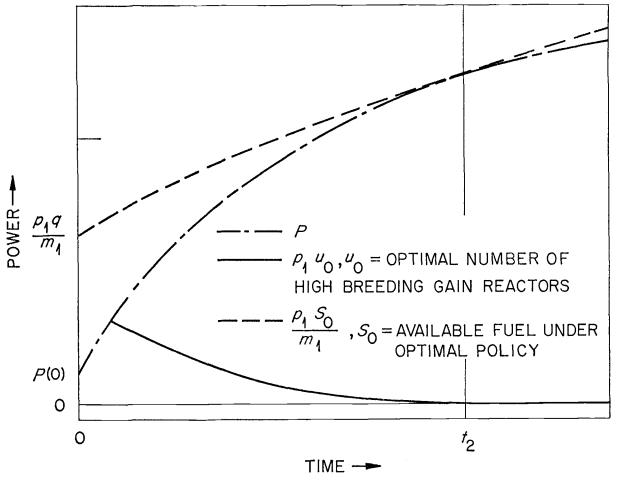


FIG. 2. *Optimal policy for the two-reactor problem,* $\kappa \geqq 0, \mu \geqq 0, \beta \geqq 0, \mu b_1 P - m_1 P'$ *increasing,* $c > \kappa/\beta$

Suppose $P = P_0 + r$, where P_0 and r satisfy the same conditions as in the preceding section. Let $x = (S/\mu) - (m_1/\mu p_1)P_0$. Then (32) and (33) are equivalent to

$$0 \leqq u(t) \leqq \min\left\{x(t) - \frac{m_1}{p_1} r(t), \frac{P(t)}{p_2}\right\}, \tag{35}$$

where x is defined by

$$\begin{aligned} x'(t) &= \frac{b_1}{p_1} P(t) - \frac{m_1}{\mu p_1} P_0'(t) + \frac{\beta}{\mu} u(t), \\ x(0) &= \frac{q}{\mu} - \left(\frac{m_1}{\mu p_1}\right) P_0(0). \end{aligned} \tag{36}$$

If $(b_1/p_1)P - (m_1/\mu p_1)P_0'$ satisfies (H4), then our general theory applies to the problem of minimizing (31) subject to (35) and (36).

The case of absolutely continuous P and $b_1 P - (m_1/\mu)P'$ increasing is of special interest. If $C \leqq (\kappa/\beta)$, then the optimal control $u_0 = n_2$ is zero, and the corresponding value of n_1 is given by (27). If $C > \kappa/\beta$, then the optimal number of high breeding gain reactors as a function of time is given by

$$u_0(t) = \begin{cases} \min\{x_0(t), P(t)/p_2\} & \text{if } t_1 \leqq t \leqq t_2, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$A = \{ t \in [0, T] \mid b_1 P(t) - (m_1/\mu)P'(t) \geqq 0\},$$

$$t_2 = \begin{cases} \inf A & \text{if } A \text{ is nonempty,} \\ T & \text{otherwise,} \end{cases}$$

$$y_0(t) = \frac{b_1}{p_1}\int_0^t P(\tau)\,d\tau - \frac{m_1}{\mu p_1}P(t), \qquad 0 \leqq t \leqq T,$$

$$R = \max\{0, y_0(t_2)\},$$

and x_0 is defined by the integral relation

$$x_0(t) = R + \frac{m_1}{\mu p_1}[P(t_2) - P(t)] + \frac{b_1}{p_1}\int_{t_2}^t P(\tau)\,d\tau + \frac{\beta}{\mu}\int_{t_2}^t \min\left\{x_0(t), \frac{P(t)}{p_2}\right\} d\tau, \qquad 0 \leqq t \leqq T,$$

$$B = \{t \in [0, t_1] \mid x_0(t) = y_0(t)\},$$

$$t_1 = \begin{cases} \sup B & \text{if } B \text{ is nonempty,} \\ 0 & \text{otherwise.} \end{cases}$$

If $t_2 = 0$, then the optimal control u_0 is everywhere 0. Thus the interesting case is when $t_2 > 0$, which implies $R = 0 > y(t_2)$ and $t_1 = 0$. The optimal number of high breeding gain reactors as a function of time for this case is illustrated qualitatively in Fig. 2.

REFERENCES

[1] R. Bellman, W. H. Fleming and D. V. Widder, *Variational problems with constraints*, Ann. Mat. Pura Appl., 41 (1956), pp. 301–323.

[2] Leonard D. Berkovitz, *Variational methods in problems of control and programming*, J. Math. Anal. Appl, 3 (1961), pp. 145–169.

[3] ———, *On control problems with bounded state variables*, Ibid., 5 (1962), pp. 488–498.

[4] Lamberto Cesari, *Existence theorems for optimal solutions in Pontryagin and Lagrange problems*, this Journal, 3 (1965), pp. 475–498.

[5] T. Guinn, *The problem of bounded space coordinates as a problem of Hestenes*, this Journal, 3 (1965), pp. 181–190.

[6] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[7] N. Levinson, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73–83.

[8] Paul Nelson, *An optimization problem in nuclear reactor economics*, ORNL-4172, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 1967.

[9] W. W. Rogosinski, *Volume and Integral*, Interscience, New York, 1962.

[10] G. Sansone and R. Conti, *Nonlinear Differential Equations*, Macmillan, New York, 1964.

[11] W. F. Tyndall, *A duality theorem for a class of continuous linear programming problems*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 644–666.

[12] ———, *An extended duality theorem for continuous linear programming problems*, Ibid., 15 (1967), pp. 1294–1298.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

A MINIMIZATION PROBLEM AND ITS APPLICATIONS TO OPTIMAL CONTROL AND SYSTEM SENSITIVITY*

WILLIAM A. PORTER†

Abstract. A Banach space minimization problem is formulated. Existence and uniqueness of the solution are discussed and the solution is completely characterized. Applications to problems of optimal control and system sensitivity are pointed out.

1. Introduction. In [1] the following basic optimization problem is considered. Let T be a bounded linear transformation between Banach spaces B and D, respectively. With T onto D and $\zeta \in D$ arbitrary, find (if one exists) a preimage of ζ with minimum norm. Of obvious interest are the questions of existence and uniqueness and the properties of the mapping from $\zeta \in D$ to a minimum norm preimage. In [2] several generalizations of this problem are treated. These results, in an expanded and extended form, provide the basis of [3, Chap. IV], which also contains several applications to lumped parameter systems.

In the present paper the following related problem is considered. With B, D, T and $\zeta \in D$ as before and $U \subset B$, the unit ball, find an element $u \in U$ (if one exists) which minimizes the functional $\| \zeta - Tu \|$. Once more the questions of existence, uniqueness and characterization of the minimizing element(s) are of apparent interest. In Hilbert spaces this problem has been studied by Rosenblum [4], Arcangeli [5], and others. Attention is called also to Rubio [6] who considered this problem for a specific Banach space (see also [11]). In this study we shall consider, without loss of generality, real spaces.

2. Some preliminaries. It is helpful to review some of the results in the papers cited in the Introduction while emphasizing the notation and assumptions to be used throughout. For the moment B and D will denote Banach spaces and T, a bounded linear transformation from B into D. The unit ball and the unit sphere of B will be denoted by U and ∂U, respectively, while $C = T(U)$ denotes the image of U in D under T. The boundary of C will be denoted by ∂C. It is easily verified that C is bounded, convex and circled. It follows from the open mapping theorem, however, that C is a neighborhood of $0 \in D$ and hence that the closure of C is a convex body (a closed convex set with nonvacuous interior) if and only if T is onto. The element $\tilde{f} \in B$ will be called an extremal (see [1] or [3]) of $f \in B^*$ if

* Received by the editors June 30, 1967, and in revised form November 7, 1967.

† Department of Electrical Engineering, University of Michigan, Ann Arbor, Michigan 48104. This research was supported in part by the United States Army Research Office-Durham under Contract DA-31-124, ARO-D-391.

$\| \bar{f} \| = 1$ and $\langle \bar{f}, f \rangle = \| f \|$ both hold. The set of all extremals for $f \in B^*$ is denoted by $\{\bar{f}\}$.

If B is reflexive, rotund and smooth and if T is onto, then (see [3]):

(i) C is closed (i.e., a convex body);

(ii) every $\zeta \in D$ has a unique minimum norm preimage;

(iii) every $f \in B^*$ has a unique extremal $\bar{f} \in B$;

(iv) C has a unique hyperplane of support at every boundary point.

The minimum preimage u_ζ of $\zeta \in D$ is moreover given by

$$u_\zeta = p(\zeta)\overline{T^*\phi_\zeta},$$

where p is the Minkowski functional of the set C, T^* is the conjugate of T, and $\phi_\zeta \in D^*$ defines the unique hyperplane of support of C at $[p(\zeta)]^{-1}\zeta \in \partial C$.

Suppose now that $K: B^* \to B$ is given by $K(f) = \| f \| \bar{f}, f \in B^*$, and that $J: D^* \to D$ is the mapping

$$J(\phi) = TKT^*\phi = \| T^*\phi \| T(\overline{T^*\phi}), \qquad \phi \in D^*.$$

The mapping J is considered in [3] and found to be one-to-one, onto, bounded and hence invertible. Furthermore, the condition

$$0 \leqq \langle \phi_1 - \phi_2, J(\phi_1) - J(\phi_2) \rangle \quad \text{for all} \quad \phi_1, \phi_2 \in D^*$$

holds. A Hilbert space operator T satisfying this relation is called *monotonic*. If the stronger relation $\epsilon \| \phi_1 - \phi_2 \|^2 \leqq \langle \phi_1 - \phi_2, T\phi_1 - T\phi_2 \rangle$ for all $\phi_1, \phi_2 \in D$ holds for some $\epsilon > 0$, then T is said to be *strongly monotonic*. The Lipschitz norm of an operator T is the number

$$\| T \| = \sup \{ \| T\phi_1 - T\phi_2 \| / \| \phi_1 - \phi_2 \| : \phi_1 \neq \phi_2 \}.$$

If $\| T \| < \infty$, then T is said to be *Lipschitzian*. In particular, if D is a Hilbert space, then J is a monotonic operator. Moreover, $J_\lambda = \lambda I + J$ is strongly monotonic for every $\lambda > 0$, and using a result of Zarantonello [7] it then follows that J_λ has a bounded Lipschitzian inverse for every $\lambda > 0$. At the point $\lambda = 0$, J_λ is invertible but not necessarily Lipschitzian unless further assumptions are made.

3. The basic problem. The assumptions that B is reflexive, rotund and smooth, and that T is onto are continued here. Also D is taken as a Hilbert space. For $\zeta \in H$ the functional Π is defined by

$$\Pi(u) = \| \zeta - Tu \|, \qquad u \in B.$$

Since D is reflexive and rotund and C is closed and convex there exists a unique element $c_\zeta \in C$ which is closest to ζ, that is,

$$\| \zeta - c_\zeta \| \leqq \| \zeta - c \| \quad \text{for all} \quad c \in C$$

with strict inequality unless $c = c_\zeta$. It is clear that $\hat{u}$ minimizes Π over U:

$$\Pi(\hat{u}) \leqq \Pi(u) \quad \text{for all} \quad u \in U$$

if and only if $T\hat{u} = c_\zeta$. If $\Pi(\hat{u}) > 0$, then $\hat{u}$ is the unique minimum norm preimage of c_ζ.

Following Rubio we consider also the functional

$$\pi_\phi(u) = \langle \zeta - Tu, \phi \rangle = \langle \zeta, \phi \rangle - \langle \phi, Tu \rangle, \qquad \phi \in D.$$

The minimum of this functional on U is obviously attained at the extremal of $T^*\phi$, in which case we have

$$\pi_\phi(\overline{T^*\phi}) = \langle \zeta - T(\overline{T^*\phi}), \phi \rangle \leqq \langle \zeta - Tu, \phi \rangle \quad \text{for all} \quad u \in U.$$

Moreover, for one $\hat{\phi} \in D$, namely, the support normal to C at c, the minimizing element for Π and π coincide. An interesting feature of Rubio's article is the characterization of $\hat{\phi}$.

Consider now the operator L on D defined by

$$L\phi = \| \zeta - T(\overline{T^*\phi}) \|^{-1}[\zeta - T(\overline{T^*\phi})], \qquad \phi \in D. \tag{1}$$

Here we assume that $\Pi(\hat{u}) > 0$. Concerning L we have the following lemma.

LEMMA 1. *If ϕ is a fixed point of L, then $\overline{T^*\phi}$ minimizes Π.*

Proof. From the above remarks we have the inequality

$$\langle \zeta - T(\overline{T^*\phi}), \phi \rangle \leqq \langle \zeta - Tu, \phi \rangle \quad \text{for all} \quad u \in U.$$

This together with the condition

$$\| \zeta - T(\overline{T^*\phi}) \| \phi = \zeta - T(\overline{T^*\phi})$$

shows that $\| \phi \| = 1$ and

$$\langle \zeta - T(\overline{T^*\phi}), \phi \rangle = \| \zeta - T(\overline{T^*\phi}) \| \leqq \langle \zeta - Tu, \phi \rangle \quad \text{for all} \quad u \in U.$$

Using the Cauchy-Schwarz inequality on the right-hand side completes the proof.

An obvious corollary is the following.

COROLLARY. *If B is reflexive, rotund, and smooth, and if T is onto, then the operator L, for each ζ at a positive distance from C, has a unique fixed point.*

In view of the definition of J, the defining equation for L may be rewritten as

$$L\phi = \| [\| T^*\phi \| \zeta - J\phi] \|^{-1}[\| T^*\phi \| \zeta - J\phi], \quad \phi \in D.$$

Letting

$$\lambda = \| [\| T^*\phi \| \zeta - J\phi] \|, \tag{2}$$

it then follows that $L\phi = \phi$ implies that $\lambda\phi = \| T^*\phi \|\zeta - J\phi$.

Since $\lambda I + J$ is invertible for $\lambda > 0$ we may solve this latter expression explicitly for ϕ:

$$\phi = \| T^*\phi \|(J + \lambda I)^{-1}\zeta. \tag{3}$$

With T onto, T^* is one-to-one. Operating on both sides of this expression with T^* and taking norms, it follows that λ must satisfy the condition

$$\| T^*(J + \lambda I)^{-1}\zeta \| = 1. \tag{4}$$

These observations lead to the next lemma.

LEMMA 2. *The conditions*:

$$\text{(i)} \qquad \phi = L\phi,$$

$$\text{(ii)} \qquad \phi = \| T^*\phi \| J_\lambda^{-1}\zeta,$$

together with $\| \phi \| = 1$, *are equivalent.*

Proof. We have seen that (i) implies (ii) for $\lambda = \| [\| T^*\phi \|\zeta - J\phi) \|$. Conversely, if (ii) holds for some λ, then clearly $(J + \lambda I)\phi = \| T^*\phi \|\zeta$, which implies

$$\| T^*\phi \|\zeta - J\phi = \lambda\phi.$$

Scalar multiplication results in

$$\| [\| T^*\phi \|\zeta - J\phi] \|^{-1}[\| T^*\phi \|\zeta - J\phi] = \lambda \| [\| T^*\phi \|\zeta - J\phi] \|^{-1}\phi.$$

Since the left-hand side of this expression is a unit vector, taking norms and using $\| \phi \| = 1$ results in the condition $\lambda = \| [\| T^*\phi \|\zeta - J\phi] \|$, which completes the proof.

In condition (i) the assumption $\| \phi \| = 1$ is superfluous but seems necessary in condition (ii). Equations (2) and (4) give two characterizations of the scalar λ. While (4) is perhaps the more useful, (2) yields also the relationship

$$\lambda = \| T^*\phi \| \Pi (\overline{T^*\phi}),$$

where ϕ is the fixed point of L.

Remark 1. In Hilbert spaces $J = TT^*$ and (3) becomes

$$\phi = \| T^*\phi \|(TT^* + \lambda I)^{-1}\zeta.$$

Consequently, $\hat{u} = T^*\phi / \| T^*\phi \|$ is identifiable as

$$\hat{u} = (T^*T + \lambda I)^{-1}T^*\zeta \tag{5}$$

and therefore,

$$c_\zeta = T\hat{u} = T(T^*T + \lambda I)^{-1}T^*\zeta. \tag{6}$$

The scalar λ is the unique solution of the equation

$$\| T^*(TT^* + \lambda I)^{-1}\zeta \| = 1. \tag{7}$$

Equations (5), (6) and (7) are essentially the characterization of Rosenblum for Hilbert spaces with T onto.

4. Some generalizations. In §2 and §3 several assumptions were imposed on B and T. Among other things it was shown that if B is reflexive, rotund and smooth the results of Rosenblum can be identified in this Banach space setting. It was assumed also that T is bounded, onto, and has a Hilbert space range. In this section we investigate the possibility of weakening these assumptions.

Consider first the assumption that B is reflexive. This condition is sufficient to guarantee that C is closed and that every $f \in B^*$ has at least one extremal $\bar{f} \in B$. In [3, §4.3], alternative conditions are given on T, B and D which guarantee that C is closed. While these other conditions are frequently more useful in the applications, they are cumbersome to discuss and we shall not take them up here. It is noted, however, that in the alternative situation extremals are actually taken from a Banach space into its dual and hence, by the Hahn-Banach theorem, extremals always exist. As a convenience we shall continue to assume in the following that B is reflexive.

The assumption that T is onto is also nonessential. In the functional Π we may always replace ζ by its orthogonal projection on the closure of the range of T without changing the minimization problem. Thus, without loss of generality we assume that T has dense but not necessarily closed range. The set C is still bounded, convex and closed, as is the set $\{\zeta\} - C$. Since the (Hilbert) space D is reflexive, $\{\zeta\} - C$ is weakly compact and Π, being a continuous functional, assumes its lower bound. Since D is rotund this lower bound is uniquely attained. That is, a unique $c' \in C$ exists such that

$$\| \zeta - c' \| \leqq \| \zeta - c \| \quad \text{for all} \quad c \in C$$

holds. Since $c' \in C$ at least one $u' \in U$ exists satisfying $c' = Tu'$ and consequently, $\Pi(u') \leqq \Pi(u)$ for all $u \in U$ is satisfied. We assume $0 < \Pi(u')$.

Suppose now that K denotes a convex set in a Banach space B. For every $f \in B^*$ we define the number f_K by

$$f_K = \sup_{x \in K} \langle x, f \rangle.$$

In general, f_K may be infinite. The following is a well-known duality theorem (see Nirenberg [8]).

THEOREM 1. *Let d denote the distance from $x_0 \in B$ to the convex set K. Then*

$$d = \inf_{x \in K} \| x_0 - x \| = \max_{\|f\| \leqq 1} \{\langle x_0, f \rangle - f_K\},$$

where the maximum on the right is achieved for some f_0.

In using this theorem B becomes the Hilbert space D and K is taken as the set C. For $\phi \in D$ it follows that

$$\phi_C = \sup_{c \in C} \langle c, \phi \rangle = \sup_{u \in U} \langle u, T^*\phi \rangle = \| T^*\phi \|.$$

Using $\zeta = x_0$ in the theorem results in

$$\inf_{c \in C} \| \zeta - c \| = \max_{\|\phi\| \leq 1} \{ \langle \zeta, \phi \rangle - \| T^*\phi \| \} = \langle \zeta, \phi' \rangle - \| T^*\phi' \|,$$

where the last equality holds at some ϕ', $\| \phi' \| = 1$. If the infimum on the left is attained at some $c' \in C$ for which $c' = Tu'$, $u' \in U$, and then

$$\| \zeta - Tu' \| = \langle \zeta, \phi' \rangle - \| T^*\phi' \| \tag{8}$$

holds. Since $\| \phi' \| = 1$ it follows that

$$\| \zeta - Tu' \| \geqq \langle \zeta - Tu', \phi' \rangle = \langle \zeta, \phi' \rangle - \langle u', T^*\phi' \rangle,$$

which together with (8) shows that

$$\| T^*\phi' \| \leqq \langle u', T^*\phi' \rangle.$$

Since $\| u' \| \leqq 1$ it follows that this last expression can hold only with equality and then only if $u' \in \{\overline{T^*\phi'}\}$.

THEOREM 2. *If* $c' \in C$ *minimizes* Π, *then*

(i) C *is supported at* c' *by some* ϕ';

(ii) $c' = T(\overline{T^*\phi})$;

(iii) *if* $\| \phi' \| = 1$, *then* $\phi' = (\zeta - c')/\| \zeta - c' \|$.

Proof. In the preceding paragraph part (ii) is established. Part (i) follows easily from the chain

$$\langle c', \phi' \rangle = \langle \overline{T^*\phi'}, T^*\phi' \rangle = \| T^*\phi' \| \geqq \langle u, T^*\phi' \rangle = \langle Tu, \phi' \rangle, \qquad u \in U.$$

To prove (iii) we rewrite (8) in the form

$$\begin{aligned} \| \zeta - c' \| &= \langle \zeta, \phi' \rangle - \| T^*\phi' \| = \langle \zeta - T(\overline{T^*\phi'}), \phi' \rangle \\ &= \langle \zeta - c', \phi' \rangle \leqq \| \zeta - c' \| \cdot \| \phi' \|. \end{aligned}$$

Since this is a Hilbert space relationship part (iii) follows.

Consider now the operator L of (1). In the present setting we shall use the following terminology. A fixed point of L is a ϕ, $\| \phi \| = 1$, such that $\{\overline{T^*\phi}\}$ is nonvacuous and for some extremal of $T^*\phi$,

$$\zeta - T(\overline{T^*\phi}) = \| \zeta - T(\overline{T^*\phi}) \| \phi.$$

With the meaning of these conditions we have the following corollary.

COROLLARY. ϕ *is a fixed point of* L *if and only if* Π *is minimized at* $T(\overline{T^*\phi})$. *Moreover, for every* ζ, L *has a unique fixed point,* Π *has a unique minimum.*

Proof. A review of Lemma 1 shows that it holds in the present setting with the same proof and this gives half of the implication of the present corollary. Conversely, if Π is minimized at some c', then by part (iii) of Theorem 2, $\phi' = (\zeta - c')/\| \zeta - c' \|$ is the outward normal to C at c'. Moreover, using part (ii) of the theorem, it is clear that $\phi' = L\phi'$. In view of part (iii) the uniqueness follows.

It is interesting to consider also the transformation K and the operator J introduced in §1. If B is reflexive and rotund, then K is a well-defined function independent of properties of T. If B is also smooth, then K is also one-to-one. Consider now the (weaker) assumption that each element in $R(T^*)$, the range of T^*, has a unique extremal; then K is still well-defined on $R(T^*)$. The mapping J is then also a well-defined operator on D. From the expansion

$$\begin{aligned}\langle \phi_1 - \phi_2, J\phi_1 - J\phi_2 \rangle = (\| T^*\phi_1 \| - \| T^*\phi_2 \|)^2 + \| T^*\phi_1 \|(\| T^*\phi_2 \| \\ - \langle T^*\phi_2, \overline{T^*\phi_1} \rangle) + \| T^*\phi_2 \|(\| T^*\phi_1 \| - \langle T^*\phi_1, \overline{T^*\phi_2} \rangle),\end{aligned} \tag{9}$$

it is clear that J is monotonic. The implication

$$\begin{aligned}0 = \langle \phi_1 - \phi_2, J\phi_1 - J\phi_2 \rangle \quad &\text{if and only if} \\ \| T^*\phi_1 \| = \| T^*\phi_2 \| \quad &\text{and} \quad \overline{T^*\phi_1} = \overline{T^*\phi_2}\end{aligned} \tag{10}$$

follows directly from (9). From these observations we have the following lemma.

LEMMA 3. *If K is well-defined on $R(T^*)$, then J is a well-defined bounded monotonic operator. If also T has dense range and K is one-to-one on $R(T^*)$, then J is also one-to-one.*

Proof. It remains only to verify the second assertion. If $J\phi_1 = J\phi_2$, then $\langle \phi_1 - \phi_2, J\phi_1 - J\phi_2 \rangle = 0$ which, using (10), implies

$$\| T^*\phi_1 \| = \| T^*\phi_2 \|, \qquad \overline{T^*\phi_1} = \overline{T^*\phi_2}.$$

Together this means that $K(T^*\phi_1) = K(T^*\phi_2)$ which, assuming that K is one-to-one on $R(T^*)$, implies $T^*\phi_1 = T^*\phi_2$. If T has dense range, then T^* is one-to-one and the lemma follows.

COROLLARY. *If K is well-defined on $R(T^*)$, then $J_\lambda = J + \lambda I$ for $\lambda > 0$ is bounded, strongly monotonic one-to-one, onto and has a Lipschitzian inverse.*

Consider now (1) with the assumption that $\Pi(\hat{u}) > 0$. We see that if K is well-defined on $R(T^*)$, then L is a well-defined operator on D. We have noted earlier that Lemma 1 remains valid with the same proof. Moreover, ϕ is a fixed point of L if and only if ϕ satisfies (3) with $\lambda > 0$ being determined by (2). When T has dense range, (4) is an alternative characteri-

zation of λ. In summary, the conditions that C is closed and K is well-defined on $R(T^*)$ suffice for the characterization of §3 to hold.

Remark. The assumption that D is a Hilbert space is also nonessential. In the above discussion it suffices that D is reflexive, rotund and smooth. For instance, D might be a suitably normed version of a finite Cartesian product of L_p and/or l_p spaces, $1 < p < \infty$. The projection of ζ onto the closest element in the closure of $R(T)$ is then of course nonlinear. Also part (iii) of Theorem 2 must read: ϕ is an extremal of $\zeta - c'$. The operator L acts on D^* and is given by $L\phi = [\zeta - T(\overline{T^*\phi})]^\wedge$, where $[\cdot]^\wedge$ denotes the extremal from D into D^*. Although the characterization involving the operator J_λ can be similarly extended, it does not appear to be fruitful to do so.

5. Some engineering applications. In [3] several optimal control problems in linear systems are modeled as minimum norm problems of the form described in §1. The results of this study generalize these problems in an obvious manner. The interested reader is referred to the reference cited for details on the physical motivation underlying this problem class.

The mathematical problems considered here relate also to certain classes of system sensitivity problems. To bring out this relationship let us consider a bounded linear system (transformation) T acting between Hilbert spaces H_1, H_2. An additive system disturbance is denoted by δT which is also a bounded linear transformation. Assume that for the nominal system T, the controller capacity (available energy) exceeds the control requirements. How can the surplus capacity be used to minimize the effects of the fluctuation, δT, in the system characteristics?

In [9] this system sensitivity problem is considered in some detail in the setting of Hilbert spaces. A control u does not exceed the controller capacity if $\| u \| \leqq k$. With $\eta \in H_2$ denoting the system task and $T^\dagger$ denoting the pseudoinverse of T, a mathematical interpretation of the physical problem is the following: for fixed $\eta \in H_2$ such that $\| T^\dagger \eta \| < k$, find $u \in T^{-1}(\eta)$ which satisfies $\| u \| \leqq k$ while minimizing $\| \delta T u \|$. In the reference cited this problem is reduced to the minimization of $\| \zeta - Vv \|$ over $\| v \| \leqq k^2 - \| T^\dagger \zeta \|^2$, where $\zeta = \delta T T^\dagger \eta$, $V = \delta T P$, P is the projection on the null space of T, and v is the solution provided $u = T^\dagger \zeta - v$ is a solution of the original problem.

The system sensitivity problem outlined above is obviously identical with the main problem considered here. The reference cited gives examples and an alternative solution of the Hilbert space problem. The present treatment not only extends the results to Banach spaces but cuts through much of the complexity of the earlier development.

Finally, it is noted that the generalization to the case where T has dense but not necessarily closed range allows applications to many distributive system problems. The interested reader is referred to [3] and [10] for further

discussion along these lines. The latter reference also discusses the case $\min \Pi = 0$.

6. Acknowledgment. The author extends his thanks to Professor James P. Williams, Department of Mathematics, University of Indiana for bringing reference [4] to his attention.

REFERENCES

[1] W. A. Porter and J. P. Williams, *A note on the minimum effort control problem*, J. Math. Anal. Appl., 13 (1966), pp. 257–264.

[2] ———, *Extensions of the minimum effort control problem*, Ibid., 13 (1966), pp. 536–549.

[3] W. A. Porter, *Modern Foundations of System Engineering*, Macmillan, New York, 1966.

[4] M. Rosenblum, *Some Hilbert space extremal problems*, Proc. Amer. Math. Soc., 16 (1965), pp. 687–691.

[5] R. Arcangeli, *Pseudo-solution de l'équation* $Ax = y$, C. R. Acad. Sci. Paris Sér. A-B, 263 (1966), pp. A282–A285.

[6] J. E. Rubio, *A fixed-point method for a minimum-norm control problem*, this Journal, 4 (1966), pp. 705–715.

[7] E. H. Zarantonello, *The closure of the numerical range contains the spectrum*, Tech. Rep. 7, Department of Mathematics, University of Kansas, Lawrence, 1964.

[8] L. Nirenberg, *Functional Analysis*, Lecture Notes 1960–1961, New York University, New York.

[9] W. A. Porter, *Improving system sensitivity via surplus controller capacity*, J. Franklin Inst., 282 (1966), pp. 366–381.

[10] ———, *Optimal control of distributive systems*, this Journal, 4 (1966), pp. 466–472.

[11] E. G. Gilbert, *A note on the fixed-point method of J. E. Rubio*, this Journal, 5 (1967), pp. 513–514.

SIAM J. CONTROL
Vol. 6, No. 2, 1968
Printed in U.S.A.

ON THE SEPARATION THEOREM OF STOCHASTIC CONTROL*

W. M. WONHAM†

1. Introduction. The object of this paper is to show that the combined problem of optimal control and filtering, for a stochastic linear dynamic system observed via a noisy linear channel, can be reduced to two independent problems of control and filtering, respectively. Under suitable conditions, solutions of the latter problems are shown to exist. This structural property of the optimal system holds whether or not the cost functional is quadratic, and whether or not the optimal feedback control happens to be linear in the system state or its expectation. In general, the optimal control depends parametrically on the intensity of channel noise; the result means, however, that channel noise plays qualitatively the same role as dynamic disturbances in determination of the feedback law.

A special result of this type, for the standard, linear stochastic regulator problem, is well known, and has been called the "separation theorem" [1], [2].

For discrete-time systems the general result can be proved by relatively straightforward application of dynamic programming [3]. In this paper attention is confined to continuous systems. The method is again dynamic programming, with appeal to the Itô-Nisio-Fleming theory of functional stochastic differential equations [4], Kalman's filter [5] and an existence theorem for parabolic equations due to Ladyjenskaya, Solonnikov and Uraltsyeva [6]. To apply the foregoing results it is necessary to impose rather stringent conditions on system coefficients. Undoubtedly the separation theorem (Theorem 2.1) is true under weaker hypotheses, more in line with requirements met in practice. In this paper our aim is to clarify some of the principles involved and to indicate the type of result to be expected.

2. Statement of the problem. The system to be controlled is described by linear stochastic differential equations:

$$\begin{aligned} dx(t) &= A(t)x(t)\,dt + b[t, u(t)]\,dt + C(t)\,dw_1(t), \qquad 0 \leqq t \leqq T, \\ x(0) &= x_0\,; \end{aligned} \tag{2.1}$$

$$\begin{aligned} dy(t) &= F(t)x(t)\,dt + G(t)\,dw_2(t), \qquad 0 \leqq t \leqq T, \\ y(0) &= 0. \end{aligned} \tag{2.2}$$

* Received by the editors December 22, 1967, and in revised form February 16, 1968.

† NASA Electronics Research Center, 575 Technology Square, Cambridge, Massachusetts 02139. This research was accomplished while the author held a National Research Council Postdoctoral Resident Research Associateship, supported by the National Aeronautics and Space Administration.

Here and below all vectors and matrices have real-valued elements. The vector $x \in R^n$ is dynamic state; u is the control vector taking values in a convex compact subset $U \subset R^m$; $y \in R^n$ is channel output; w_1, w_2 are independent, standard Wiener processes in R^{d_1}, R^{d_2}, respectively; and stochastic differentials are understood in the sense of Itô. Below, the symbol $|\cdot|$ means Euclidean norm, a prime ($'$) denotes transpose, and $\mathcal{E}$ the expectation; $c_1, c_2, \cdots$ are positive constants. An underlying probability triple $(\Omega, \mathfrak{F}, P)$ which carries x_0 and the $w_i(t)$ processes is assumed to be given; the variable $\omega \in \Omega$ will usually be suppressed from notation. If $\mathfrak{R}$ is a family of random variables, $\sigma\{\mathfrak{R}\} \subset \mathfrak{F}$ denotes the smallest σ-algebra of ω-sets relative to which $\mathfrak{R}$ is measurable.

In practical terms, the problem is to control $x(\cdot)$ in such a way as to minimize a real-valued functional

$$J[u] = \mathcal{E}\left\{\int_0^T L[t, x(t), u(t)]\, dt\right\}. \tag{2.3}$$

Control is based on the (a priori) distribution of x_0 and on information provided by the channel output $y(\cdot)$. Since the controller is not clairvoyant, $u(t)$ must be assumed to depend only on the $y(s)$ for $0 \leqq s \leqq t$. To express this *nonanticipative* dependence we introduce, following Fleming and Nisio [4], a suitable class of control functionals. Let $\mathfrak{C}$ denote the class of functions $f(t)$ continuous on $[0, T]$ with values in R^n; and write, for the *past of f at time t*,

$$(\pi_t f)(s) = \begin{cases} f(s) & \text{for } 0 \leqq s \leqq t, \\ f(t) & \text{for } t \leqq s \leqq T. \end{cases} \tag{2.4}$$

Clearly $\pi_t f \in \mathfrak{C}$ if $f \in \mathfrak{C}$. Let $\|\cdot\|$ denote sup norm in $\mathfrak{C}$ and let

$$\psi : [0, T] \times \mathfrak{C} \to U$$

be a mapping with the properties: $\psi(t, f)$ is Hölder continuous in t for each $f \in \mathfrak{C}$ and satisfies a uniform Lipschitz condition

$$|\psi(t, f) - \psi(t, g)| < c_1 \|f - g\|, \tag{2.5}$$

where $t \in [0, T]$ and $f, g \in \mathfrak{C}$. Let Ψ denote the class of functionals ψ. We call the control $u(\cdot)$ *admissible* and write $u \in \mathfrak{U}$ if

$$u(t) = \psi(t, \pi_t y), \qquad 0 \leqq t \leqq T,$$

for some $\psi \in \Psi$. The problem is *to find $u^0 \in \mathfrak{U}$ such that*

$$J[u^0] = \min\{J[u] : u \in \mathfrak{U}\}.$$

The corresponding functional ψ^0 is *optimal*. It will be verified later that $J[u]$ is well-defined.

The separation theorem states that an optimal control exists in a subclass

$\hat{\mathfrak{u}}$ of controls which depend only on the expected value of the current state given the past of y. More precisely, let

$$\mathcal{Y}_t = \sigma\{y(s), 0 \leqq s \leqq t\},$$
$$\hat{x}(t) = \mathcal{E}\{x(t) | \mathcal{Y}_t\}.$$

Write $\hat{\Psi}$ for the class of functions

$$\hat{\psi}: [0, T] \times R^n \to U$$

such that

$$|\hat{\psi}(t, \xi) - \hat{\psi}(s, \xi)| + |\hat{\psi}(t, \xi) - \hat{\psi}(t, \eta)| \leqq c_2(R)|t - s|^\alpha + c_3|\xi - \eta| \tag{2.6}$$

in every domain $0 \leqq s, t \leqq T$, $|\xi| < R$, $|\eta| < R$, where c_3 and $\alpha \in (0, \frac{1}{2})$ are independent of R. We write $u \in \hat{\mathfrak{u}}$ if

$$u(t) = \hat{\psi}[t, \hat{x}(t)], \qquad t \in [0, T],$$

for some $\hat{\psi} \in \hat{\Psi}$. It will be shown later than $\hat{\mathfrak{u}} \subset \mathfrak{u}$.

The following additional assumptions will be made. (We write u.h.c. (α) for "uniformly Hölder continuous (exponent α)," and u.l.c. for "uniformly Lipschitz continuous," where the uniformity is to hold over the whole range of the relevant arguments, unless otherwise stated. A subscript denotes differentiation. If P, Q are symmetric matrices, $P > Q$ ($P \geqq Q$) means $P - Q$ is positive (semi-)definite.)

(A.1) The matrices A, C are u.h.c. (α) in t, and F, G are continuously differentiable in $[0, T]$.

(A.2) $G(t)G(t)' \geqq c_4 I$, $t \in [0, T]$.

(A.3) $|\det [F(t)]| \geqq c_5$, $t \in [0, T]$.

(A.4) b, b_u, b_{uu} are continuous on $[0, T] \times U$ and b, b_u are u.h.c.(α) in t.

(A.5) L and L_u are bounded, u.h.c. (α) in t and u.l.c. in x. L_{uu} is bounded and continuous on $[0, T] \times R^n \times U$.

(A.6) $[b(t, u)'p + L(t, x, u)]_{uu} \geqq c_6 I$ for all $(t, x, u, p) \in [0, T] \times R^n \times U \times \{p : |p| \leqq \pi\}$, where π is defined by (6.6) below.

(A.7) x_0 is a Gaussian random variable independent of the processes $w_1(t)$, $w_2(t)$ and with positive definite covariance matrix Q_0.

The foregoing restrictions are mainly technical. Assumption (A.3) would rarely be met in practice, where typically dim $y <$ dim x; the condition is needed below to guarantee that a certain elliptic operator be nondegenerate. A square nonsingular matrix F could be constructed artificially, if necessary, by adjoining to the channel equation (2.2) a suitable term of form

$$d\tilde{y} = \epsilon \tilde{F} x\, dt + \tilde{G}\, d\tilde{w}_2.$$

If $\epsilon > 0$ is sufficiently small, then from a practical viewpoint the components $\tilde{y}$ of the observation vector contribute negligible information to the controller. However, details of such an approximation have yet to be worked out.

The number π in (A.6) is an a priori bound on the space-derivative of the solution of Bellman's equation. In the special but important case where $b(t, u)$ is linear in u, the estimate π is not required, and (A.6) can be replaced by

$$(A.6)' \quad L_{uu}(t, x, u) \geqq c_6 I, \ (t, x, u) \in [0, T] \times R^n \times U.$$

The crucial assumptions for Theorem 2.1 are that: (i) the basic dynamic equations have the form (2.1), (2.2); (ii) the (formal) perturbations dw_i/dt, $i = 1, 2$, be "white Gaussian noise"; (iii) x_0 be Gaussian and independent of the w_i; (iv) $J[u]$ be a functional additive in t.

THEOREM 2.1 (Separation theorem). *Subject to the assumptions stated, an admissible optimal control exists of the form*

$$u^0(t) = \hat{\psi}^0[t, \hat{x}(t)], \qquad t \in [0, T],$$

for some $\hat{\psi}^0 \in \hat{\Psi}$.

An optimal feedback law $\hat{\psi}^0$ is given by (6.7) below.

The theorem will be proved in several steps. In §3 we verify that the solution of (2.1), (2.2) is well-defined. Kalman's equations for $\hat{x}(t)$ are introduced in §4, and it is shown that $\hat{x}(t)$ is a diffusion process when $u = \hat{\psi} \in \hat{\Psi}$. In §5 we prove that Bellman's equation provides a sufficient condition for optimality, and in §6, that an optimal control exists. The standard linear regulator problem lies outside the scope of Theorem 2.1 and is discussed separately in §7.

3. Solution of (2.1) and (2.2). As usual, (2.1) and (2.2) are to be interpreted as stochastic integral equations, with integrals defined in the sense of Itô. Let $\psi \in \Psi$ and $f, g \in \mathcal{C}$. From (2.4), (2.5) and (A.4) it follows that

$$\begin{aligned} |\, b[t, \psi(t, \pi_t f)] - b[t, \psi(t, \pi_t g)]| &\leqq \max_{\substack{t \in [0,T] \\ u \in U}} |\, b_u(t, u)| \cdot c_7 \,\|\, \pi_t f - \pi_t g \,\| \\ &\leqq c_8 \,\|\, f - g \,\|. \end{aligned} \tag{3.1}$$

Clearly $b[t, \psi(t, \pi_t f)]$ is bounded on $[0, T] \times \mathcal{C}$. It follows by Theorem 1 of [4] that the system (2.1), (2.2) has exactly one continuous solution $\{x(t), y(t) : 0 \leqq t \leqq T\}$ with bounded second moment. Furthermore, this solution has the property that $x(t)$, $y(t)$ are measurable relative to $\sigma\{x_0, w_1(s), w_2(s), 0 \leqq s \leqq t\}$ and are independent of $\sigma\{w_i(t'') - w_i(t'), t \leqq t' \leqq t'' \leqq T, i = 1, 2\}$. By continuity and [7, p. 60, Theorem 2.5], $\{x(t), y(t)\}$ is a measurable process.

4. Representation of $\hat{x}(t)$. Suppose $b \equiv 0$. The problem of representing $\hat{x}(t)$ as the solution of a stochastic differential equation was solved formally by Kalman [5]. Since the derivation in [5] is easily made rigorous we shall justify only the extension for $b \not\equiv 0$. Let $u(\cdot)$ be admissible and write

$$\begin{aligned} \beta(t) &= b[t, u(t)] \\ &= b[t, \psi(t, \pi_t y)]. \end{aligned} \tag{4.1}$$

Observe first that the random variable $\beta(t)$ is $\mathcal{Y}_t$-measurable. Indeed, if S is open in R^m, then $\tilde{S} = \{f: b[t, \psi(t, f)] \in S\}$ is open in $\mathcal{C}$, and using the fact that $\mathcal{C}$ is a separable metric space, we see that

$$\{\omega: \pi_t y(\cdot, \omega) \in \tilde{S}\} \in \mathcal{Y}_t .$$

The assertion follows by extension to the Borel sets of R^m. Next, let

$$x(t) = \tilde{x}(t) + x^*(t),$$

where $\tilde{x}(t)$ is the diffusion process determined by

$$\begin{aligned} d\tilde{x}(t) &= A(t)\tilde{x}(t)\, dt + C(t)\, dw_1(t), \qquad t \in [0, T], \\ \tilde{x}(0) &= x_0 ; \end{aligned} \tag{4.2}$$

and $x^*(t)$ is defined by

$$\begin{aligned} \frac{dx^*(t)}{dt} &= A(t)x^*(t) + \beta(t), \qquad t \in [0, T], \\ x^*(0) &= 0. \end{aligned} \tag{4.3}$$

Since $x^*(t)$ is $\mathcal{Y}_t$-measurable there follows

$$\hat{x}(t) = \mathcal{E}\{\tilde{x}(t) \mid \mathcal{Y}_t\} + x^*(t). \tag{4.4}$$

Now define a process $\tilde{y}(t)$ according to

$$d\tilde{y}(t) \equiv dy(t) - F(t)x^*(t)\, dt \tag{4.5a}$$

$$\begin{aligned} &= F(t)\tilde{x}(t)\, dt + G(t)\, dw_2(t), \qquad t \in [0, T], \\ \tilde{y}(0) &= 0; \end{aligned} \tag{4.5b}$$

and let

$$\tilde{\mathcal{Y}}(t) = \sigma\{\tilde{y}(s), 0 \leqq s \leqq t\}.$$

By (4.3), $x^*(t)$ is $\mathcal{Y}_t$-measurable; then, by (4.5), $\tilde{y}(t)$ is $\mathcal{Y}_t$-measurable.

It will be shown that $y(t)$ is $\tilde{\mathcal{Y}}_t$-measurable, and thus that $\mathcal{Y}_t = \tilde{\mathcal{Y}}_t$. In view of (4.1) and (4.3), (4.5) can be written

$$\begin{aligned} y(t) &= \tilde{y}(t) + \int_0^t F(s)x^*(s)\, ds \\ &= \tilde{y}(t) + \phi(t, \pi_t y), \end{aligned} \tag{4.6}$$

where $\phi \in \Psi$. Since $\phi(t\ , f)$ is bounded on $[0, T] \times \mathfrak{C}$ and

$$|\phi(t, f) - \phi(t, g)| \leqq c_9 t \|f - g\|,$$

the functional equation (4.6) can be solved by successive approximations uniquely for y in $\mathfrak{C}$. Setting $y^{(0)}(t) \equiv 0$ and

$$y^{(\nu)}(t) = \tilde{y}(t) + \phi(t, \pi_t y^{(\nu-1)}), \qquad \nu = 1, 2, \cdots,$$

we see that $y^{(\nu)}(t)$ is $\tilde{\mathfrak{Y}}_t$-measurable for each ν, and the conclusion follows.

We now have, from (4.4),

$$\hat{x}(t) = \bar{x}(t) + x^*(t), \tag{4.7}$$

where

$$\bar{x}(t) = \mathcal{E}\{\tilde{x}(t) | \tilde{\mathfrak{Y}}(t)\}.$$

It remains to compute $\bar{x}(t)$. Equations (4.2) and (4.5b) have the form of (2.1) and (2.2) with $b \equiv 0$, and Kalman's result [5] applies. Introduce the conditional covariance matrix

$$\begin{aligned} Q(t) &= \mathcal{E}\{[x(t) - \hat{x}(t)][x(t) - \hat{x}(t)]' \mid \mathfrak{Y}_t\} \\ &= \mathcal{E}\{[\tilde{x}(t) - \bar{x}(t)][\tilde{x}(t) - \bar{x}(t)]' \mid \tilde{\mathfrak{Y}}_t\}, \end{aligned}$$

where the second equality holds because $x(t) = \tilde{x}(t) + x^*(t)$ and $\mathfrak{Y}_t = \tilde{\mathfrak{Y}}_t$. By the results of [5], applied to (4.2) and (4.5b), $Q(t)$ is the unique solution of the Riccati equation

$$\begin{aligned} &\frac{dQ}{dt} = AQ + QA' + CC' - QF'(GG')^{-1}FQ, \qquad t \in [0, T], \\ &Q(0) = Q_0. \end{aligned} \tag{4.8}$$

Then (see [5]) $\bar{x}$ is determined by

$$d\bar{x} = A\bar{x}\,dt + QF'(GG')^{-1}(d\tilde{y} - F\bar{x}\,dt), \qquad t \in [0, T], \tag{4.9}$$

with initial condition

$$\begin{aligned} \bar{x}(0) &= \mathcal{E}\{\tilde{x}(0) | \tilde{\mathfrak{Y}}_0\} \\ &= \mathcal{E}\{\tilde{x}(0)\} \\ &= \mathcal{E}\{x_0\}. \end{aligned}$$

Combining (4.2)–(4.5) and (4.7)–(4.9) we finally obtain

$$\begin{aligned} &d\hat{x} = A\hat{x}\,dt + \beta(t)\,dt + QF'(GG')^{-1}(dy - F\hat{x}\,dt), \quad t \in [0, T], \\ &\hat{x}(0) = \mathcal{E}\{x_0\}. \end{aligned} \tag{4.10}$$

Equation (4.10) exhibits the $\hat{x}(t)$-process as the solution of an equation "forced" by the channel output increments dy and by the control term β. It

is possible—and for later purposes necessary—to replace the differential $dy - F\hat{x}\,dt$ by the suitably scaled differential of a Wiener process. This can be justified by the observation that linear least squares estimation is equivalent to an orthogonal projection of the estimated variable on the data, but we shall use a different argument.

Define the process $z(t)$ by

$$\begin{aligned} dz &= dy - F\hat{x}\,dt \\ &= F(x - \hat{x})\,dt + G\,dw_2, \qquad t \in [0, T], \\ z(0) &= 0. \end{aligned} \tag{4.11}$$

Evidently $z(t)$ is $\mathcal{Y}_t$-measurable. The relation

$$z(t_2) = z(t_1) + \int_{t_1}^{t_2} F(t)[x(t) - \hat{x}(t)]\,dt + \int_{t_1}^{t_2} G(t)\,dw_2(t), \qquad 0 \leqq t_1 \leqq t_2 \leqq T, \tag{4.12}$$

implies that $z(\cdot)$ is continuous and that

$$\begin{aligned} \mathcal{E}\{z(t_2) \mid \mathcal{Y}_{t_1}\} &= z(t_1) + \int_{t_1}^{t_2} F(t)\mathcal{E}\{\mathcal{E}\{x(t) - \hat{x}(t) \mid \mathcal{Y}_t\}\mathcal{Y}_{t_1}\}\,dt \\ &= z(t_1), \end{aligned}$$

i.e., the $z(t)$-process is a continuous martingale relative to the $\mathcal{Y}_t$. Furthermore, one verifies easily, from (2.1), (2.2) and (4.10)–(4.12), that

$$\lim_{t_2 \downarrow t_1} (t_2 - t_1)^{-1}\mathcal{E}\{[z(t_2) - z(t_1)][z(t_2) - z(t_1)]' \mid \mathcal{Y}_{t_1}\} = G(t_1)G(t_1)'.$$

But then, by a representation theorem of Doob [7, p. 287, Theorem 3.3], there exists a Wiener process $\{\hat{w}(t) : 0 \leqq t \leqq T\}$ in R^n such that

$$dz(t) = [G(t)G(t)']^{1/2}\,d\hat{w}(t). \tag{4.13}$$

Since GG' is positive definite, the $\hat{w}(t)$-process (normalized by setting $\hat{w}(0) = 0$) is carried by $(\Omega, \mathfrak{F}, P)$; indeed (4.13) shows that $\hat{w}(t)$ is $\mathcal{Y}_t$-measurable. Combining (4.10), (4.11) and (4.13) we obtain finally

$$\begin{aligned} d\hat{x} &= A\hat{x}\,dt + \beta(t)\,dt + QF'(GG')^{-1/2}\,d\hat{w}(t), \\ \hat{x}(0) &= \mathcal{E}\{x_0\}. \end{aligned} \tag{4.14}$$

Suppose now that

$$u(t) = \hat{\psi}[t, \hat{x}(t)], \tag{4.15}$$

where $\hat{\psi} \in \hat{\Psi}$. Under the regularity conditions (2.6) and (A.4), the Itô

equation

$$\begin{aligned} d\hat{x} &= A\hat{x}\,dt + b[t, \hat{\psi}(t, \hat{x})]\,dt + QF'(GG')^{-1/2}\,d\hat{w}, \\ \hat{x}(0) &= \mathcal{E}\{x_0\}, \end{aligned} \tag{4.16}$$

determines a diffusion process on $[0, T]$. Let $\xi \in R^n$ denote a value of $\hat{x}$ and let $V: R^n \to R^1$ have continuous derivatives up to second order. The differential generator of the $\hat{x}$-process is the elliptic operator $\hat{\mathfrak{L}}(\hat{\psi})$ given by (see [8])

$$\hat{\mathfrak{L}}(\hat{\psi})V(\xi) \equiv \tfrac{1}{2}\operatorname{tr}\{\hat{C}'V_{\xi\xi}(\xi)\hat{C}\} + (A\xi + b[t, \hat{\psi}(t, \xi)])'V_\xi(\xi). \tag{4.17}$$

In (4.17), $\hat{C} = QF'(GG')^{-1/2}$ and $V_\xi(V_{\xi\xi})$ denotes the vector (matrix) of first (second) partial derivatives of V.

Next, we show that $\hat{\mathfrak{L}}$ is uniformly elliptic. Observe first that $Q(t) > 0$; in fact, (4.8) can be written

$$\frac{dQ}{dt} = \hat{A}Q + Q\hat{A}' + CC', \tag{4.18}$$

where

$$\hat{A} = A - \tfrac{1}{2}QF'(GG')^{-1}F.$$

If $\mathfrak{A}(t)$ is the matrix determined by

$$\frac{d\mathfrak{A}(t)}{dt} = \hat{A}(t)\mathfrak{A}(t), \quad \mathfrak{A}(0) = I,$$

then (4.18) and (A.7) imply

$$\begin{aligned} Q(t) &\geqq \mathfrak{A}(t)Q_0\mathfrak{A}(t)' \\ &\geqq c_{10}I, \qquad t \in [0, T]. \end{aligned}$$

This fact, combined with (A.2) and (A.3), shows that

$$\hat{C}(t)\hat{C}(t)' \geqq c_{11}I, \qquad t \in [0, T]. \tag{4.19}$$

We verify now that controls of the class $\hat{\mathfrak{U}}$, i.e.,

$$u(t) = \hat{\psi}[t, \hat{x}(t)], \qquad \hat{\psi} \in \hat{\Psi},$$

are admissible. In view of (2.6) it is enough to check that $\hat{x}(t)$ is a uniformly Lipschitz-continuous functional of $\pi_t y$. By (A.1) and (4.8) the matrix

$$K(t) = Q(t)F(t)'[G(t)G(t)']^{-1}$$

is continuously differentiable in $[0, T]$, so that

$$\int_0^t K(s)\,dy(s) = K(t)y(t) - \int_0^t \frac{dK(s)}{ds}\,y(s)\,ds. \tag{4.20}$$

Clearly the right side of (4.20) is a continuous functional of $\pi_t y$. Consider now (4.10) with $\beta(t) = b[t, \hat{\psi}(t, \hat{x}(t))]$. Integrating (4.10) and substituting (4.20), one obtains for $\hat{x}(\cdot)$ a Volterra equation with kernel uniformly Lipschitz in the variable $\hat{x}$. From this it is easy to establish the required continuity in $\mathcal{C}$ of the mapping $\pi_t y(\cdot) \to \hat{x}(\cdot)$.

To conclude this section, we verify that the conditional distribution of $x(t)$ given $\mathcal{Y}_t$ is Gaussian and that, if $0 \leqq t_1 \leqq t_2 \leqq t_3 \leqq T$, the increments $\hat{w}(t_3) - \hat{w}(t_2)$ are independent of $\mathcal{Y}_{t_1}$. These facts are crucial to the later development. To prove the first statement recall that $x(t) = \tilde{x}(t) + x^*(t)$, where $x^*(t)$ is $\mathcal{Y}_t$-measurable. Furthermore, by the linearity of (4.2) and (4.5b), the conditional distribution of $\tilde{x}(t)$ given $\tilde{\mathcal{Y}}_t$ is Gaussian. Since $\tilde{\mathcal{Y}}_t = \mathcal{Y}_t$ the assertion follows. Next, the relation $e(t) \equiv x(t) - \hat{x}(t)$ shows that the conditional distribution of $e(t)$ given $\mathcal{Y}_t$ is Gaussian, with covariance $Q(t)$ depending only on t. Also, by (2.1) and (4.1),

$$de = [A - \hat{C}(GG')^{-1/2}F]e\,dt + C dw_1 - \hat{C}(GG')^{-1/2}G dw_2\,,$$

$$e(0) = x_0 - \mathcal{E}\{x_0\}. \qquad t \in [0, T], \tag{4.21}$$

Combining (4.12) and (4.21) we see that $z(t_2) - z(t_1)$ can be written as a linear functional of $e(t_1)$ and the $w_i(s)$ increments, $i = 1, 2$, for $t_1 \leqq s \leqq t_2$. This implies that the conditional distribution of $z(t_3) - z(t_2)$ given $\mathcal{Y}_{t_1}$ is Gaussian, with zero mean, and covariance independent of $\mathcal{Y}_{t_1}$; hence the same is true of $\hat{w}(t_3) - \hat{w}(t_2)$, and the independence assertion follows. Finally, if the Itô equation (4.16) holds, then $\hat{x}(t_2)$ is measurable relative to the sample space of $\hat{x}(t_1)$ and the $\hat{w}(s)$ increments for $t_1 \leqq s \leqq t_2$. Thus if $f: R^n \to R^1$ is an arbitrary bounded measurable function, then

$$\mathcal{E}\{f[\hat{x}(t_2)] \mid \mathcal{Y}_{t_1}\} = \mathcal{E}\{f[\hat{x}(t_2)] \mid \hat{x}(t_1)\} \tag{4.22}$$

with probability 1.

5. A sufficient condition for optimality. Let $\mathcal{G}(x;\, t,\, \xi)$ be the Gaussian probability density in R^n with mean ξ and covariance matrix $Q(t)$:

$$\mathcal{G} = (2\pi)^{-n/2}[\det Q(t)]^{-1/2} \exp\,[-\tfrac{1}{2}(x - \xi)'Q(t)^{-1}(x - \xi)]. \tag{5.1}$$

By the results of §4, if u is a fixed vector of U, then

$$\begin{aligned}\hat{L}(t, \xi, u) &\equiv \mathcal{E}\{L[t, x(t), u] \mid \hat{x}(t) = \xi\} \\ &= \int_{R^n} L(t, x, u)\mathcal{G}(x; t, \xi)\,dx.\end{aligned}$$

It is verified in §6 that $\hat{L}$ satisfies the conditions imposed on L in (A.5). On this assumption we establish the following optimality criterion (cf. [9] and [10]).

LEMMA 5.1 (Optimality criterion). *Suppose there exist an element* $\hat{\psi}^0 \in \hat{\Psi}$ *and a function* $V:[0, T] \times R^n \to R^1$ *such that*

(i) $V, V_t, V_\xi, V_{\xi\xi}$ *are continuous, and*

$$|V| + |V_t| + |\xi||V_\xi| + |V_{\xi\xi}| \leqq c_{12}(1 + |\xi|^2), \tag{5.1}$$

(ii)

$$0 = V_t(t, \xi) + \hat{\mathfrak{L}}(\hat{\psi}^0)V(t, \xi) + \hat{L}[t, \xi, \hat{\psi}^0(t, \xi)] \tag{5.2a}$$

$$\leqq V_t(t, \xi) + \hat{\mathfrak{L}}(u)V(t, \xi) + \hat{L}(t, \xi, u) \tag{5.2b}$$

for all $(t, \xi, u) \in [0, T] \times R^n \times U$, *and*

$$V(T, \xi) = 0, \qquad \xi \in R^n. \tag{5.2c}$$

Then the control $u = \hat{\psi}^0$ *is optimal in* $\mathfrak{U}$.

For the proof, introduce the random variable

$$W(t) = \mathcal{E}\left\{\int_t^T L[s, x(s), \psi^0(s, \hat{x}(s))\, ds \mid \mathcal{Y}_t\right\},$$

where $x(t)$ is the solution of (2.1) with $u(t) = \hat{\psi}^0[t, \hat{x}(t)]$ and $\hat{x}(t)$ is given by (4.16) with $\hat{\psi} = \hat{\psi}^0$. Now

$$\begin{aligned} W(t) &= \mathcal{E}\left\{\int_t^T \mathcal{E}\{L[s, x(s), \hat{\psi}^0(s, \hat{x}(s))] \mid \mathcal{Y}_s\}\, ds \mid \mathcal{Y}_t\right\} \\ &= \mathcal{E}\left\{\int_t^T \hat{L}[s, \hat{x}(s), \hat{\psi}^0(s, \hat{x}(s))]\, ds \mid \mathcal{Y}_t\right\} \qquad (5.3) \\ &= \mathcal{E}\left\{\int_t^T \hat{L}[s, \hat{x}(s), \hat{\psi}^0(s, \hat{x}(s))]\, ds \mid \hat{x}(t)\right\}, \end{aligned}$$

where we have used (4.22). By (5.1), (5.2a), (5.3) and Itô's integration formula (see [11]),

$$\begin{aligned} W(t) &= -\mathcal{E}\left\{\int_t^T (V_t[s, \hat{x}(s)] + \hat{\mathfrak{L}}(\hat{\psi}^0)V[s, \hat{x}(s)])\, ds \mid \hat{x}(t)\right\} \\ &= V[t, \hat{x}(t)]. \end{aligned}$$

In particular,

$$V[0, \hat{x}(0)] = J[\hat{\psi}^0]. \tag{5.4}$$

To show that $\hat{\psi}^0$ is optimal, let $u(t)$ be an arbitrary control

$$u(t) = \psi(t, \pi_t y),$$

where $\psi \in \Psi$; and now write $x(t)$, $\hat{x}(t)$ for the corresponding solution of (2.1) and (4.10). Since the moment $\mathcal{E}\{|x(t)|^2\}$ is integrable on $[0, T]$ (see

[4]) and since (5.1) holds, we may again apply Itô's integration formula to obtain

$$\begin{aligned}
\mathcal{E}\{V[t, \hat{x}(t)] \mid \mathfrak{Y}_t\} &= -\mathcal{E}\left\{\int_t^T [V_s[s, \hat{x}(s)] + \hat{\mathfrak{L}}(\psi)V[s, \hat{x}(s)]]\, ds \mid \mathfrak{Y}_t\right\} \\
&\leqq -\mathcal{E}\left\{\int_t^T [V_s[s, \hat{x}(s)] + \hat{\mathfrak{L}}(\hat{\psi}^0)V[s, \hat{x}(s)]\right. \\
&\quad \left. + \hat{L}[s, \hat{x}(s), \hat{\psi}^0(s, \hat{x}(s))] - \hat{L}[s, \hat{x}(s), \psi(s, \pi_s y)]]\, ds \mid \mathfrak{Y}_t\right\} \qquad (5.5) \\
&= \mathcal{E}\left\{\int_t^T \hat{L}[s, \hat{x}(s), \psi(s, \pi_s y)]\, ds \mid \mathfrak{Y}_t\right\} \\
&= \mathcal{E}\left\{\int_t^T L[s, x(s), \psi(s, \pi_s y)]\, ds \mid \mathfrak{Y}_t\right\}.
\end{aligned}$$

Here the inequality results from (5.2b) with $u = \psi$ in the right side, and the last equality follows as in (5.3). Setting $t = 0$ in (5.5) and using (2.3) and (5.4), we obtain

$$J[\hat{\psi}^0] \leqq J[\psi].$$

This inequality states that $\hat{\psi}^0$ is optimal.

It remains to prove that $\hat{\psi}^0$ and V exist. We shall do this by solving Bellman's equation.

6. Solution of Bellman's equation. Observe that (5.2) is formally equivalent to Bellman's functional equation

$$\min_{u \in U} [V_t(t, \xi) + \hat{\mathfrak{L}}(u)V(t, \xi) + \hat{L}(t, \xi, u)] = 0, \qquad (6.1a)$$

$$(t, \xi) \in [0, T] \times R^n,$$

$$V(T, \xi) = 0. \qquad (6.1b)$$

The minimization in (6.1a) is to be done at each fixed $(t, \xi) \in [0, T] \times R^n$. It will be shown first that this is possible. Write $V_\xi = p$ and consider the function

$$\lambda(t, \xi, p, u) = b(t, u)'p + \hat{L}(t, \xi, u) \qquad (6.2)$$

defined for $(t, \xi, p, u) \in [0, T] \times R^n \times \{p : |p| \leqq \pi\} \times U$.

We shall verify that $\hat{L}$ satisfies the conditions imposed on L in (A.5). Clearly $\hat{L}$ is bounded. From the elementary relations

$$\begin{aligned}
|e^a - e^b| &\leqq \tfrac{1}{2}|a - b|(e^a + e^b), \\
|a^{-1/2} - b^{-1/2}| &= (ab)^{-1/2}(a^{1/2} + b^{1/2})^{-1}|a - b|, \quad a, b > 0, \\
|ac - bd| &\leqq |c||a - b| + |b||c - d|,
\end{aligned}$$

together with the fact that

$$c_{12}I \leqq Q(t) \leqq c_{13}I, \qquad t \in [0, T],$$

and the continuity of dQ/dt, there results

$$|\mathcal{G}(x; t_1, \xi_1) - \mathcal{G}(x; t_2, \xi_2)|$$
$$\leqq c_{14}(|\xi_1 - \xi_2| + |t_1 - t_2|)(\exp(-c_{15}|x - \xi|^2) + \exp(-c_{15}|x - \xi_2|^2)).$$

The assertion now follows by simple computations.[1] Henceforth (Â.5) denotes (A.5) with L, x replaced by $\hat{L}$, ξ.

Next, observe that the inequality in (A.6) can be integrated over $x \in R^n$ with respect to $\mathcal{G}$ to yield:

(Â.6) $\quad [b(t, u)'p + \hat{L}(t, \xi, u)]_{uu} \geqq c_{16}I.$

By virtue of (Â.5), (Â.6), the problem of minimizing with respect to u the function λ in (6.2) has the following solution.

LEMMA 6.1. *There exists a (unique) function $\mu(t, \xi, p)$ with values in U such that*

(i) $\lambda[t, \xi, p, \mu(t, \xi, p)] \leqq \lambda(t, \xi, p, u)$ *for all* $(t, \xi, p, u) \in [0, T] \times R^n \times \{p : |p| \leqq \pi\} \times U$,

(ii) μ *is* u.h.c. (α) *in* t *and is* u.l.c. *in* (ξ, p) *in the domain* $(t, \xi, p) \in [0, T] \times R^n \times \{p : |p| \leqq \pi\}$.

This result is due to Fleming [12, Lemma 2.1].

Write

$$\Lambda(t, \xi, p) = \xi' A(t)'p + \lambda[t, \xi, p, \mu(t, \xi, p)].$$

With the substitution $u = \mu$ in (5.2) we obtain the semilinear parabolic equation

$$V_t(t, \xi) + \tfrac{1}{2} \operatorname{tr}\{\hat{C}(t)' V_{\xi\xi}(t, \xi)\hat{C}(t)\} + \Lambda[t, \xi, V_\xi(t, \xi)] = 0,$$

$$(t, \xi) \in [0, T] \times R^n, \tag{6.3}$$

$$V(T, \xi) = 0, \qquad \xi \in R^n.$$

It remains to verify that the Cauchy problem (6.3) has a suitably smooth solution V. This conclusion follows by a theorem of Ladyjenskaya, Solonnikov, and Uraltsyeva [6, p. 564, Theorem 8.1]. For ease of reference we check the hypotheses of the theorem in detail (page numbers refer to [6]).

(i) Condition (b), p. 564: by (6.2) and boundedness of $\hat{L}$,

$$\Lambda(t, \xi, 0)V = \hat{L}[t, \xi, \mu(t, \xi, 0)]V \leqq c_{17}V^2 + c_{18}$$

for all $(t, \xi, V) \in [0, T] \times R^n \times R^1$.

[1] These show that the Lipschitz condition in x on L and L_u could be relaxed.

(ii) Condition (b), p. 513: by (A.1), (A.4), ($\hat{\text{A}}$.5) and Lemma 6.1, Λ is continuous in (t, ξ, p); as shown in §4, $\hat{\mathfrak{L}}$ is uniformly elliptic; and it is clear that $\hat{C}\hat{C}'$ is bounded on $[0, T]$. Finally,

$$|\Lambda(t, \xi, p)| \leqq c_{19}(R)(1 + |p|)$$

in every cylinder $t \in [0, T]$, $|\xi| \leqq R$.

(iii) Condition (c), p. 513: by (A.1), (A.4), ($\hat{\text{A}}$.5) and Lemma 6.1, $\Lambda(t, \xi, p)$ is u.h.c. (α) in t, in every domain $t \in [0, T]$, $|\xi| \leqq R$, $|p| \leqq \pi$, and is u.l.c. in ξ (respectively p) in the domain $t \in [0, T]$, $|p| \leqq \pi$ (respectively $t \in [0, T]$, $|\xi| \leqq R$). By (A.1) and (4.8) it is clear also that $\hat{C}(t)\hat{C}(t)'$ is Hölder (even Lipschitz) continuous on $[0, T]$.

It follows from [6, Theorem 8.1] that (6.3) has a solution $V(t, \xi)$, defined and bounded for $(t, \xi) \in [0, T] \times R^n$, such that V, V_t, V_ξ and $V_{\xi\xi}$ are u.h.c.(α) in t and u.h.c.(2α) in ξ, in every finite cylinder $t \in [0, T]$, $|\xi| \leqq R$. We shall show that V_ξ is actually u.l.c. in ξ on $[0, T] \times R^n$. It is enough to show that $V_{\xi\xi}$ is bounded. To this end, introduce the change of variables

$$\eta = S(t)\xi, \quad t = t,$$

where S is determined by

$$\frac{dS(t)}{dt} = -S(t)A(t), \qquad t \in [0, T], \tag{6.4}$$

$$S(0) = I.$$

Setting $\tilde{V}(t, \eta) = V(t, S(t)^{-1}\eta)$, we obtain

$$\tilde{V}_t + \tfrac{1}{2} \operatorname{tr} \{\hat{C}(t)'S(t)'\tilde{V}_{\eta\eta}S(t)\hat{C}(t)\} + \tilde{\Lambda}(t, \eta, \tilde{V}_\eta) = 0, \quad t \in [0, T], \tag{6.5a}$$

$$\tilde{V}(T, \eta) = 0, \tag{6.5b}$$

where

$$\tilde{\Lambda}(t, \eta, \tilde{p}) = \lambda[t, \xi, p, \mu(t, \xi, p)]|_{\substack{\xi = S^{-1}\eta \\ p = S'\tilde{p}}}.$$

Observe first that $\tilde{\Lambda}(t, \eta, \tilde{p})$ is u.h.c. (α) in t and u.l.c. in $(\eta, \tilde{p})$, in the domain

$$(\eta, \tilde{p}) \in R^n \times S(t)'^{-1}\{p : |p| \leqq \pi\}, \qquad 0 \leqq t \leqq T.$$

The operator in (6.5) is uniformly parabolic on $[0, T] \times R^n$. Consider the fundamental solution corresponding to the operator defined by the first two terms in (6.5a). By direct calculation it follows, since $\tilde{\Lambda}$ is bounded, that $\tilde{V}_\eta$ is bounded, and π can be chosen a priori such that

$$|V_\xi(t, \xi)| \leqq \pi, \qquad (t, \xi) \in [0, T] \times R^n. \tag{6.6}$$

Again, boundedness of $\tilde{\Lambda}$ implies that $\tilde{V}_\eta(t, \eta)$ is u.h.c. in η on $[0, T] \times R^n$ (cf. [13, p. 193, Lemma 2]). Hence $\tilde{\Lambda}[t, \eta, \tilde{V}_\eta(t, \eta)]$ is u.h.c. on $[0, T] \times R^n$. But then one sees easily (cf. [6, p. 308]) that $\tilde{V}_{\eta\eta}$ is uniformly bounded; hence so is $V_{\xi\xi}$.

With $V_\xi(t, \xi)$ u.l.c. in ξ and continuous in t, we have that

$$\hat{\psi}^0(t, \xi) \equiv \mu[t, \xi, V_\xi(t, \xi)] \tag{6.7}$$

enjoys the same properties, i.e., $\hat{\psi}^0 \in \hat{\Psi}$.

Returning to (6.5) and using as before the fundamental solution for the linear part, one obtains that $\tilde{V}_t$ is bounded, so that

$$|V_t| = |\tilde{V}_t - \xi' A' S' \tilde{V}_\eta| \leqq c_{20}(1 + |\xi|).$$

It is now clear that $V(t, \xi)$ and $\hat{\psi}^0(t, \xi)$ satisfy the conditions of Lemma 5.1. The proof of Theorem 2.1 is complete.

7. Linear regulator. If Bellman's equation can be solved explicitly for functions V and $\hat{\psi}^0$ which satisfy the hypotheses of Lemma 5.1, then, of course, many of the restrictive conditions imposed in the general discussion become irrelevant. A well-known example is the following (cf. [1], [10]). In (2.1) let

$$b[t, u(t)] = B(t)u(t),$$

let u range over R^m, and let

$$L(t, x, u) = x'M(t)x + u'N(t)u,$$

where $M(t)$ and $N(t)$ are respectively positive semidefinite and positive definite, with $N(t)^{-1}$ bounded on $[0, T]$. In this case Bellman's equation (6.3) has a quadratic solution

$$V(t, \xi) = \xi' P(t)\xi + p(t),$$

where P is the (unique) solution of a certain matrix Riccati equation and $p(t)$ is given by a quadrature. The optimal control is then

$$\hat{\psi}^0(t, \xi) = -N(t)^{-1}B(t)'P(t)\xi.$$

Here P, and hence $\hat{\psi}^0$, are actually independent of the channel coefficient matrices F, G. For this solution of (6.3) to exist it is sufficient, with the stated conditions on M and N, that all parameter matrices be piecewise continuous and that (A.2) hold.

REFERENCES

[1] P. D. Joseph and J. T. Tou, *On linear control theory*, AIEE Trans. Applications and Industry, 80 (1961), pp. 193–196

[2] J. E. Potter, *A guidance-navigation separation theorem*, Rep. Re-11, Experi-

mental Astronomy Laboratory, Massachusetts Institute of Technology, Cambridge, 1964.

[3] C. STRIEBEL, *Sufficient statistics in the optimum control of stochastic systems*, J. Math. Anal. Appl., 12 (1965), pp. 576–592.

[4] W. H. FLEMING AND N. NISIO, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.

[5] R. E. KALMAN, *New methods in Wiener filtering theory*, Proc. First Symposium on Engineering Applications of Random Function Theory and Probability, John Wiley, New York, 1963, pp. 270–388.

[6] O. A. LADYJENSKAYA, V. A. SOLONNIKOV AND N. N. URALTSYEVA, *Linear and Quasilinear Equations of Parabolic Type*, Izd.-vo "Nauka," Moscow, 1967.

[7] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[8] E. B. DYNKIN, *Markov Processes*, Academic Press, New York, 1965.

[9] N. N. KRASOVSKII, *On optimal control in the presence of random disturbances*, J. Appl. Math. Mech., 24 (1960), pp. 82–102.

[10] W. M. WONHAM, *Stochastic problems in optimal control*, Tech. rep. 63-14, Research Institute for Advanced Studies, Baltimore, 1963.

[11] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, Izd.-vo "Nauka," Moscow, 1965.

[12] W. H. FLEMING, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279.

[13] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

REGULATION OF INCOMPLETELY IDENTIFIED LINEAR SYSTEMS*

A. CHANG AND J. RISSANEN†

Abstract. This paper is concerned with regulation of linear systems disturbed by stationary Gaussian processes. Neither the noise nor system characteristics are assumed to be known a priori, but the input and output signals can be observed for the purpose of identification.

The problem is to find a feedback law, a linear function of finitely many past observations, which minimizes an appropriate objective function measuring the goodness of the regulation.

Precise conditions under which the problem can be solved with a stable control law are stated, and an algorithm for finding the solution with arbitrary accuracy is given.

1. Introduction. This paper is concerned with the problem of regulating linear systems which are disturbed by stationary Gaussian random processes (see Fig. 1). The variables u, y and v are scalar-valued and defined only at discrete times $t \in T = \{0, \pm 1, \pm 2, \cdots\}$. Both the input u and the output y—but not v, a stationary Gaussian r.p.—are assumed to be directly observable quantities. The regulation problem is to keep the output as near as possible, in some sense, to a prescribed constant y_d, say, $y_d = 0$, by suitable manipulation of the input u_t. The input at time t, u_t, is of the form

$$u_t = u_t' + w_t, \quad t \in T, \tag{1.1}$$

where u_t' is the component of the input that we can choose, and w_t is a noise term. The noise takes into account the deficiencies in the physical device needed to implement the desired control process u'. The control u_t' at each $t \in T$ is to be computed as a function of the past observations

$$z_{t,n} = (y_t, u_{t-1}, y_{t-1}, \cdots, u_{t-n+1}), \tag{1.2}$$

where n is a fixed positive integer or $+\infty$.

The difficulty of the regulation problem depends on what assumptions can be made about the process; and this is influenced by the amount of information we have initially, or can obtain about it. Wishing to study the problem for a wide enough class of systems to include commonly encountered industrial processes, we want to impose as few restrictive assumptions as possible. We shall assume that the system is linear and time-invariant, and that the disturbance is a stationary Gaussian process. Thus, in particular, the system is not required to be of finite order; when the structure of the system is unknown, such an assumption is often difficult to

* Received by the editors July 19, 1967, and in revised form November 13, 1967.
† IBM Research Laboratory, San Jose, California 95114.

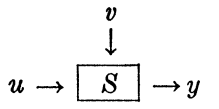


FIG. 1

justify.[1] The assumption of linearity, however, is usually justified, because if regulation is possible at all, both y and u tend to remain near their nominal operating values. In practice, the assumption that the disturbance is Gaussian is harder to justify, but the material presented here is useful even in the non-Gaussian case if only second order statistical properties are considered.

The very fact that we permit processes of infinite order introduces interesting conceptual problems which also are of practical importance. The usual approach of first determining a finite parameter model of the system and then designing a control law on the basis of that becomes questionable. Indeed, since the model cannot, in general, exactly describe the system, the value of a control law so determined is not immediately clear. Furthermore, while the parameters in the model can be chosen to fit the observations for a particular input process, once a new control law is implemented, the input process will be different. Therefore, the parameters chosen may no longer be optimal for the closed loop system obtained with a control law determined on the basis of this model.

In this paper, we study these questions for a particular control problem. We shall show that the solution of the "finite memory" versions of the problem is expressible in terms of a finite number of parameters. These parameters constitute a natural model for each such problem: they define a model which is optimal for the closed loop system obtained when the optimal control law for the model itself is implemented. Moreover, these parameters can be determined with arbitrary accuracy by sufficiently long observations of the input and output processes, so that the problem of determining a control law can be solved for an a priori unknown system without identifying the exact characteristics of the system, which would be impossible for infinite order systems.

The control problem to be studied is a mathematical optimization problem. We consider the quantity

$$L_{t+1} = y_{t+1}^2 + \rho u_t^2, \qquad \rho \geqq 0, \tag{1.3}$$

as the measure of the deviation from perfect regulation at time $t + 1$. Let

$$J(u_t', z_{t,n}) = E(L_{t+1} \mid z_{t,n}) \tag{1.4}$$

[1] Alternative approaches to the regulation problem which make use, in an essential manner, of the assumed finite order structure of the process are discussed in [1], [2] and [3].

denote the conditional expectation of L_{t+1} given $z_{t,n}$. J may be interpreted as the expected loss at time $t+1$ given $z_{t,n}$. The control u_t' is to be computed as a function of $z_{t,n}$.

DEFINITION 1.1. A control u_t' is *admissible* if it is a function only of $z_{t,n}$. The set of admissible controls is $U_{t,n}$, and an *admissible control law* is any algorithm for computing u_t' from the data $z_{t,n}$.

The problem to be considered is essentially the following one (see §3 for a complete statement).

PROBLEM. For each $t \in T$, choose $u_t' \in U_{t,n}$ such that

$$J(u_t', z_{t,n}) = \min_{v \in U_{t,n}} J(v, z_{t,n}). \tag{1.5}$$

To complete the problem formulation, we have to specify S, v and w. This is done in the next two sections.

2. Mathematical preliminaries. This section is a short description of the material concerning convolutions, z-transforms and stationary random processes pertinent to the subsequent discussion.

Let $T = \{0, \pm 1, \pm 2, \cdots\}$ and l be the set of all real-valued sequences $f = (\cdots, f_{-1}, f_0, f_1, \cdots)$ on T which satisfy the inequality $\sum_i |f_i| < \infty$. An important subset of l is l^+, the set of all $f \in l$ which satisfy the condition $f_i = 0$ for all $i < 0$. Elements of l^+ will usually be written as $f = (f_0, f_1, \cdots)$, we omit the zero components corresponding to negative indices. With the usual linear vector operations and $\|f\| = \sum_i |f_i|$ as the norm, l and l^+ are both Banach spaces.

If $f, g \in l\ (l^+)$, the convolution of f and g, denoted $h = f * g$, is the element of $h \in l\ (l^+)$ defined by

$$h_n = \sum_i f_{n-i} g_i, \qquad n \in T. \tag{2.1}$$

The norms of f, g and h satisfy the inequality

$$\|h\| \leqq \|f\| \|g\|. \tag{2.2}$$

With convolution as the product, both l and l^+ are commutative Banach algebras (see [4]).

The element $e \in l\ (l^+)$ defined by

$$e_n = \begin{cases} 0 & \text{if } n \neq 0, \\ 1 & \text{if } n = 0, \end{cases} \qquad n \in T, \tag{2.3}$$

is the identity in $l\ (l^+)$: $e * f = f$ for all $f \in l\ (l^+)$. An element $f \in l\ (l^+)$ is said to be a *unit* or *regular* if there exists $g \in l\ (l^+)$ such that $f * g = e$. Such a g is unique and is called the inverse of f and is written f^{-1}. The (real) spectrum of $f \in l\ (l^+)$, $\sigma(f)$, is the set of real numbers λ for which $f - \lambda e$ has no inverse in $l\ (l^+)$.

The study of convolutions and regularity can occasionally be simpler with z-transforms. Let $f \in l$ and let D_f be the set of all complex numbers z such that $\sum_i |f_i z^i| < \infty$. The z-transform of f is the complex-valued function $\hat{f}$ defined by

$$\hat{f}(z) = \sum_i f_i z^i, \qquad z \in D_f. \tag{2.4}$$

Since $f \in l$, D_f contains the unit circle $C = \{z: |z| = 1\}$. If moreover, $f \in l^+$, D_f contains the disk $D = \{z: |z| \leqq 1\}$. Given the z-transform of an element f of l, we can recover f by the inversion formula

$$f_n = \frac{1}{2\pi i} \int_C \hat{f}(z) z^{-n-1}\, dz, \qquad n \in T. \tag{2.5}$$

A connection between convolutions, regularity, and z-transforms is given in the following two theorems, which we state without proof.

THEOREM 2.1. *Let $f, g \in l$ and let $h = f * g$. Then $\hat{h}(z) = \hat{f}(z)\hat{g}(z)$ for all $z \in D_f \cap D_g$.*

THEOREM 2.2 (Wiener). *Let $f \in l$ (l^+). Then f is a unit if and only if $\hat{f}(z) \neq 0$ for all $z \in C(D)$.*

The following linear maps of l into l will be used frequently in the subsequent discussion. In their definitions, f is any element of l.

DEFINITION 2.1. The projections E_n and E_+:

$$(E_n f)_i = \begin{cases} f_i & \text{for } 0 \leqq i \leqq n-1, \\ 0 & \text{otherwise;} \end{cases} \tag{2.6}$$

$$(E_+ f)_i = \begin{cases} f_i & \text{for } i \geqq 0, \\ 0 & \text{for } i < 0. \end{cases} \tag{2.7}$$

$E_+ f$ is called the *positive* part of f. We also write $(f)_+$ instead of $E_+ f$.

DEFINITION 2.2. The conjugate of f, denoted f^*:

$$(f^*)_n = f_{-n}, \qquad n \in T. \tag{2.8}$$

DEFINITION 2.3. The unit shift operator U^{-1}:

$$(U^{-1} f)_n = f_{n+1}, \qquad n \in T. \tag{2.9}$$

We construct next a class of random processes. Let $w = \{w_n\}$, $n \in T$, be a sequence of independent Gaussian random variables with

$$\begin{aligned} E(w_n) &= 0, \\ E(w_n w_m) &= \delta_{nm}, \end{aligned} \qquad n, m \in T, \tag{2.10}$$

where $E(x)$ denotes the expectation of x. If $a \in l^+$, the sum

$$v_n = \sum_i a_i w_{n-i}, \qquad n \in T, \tag{2.11}$$

exists in the mean square sense and defines a Gaussian random variable [5].

We define the convolution of a and w to mean the random process $v = a * w$, where v is defined by (2.11). Using (2.1) and (2.11), we can easily show that the covariance function of v is

$$\begin{aligned} r_v(m) &= E(v_n v_{n+m}) \\ &= \sum a_i a_{m+i} \\ &= (a^* * a)_m, \qquad m \in T, \end{aligned} \tag{2.12}$$

where a^* is the conjugate of a. Since $r_v(m)$ is independent of n, v is a stationary process; in addition, $r_v \in l$.

The notations in later sections are simplified by the use of matrices of l-elements or a random process of type (2.11). Only 2×2 matrices and 2-vectors are needed. The 2×2 matrices of elements in l form a noncommutative Banach algebra, if the sum is defined in the usual fashion and the product as the convolution:

$$(A * B)_n = \sum A_{n-i} B_i, \qquad n \in T. \tag{2.13}$$

In (2.13), $A_{n-i}B_i$ is the ordinary product of the 2×2 matrices A_{n-i} and B_i. We use the norm $\| A \| = \max_{i=1,2} \{ \| a_{i1} \| + \| a_{i2} \| \}$, and in analogy with (2.2), have the inequality

$$\| A * B \| \leqq \| A \| \, \| B \|. \tag{2.14}$$

The previously given definitions extend in an obvious way to matrices. In particular, the z-transform of a matrix is the matrix of the z-transformed elements. Similarly, if Q is any one of the operators (2.6)–(2.9) and $A = (a_{ij})$, then $QA = (Qa_{ij})$. The special matrix $\begin{pmatrix} x^1 & 0 \\ x^2 & 0 \end{pmatrix}$ is written as the 2-vector $x = \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}$, and thus the convolution $y = A * x$ yields a 2-vector $y = \begin{pmatrix} y^1 \\ y^2 \end{pmatrix}$, for which, in particular, (2.14) applies.

Analogously, if w^1 and w^2 are processes of the type (2.10), the convolution of a 2×2 matrix A with $w = \begin{pmatrix} w^1 \\ w^2 \end{pmatrix}$ is the vector random process:

$$A * w = \begin{pmatrix} a_{11} * w^1 + a_{12} * w^2 \\ a_{21} * w^1 + a_{22} * w^2 \end{pmatrix}. \tag{2.15}$$

3. Regulation problem. This section contains a precise statement of our regulation problem and a characterization of its solution in terms of a prediction formula. This characterization will be examined in detail in subsequent sections.

The system that we want to control is assumed to be linear in the sense that y, u and v satisfy the equations

$$\begin{aligned} y &= a * u + v, \\ v &= b * w^1, \end{aligned} \tag{3.1}$$

where

(a) $a \in l^+$ and $a_0 = 0$,
(b) $b \in l^+$ and is a unit,
(c) w^1 is a sequence of independent random variables with

$$E(w_n^1) = 0, \quad E(w_n^1 w_m^1) = \delta_{nm}$$

for all $n, m \in T$.

Physically, (3.1) says that the output is the sum of a term depending linearly on the input u and a stationary Gaussian random process.

The quantity u appearing in (3.1) is generated by some linear function of the past $2n - 1$ observations $z_{t,n}$ (see (1.2)) and is corrupted by noise (1.1):

$$u_t = \sum_0^{n-1} h_i^1 y_{t-i} + \sum_0^{n-1} h_i^2 u_{t-i} + \sigma w_t^2, \quad \sigma \neq 0. \tag{3.2}$$

In (3.2), w^2 is independent of w^1 but otherwise has the properties (c) of (3.1). Let

$$\begin{aligned} h^1 &= (h_0^1, h_1^1, \cdots, h_{n-1}^1, 0, \cdots), \\ h^2 &= (0, h_1^2, \cdots, h_{n-1}^2, 0, \cdots). \end{aligned} \tag{3.3}$$

We call the control law (3.2) stable if $e - h^2$ is a unit. The reason for this terminology is that the control law defined by (3.2) may be put in the form

$$u = c * y + d * w^2, \tag{3.4}$$

where

$$\begin{aligned} c &= h^1 * (e - h^2)^{-1}, \\ d &= \sigma(e - h^2)^{-1}, \end{aligned} \tag{3.5}$$

when $e - h^2$ is a unit, and the linear system (3.4) is stable in the input-output sense [6]. We also say then that c is stable.

The following proposition shows when the pair of equations (3.1) and

(3.4) define meaningful random processes. Let

$$x = \begin{pmatrix} y \\ u \end{pmatrix} \quad \text{and} \quad w = \begin{pmatrix} w^1 \\ w^2 \end{pmatrix}.$$

PROPOSITION 3.1. *Suppose* $\Delta = e - a * c$ *is a unit and* c *is stable. Then the following three properties hold*:

(i) *The pair of equations* (3.1) *and* (3.4) *define a vector random process*

$$x = A * w,$$

where

$$A = \Delta^{-1} * \begin{pmatrix} b & a * d \\ b * c & d \end{pmatrix}.$$

(ii) *The covariance matrices of* x *are given by*

$$R(m) = E(x_n x'_{n+m}) = \sum_i A_i A'_{m+i},$$

where the prime denotes transposition; or equivalently, by the convolution equation

$$R = A^* * A',$$

where A^* *is the conjugate of* A.

(iii) A *has an inverse* (*i.e.*, $A^{-1} * A = I$, *the identity*):

$$A^{-1} = (b * d)^{-1} * \begin{pmatrix} d & -a * d \\ -b * c & b \end{pmatrix}.$$

Proof. (i) This follows from (3.1) and (3.4) after some straightforward algebra.

(ii) This follows readily from (i) and the assumed statistical properties w^1 and w^2.

(iii) The stability of c implies that d is a unit, and the formula for A^{-1} may be verified by a direct calculation.

We can now give a more precise reformulation of the regulator problem.

REGULATOR PROBLEM (R.P.). Let n be a positive integer or $+\infty$. Let $z_{t,n}$, L_{t+1} and $J(u_t', z_{t,n})$ be defined by (1.2), (1.3) and (1.4). With the system (3.1) the problem is to find the vectors $h^1 = (h_0^1, h_1^1, \cdots, h_{n-1}^1, 0, \cdots)$ and $h^2 = (0, h_1^2, \cdots, h_{n-1}^2, 0, \cdots)$ such that:

(i) $u_t' = \sum_0^{n-1} h_i^1 y_{t-i} + \sum_1^{n-1} h_i^2 u_{t-i}$ minimizes $J(u_t', z_{t,n})$; i.e.,

$$J(u_t', z_{t,n}) = \min_{w \in U_{t,n}} J(w, z_{t,n}), \qquad t \in T;$$

(ii) $c = h^1 * (e - h^2)^{-1}$ is stable;

(iii) $e - a * c$ is a unit.

The conditional expectation defining $J(u_t', z_{t,n})$ is for the statistics of the process (3.1) with c as the control law.

We shall speak of $h = \begin{pmatrix} h^1 \\ h^2 \end{pmatrix}$ or c given by (ii) as being solutions of R.P. We want to develop conditions on solutions in terms of the random processes y and u. If c is a solution, the conditions of Proposition 3.1 are satisfied, and both y and u are well-defined stationary Gaussian random processes. Introduce the vector

$$x_{t,n} = (y_t, u_t, y_{t-1}, u_{t-1}, \cdots, u_{t-n+1}), \tag{3.6}$$

which differs from $z_{t,n}$ only in the appearance of the term u_t. There are then vectors

$$p^1 = (p_0^1, p_1^1, \cdots, p_{n-1}^1),$$
$$p^2 = (p_0^2, p_1^2, \cdots, p_{n-1}^2),$$

such that

$$\hat{y}_{t+1} = E(y_{t+1} \mid x_{t,n}) = \sum_0^{n-1} p_i^1 y_{t-i} + \sum_0^{n-1} p_i^2 u_{t-i}. \tag{3.7}$$

The value of $p = \begin{pmatrix} p^1 \\ p^2 \end{pmatrix}$ in general, depends on c. The quantity $\hat{y}_{t+1}$ is called the predicted value of y_{t+1}, and it is related to y_{t+1} by the equation

$$y_{t+1} = \hat{y}_{t+1} + \epsilon_{t+1}, \tag{3.8}$$

where ϵ_{t+1} is a random variable that is independent of the conditioning random variables in $x_{t,n}$.

We now invoke a so-called "smoothing" property of conditional expectations (see [7]) and write

$$\begin{aligned} J &= E(L_{t+1} \mid z_{t,n}) \\ &= E(E(L_{t+1} \mid x_{t,n}) \mid z_{t,n}). \end{aligned} \tag{3.9}$$

Using (3.7), (3.8) and the independence of ϵ_{t+1} and $x_{t,n}$, we can express the term $E(L_{t+1} \mid x_{t,n})$ in the following way:

$$E(L_{t+1} \mid x_{t,n}) = \hat{y}_{t+1}^2 + \rho u_t^2 + E(\epsilon_{t+1}^2). \tag{3.10}$$

Then, putting $u_t = u_t' + w_t$, substituting (3.10) in (3.9) and using the independence of w_t and $z_{t,n}$, we obtain

$$\begin{aligned} J = \left(\sum_0^{n-1} p_i^1 y_{t-i} + \sum_1^{n-1} p_i^2 u_{t-i} + p_0^2 u_t' \right)^2 \\ + \rho (u_t')^2 + \rho\sigma^2 + (p_0^2)^2 + E(\epsilon_{t+1}^2). \end{aligned} \tag{3.11}$$

The last term being independent of $u_t' \in U_{t,n}$, the minimum of J is achieved by setting

$$u_t' = \mu\left(\sum_0^{n-1} p_i^1 y_{t-i} + \sum_1^{n-1} p_i^2 u_{t-i}\right),$$

$$\mu = -\frac{p_0^2}{\rho + (p_0^2)^2}. \tag{3.12}$$

Equation (3.12) defines a control law of the form (3.4):

$$c = \alpha p^1 * (e - \alpha p^2)^{-1}, \quad \alpha = -p_0^2/\rho,$$

providing $(e - \alpha p^2)$ is a unit.[2]

In deriving (3.12), we have tacitly assumed that $(p_0^2)^2 + \rho > 0$. This is reasonable, for we must assume $p_0^2 \neq 0$ to have a physically meaningful problem. If this condition is not satisfied, u_t has no influence on y_{t+1}, and the problem set up must be modified.

We have thus shown that if c is a solution, u_t' must satisfy (3.12), where p is defined by (3.7). On the other hand, let a c, satisfying conditions (ii) and (iii) of the problem statement, be determined by (3.12) for some p. Then if p satisfies (3.7) for the processes y and u that correspond to c, the expression (3.11) is minimized and thus condition (i) is satisfied, and c is a solution of R.P. Thus, we have proven the following theorem.

THEOREM 3.2. *A control law c solves R.P. if and only if for some vectors*

$$p^1 = (p_0^1, p_1^1, \cdots, p_{n-1}^1, 0, \cdots),$$

$$p^2 = (p_0^2, p_1^2, \cdots, p_{n-1}^2, 0, \cdots),$$

the following conditions are satisfied:

(i) $E(y_{t+1} \mid x_{t,n}) = \sum_0^{n-1} p_i^1 y_{t-i} + \sum_0^{n-1} p_i^2 u_{t-i}$;

(ii) $c = \alpha p^1 * (e - \alpha p^2)^{-1} \in l^+, \quad \alpha = -p_0^2/\rho$;

(iii) $e - a * c$ *is a unit.*

The conditional expectation in (i) is defined for the random processes obtained with c as the control law.

4. Wiener-Hopf equation in l. This section contains a discussion of the Wiener-Hopf (W-H) equation in l that arises in the Wiener-Kolmogorov theory of prediction. The results obtained are used extensively in later sections to solve R.P. If the reader wishes to omit the proofs, he can without loss of continuity skip the material following (4.10), and refer to the results only when needed.

We begin by defining the W-H equation of interest. Let A be a regular 2×2 matrix with entries in l^+ (i.e., A has an inverse also with entries

[2] If $\rho = 0$, then $c = -p^1 * (p^2)^{-1}$.

in l^+). Let

$$R = A^* * A'. \tag{4.1}$$

Let $p^1, p^2, q^1, q^2 \in l^+$ and let

$$p = \begin{pmatrix} p^1 \\ p^2 \end{pmatrix}, \quad q = \begin{pmatrix} q^1 \\ q^2 \end{pmatrix}. \tag{4.2}$$

We consider the W-H equation:

$$(R * p - q)_+ = 0. \tag{4.3}$$

Occasionally, a different form for this equation is preferable. Let

$$\Gamma = \begin{pmatrix} R(0) & R(-1) & R(-2) & \cdots \\ R(1) & R(0) & R(-1) & \cdots \\ R(2) & R(1) & R(0) & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}. \tag{4.4}$$

The sum of the absolute values of the elements in each row of Γ is uniformly bounded, and therefore Γ, considered as a linear map of l^+ into l^+, is a bounded operator. If we arrange the elements of the vectors (4.2) as follows (keeping the same notation p and q):

$$p = \begin{pmatrix} p_0^1 \\ p_0^2 \\ p_1^1 \\ p_1^2 \\ \vdots \end{pmatrix}, \qquad q = \begin{pmatrix} q_0^1 \\ q_0^2 \\ q_1^1 \\ q_1^2 \\ \vdots \end{pmatrix}, \tag{4.5}$$

then (4.3) is equivalent to

$$\Gamma p - q = 0. \tag{4.6}$$

Finite-dimensional analogues of (4.3) are obtained by requiring that the solution p belong to proper subspaces of l^+. Let

$$M_n = \left\{ p^n : p^n = E_n \begin{pmatrix} p^1 \\ p^2 \end{pmatrix}, p^1, p^2 \in l^+ \right\}.$$

The corresponding equations are

$$E_n(R * p^n - q^n) = 0, \qquad p^n, q^n \in M_n. \tag{4.7}$$

The matrix analogue of (4.7) is obtained by defining

$$\Gamma_n = \begin{pmatrix} R(0) & R(-1) & \cdots & R(1-n) \\ R(1) & R(0) & & \\ \vdots & & & \vdots \\ R(n-1) & & & R(0) \end{pmatrix} \tag{4.8}$$

and making the identification:

$$p^n = \begin{pmatrix} p_0^1 \\ p_0^2 \\ p_1^1 \\ \vdots \\ p_{n-1}^2 \end{pmatrix}, \qquad q^n = \begin{pmatrix} q_0^1 \\ q_0^2 \\ \\ \vdots \\ q_{n-1}^2 \end{pmatrix}. \tag{4.9}$$

Then (4.7) is equivalent to

$$\Gamma_n p^n - q^n = 0. \tag{4.10}$$

Although the same notation p^n, q^n is used both for the $2n$-dimensional vectors (4.9) and the 2-vectors in M_n, no confusion should occur. With this understanding, Γ_n can be considered as a map of M_n into M_n, and in subsequent discussions it will be convenient to use this interpretation. Similar remarks apply to p, q and Γ.

In the W-H equations above, we are given R and q and want to solve for p. The solution of (4.3) is given in the following proposition.

PROPOSITION 4.1. *The W-H equation* (4.3) *has a unique solution* p *given by*

$$p = (A'^{-1}) * (A^{*-1}q)_+ .$$

Proof. The proof consists of an application of Wiener's spectral factorization method (see [8]). Define g by the equation

$$A^* * g = q. \tag{4.11}$$

This is legitimate since A, by hypothesis, is regular. Substituting (4.11) into (4.3) and using (4.1), we obtain

$$(A^* * (A' * p - g))_+ = 0. \tag{4.12}$$

Write $g = (g - g_+) + g_+$ in (4.12), where $g_+ = E_+ g$. Then

$$(A^* * (A' * p - g_+))_+ - (A^* * (g - g_+))_+ = 0.$$

In the second term, note that $(g - g_+)_i = 0$ for $i \geqq 0$. Since $(A^*)_i = 0$ for $i > 0$, the positive part of $A^* * (g - g_+)$ is zero. Hence we have

$$(A^* * (A' * p - g_+))_+ = 0. \tag{4.13}$$

Equation (4.13) can be satisfied by setting

$$A' * p - g_+ = 0 \tag{4.14}$$

or

$$p = (A')^{-1} * g_+ . \tag{4.15}$$

Since the components of g_+ and $(A')^{-1} \in l^+$, p given by (4.15) is a solution of the W-H equation.

To see that (4.15) gives the only solution, let w be another, and let $z = p - w$. Then the components of z and of $y = A' * z$ belong to l^+: $y_i = 0$ for $i < 0$. Since both p and w satisfy (4.13), we must have $(A^* * y)_+ = 0$. This implies that $(A^* * y)_i = 0$, $i \geqq 0$. Since also $(A^{*^{-1}})_i = 0$ for $i \geqq 0$, it follows from the identity $A^{*^{-1}} * (A^* * y) = y$ that $y_i = 0$, $i \geqq 0$. But since $y_i = 0$ for $i < 0$, we must have $y = 0$, and thus $z = 0$, since A' is invertible.

Regarding the finite-dimensional analogues (4.7) and (4.10) of the W-H equation, the following proposition implies that they can be solved, too.

PROPOSITION 4.2 (Szegö). *There exist constants δ_1, δ_2 such that the eigenvalues λ_m of Γ_n satisfy the inequalities*

$$0 < \delta_1 \leqq \lambda_m \leqq \delta_2, \qquad m = 1, 2, \cdots, 2n.$$

The constants δ_1, δ_2 may be chosen independently of n.

Proof. Let $A(z)$ be the z-transform of A. Then from (4.1),

$$\hat{R}(z) = \hat{A}\left(\frac{1}{z}\right)\hat{A}'(z).$$

Combining the hypothesis that A is regular with Theorem 2.2, we can conclude that there exist δ_1, $\delta_2 > 0$ such that

$$\delta_1 I \leqq \hat{R}(z) \leqq \delta_2 I, \qquad z \in C, \tag{4.16}$$

where I is the 2×2 identity matrix.

From this point on, our proof is an obvious extension of the proof of Szegö [9] to vector-valued processes. Let $u_i = (u_i^1, u_i^2)$, $0 \leqq i \leqq n-1$, be arbitrary 2-vectors of complex numbers, and let $u = (u_0, u_1, \cdots, u_{n-1})$. Let $\hat{u}(z) = \sum_0^{n-1} u_i z^i$. Then as can be verified by direct integration, we have the identity

$$(u, \Gamma_n u) = \frac{1}{2\pi i}\oint_C \left(\hat{u}\left(\frac{1}{z}\right), \hat{R}(z)\hat{u}(z)\right)\frac{dz}{z}, \tag{4.17}$$

where (x, y) is the scalar product of x and y. Applying (4.16) to the right-hand side of (4.17), we obtain

$$\delta_1(u, u) \leqq (u, \Gamma_n u) \leqq \delta_2(u, u),$$

from which the result follows.

Proposition 4.2 implies that Γ_n has an inverse, Γ_n^{-1}, for every n. We shall need the following sharper form of Proposition 4.2 in subsequent sections, the proof of which is considerably more difficult.

LEMMA 4.3. *There exists a constant M such that*

$$\sup_n \| \Gamma_n^{-1} \| \leqq M.$$

Proof. Suppose that the Γ_n^{-1} are not uniformly bounded. Then there is a sequence of $x^n \in M_n$, $n = 1, 2, \cdots$, such that

$$\| x^n \| = 1, \quad E_n(R * x^n) \to 0. \tag{4.18}$$

Let $y^n = E_n(R * x^n)$ and let $x^n = \begin{pmatrix} x^{n,1} \\ x^{n,2} \end{pmatrix}$ and $y^n = \begin{pmatrix} y^{n,1} \\ y^{n,2} \end{pmatrix}$. Then from (4.18) it is clear that

$$S_n = \sum_0^n (y_i^{n,1})^2 + (y_i^{n,2})^2 \to 0.$$

But by Proposition 4.1, since $x^n = \Gamma_n^{-1} y^n$,

$$\sum_0^n (x_i^{n,1})^2 + (x_i^{n,2})^2 \leqq \frac{1}{\delta_1} S_n .$$

Hence we have the estimate

$$\epsilon_n = \max_{0 \leqq i \leqq n-1} \| x_i^n \| \to 0. \tag{4.19}$$

Let $z_n = (R * x^n)_+$. By Proposition 4.1 we have

$$x^n = (A'^{-1}) * (A^{*-1} * z_n)_+ ,$$

and since $\| x^n \| = 1$, there is a δ such that

$$\inf_n \| z_n \| \geqq \delta > 0. \tag{4.20}$$

Now using (4.18) we have

$$\delta \leqq \| z_n \| = \| (E_+ - E_n) z_n \| + \| E_n z_n \| \to \| (E_+ - E_n) z_n \|.$$

We shall obtain a contradiction by showing that $w_n = (E_+ - E_n) z_n \to 0$.

First we introduce the 2×2 matrices U^m and V^m defined by the equations:

$$U^m(n) = \begin{cases} R(n) & \text{for } |n| \leqq m-1, \\ 0 & \text{for } |n| > m-1, \end{cases}$$

$$V^m(n) = \begin{cases} 0 & \text{for } |n| \leqq m-1, \\ R(n) & \text{for } |n| > m-1. \end{cases}$$

We obviously have

$$\begin{aligned} R &= U^m + V^m, \\ \| V^m \| &\to 0, \quad m \to \infty. \end{aligned} \tag{4.21}$$

In terms of U^m and V^m,

$$
\begin{aligned}
w_n &= (E_+ - E_n)(U^m + V^m) * x^n \\
&= (E_+ - E_n)(U^m * x^n) + \delta_m ,
\end{aligned} \tag{4.22}
$$

where $\delta_m \to 0$ uniformly in m as $m \to \infty$.

For $n > m$ write $x^n = (E_n - E_{n-m})x^n + E_{n-m}x^n$. It is clear that

$$E_+(U^m * (E_{n-m}x^n)) \in M_n$$

and therefore, since $(E_+ - E_n)v = 0$ for $v \in M_n$, we have from (4.22),

$$w_n = (E_+ - E_n)(U^m * (E_n - E_{n-m})x^n) + \delta_m .$$

Finally using (4.19) and (4.21), we have

$$\| w_n \| \leqq \| R \| m\epsilon_n + \| \delta_m \|,$$

which can be made arbitrarily small. This completes the proof.

5. Fixed-point condition. In this section, the criteria in Theorem 3.2 which characterize solutions of R.P. are restated as a fixed-point condition.

Let G be the set of control laws defined as follows:

$$G = \{c : c \in l^+,\ e - a * c \text{ is a unit}\}.$$

Then, for each $c \in G$, the prediction formula

$$\hat{y}_{t+1} = \sum_0^{n-1} p_i^1 y_{t-i} + \sum_0^{n-1} p_i^2 u_{t-i} \tag{3.7}$$

is uniquely determined by the condition that the prediction error is orthogonal to $x_{t,n}$:

$$
\begin{aligned}
E((y_{t+1} - \hat{y}_{t+1}) \cdot y_{t'}) &= 0, \qquad t - n + 1 \leqq t' \leqq t, \\
E((y_{t+1} - \hat{y}_{t+1}) \cdot u_{t'}) &= 0, \qquad t - n + 1 \leqq t' \leqq t.
\end{aligned} \tag{5.1}
$$

Substitution of (3.7) into these equations yields a system of linear equations for the unknown p_i's in terms of the covariance matrix of the process. These turn out to be identical to (4.3) and (4.7) in the case $n = +\infty$ and $n < \infty$, respectively, if we set:

$$
\begin{aligned}
A &= \Delta^{-1} * \begin{pmatrix} b & a * d \\ b * c & d \end{pmatrix}, \\
R &= (r_{ij}) = A^* * A', \\
p^1 &= (p_0^1, p_1^1, \cdots), \quad p^2 = (p_0^2, p_1^2, \cdots), \\
q^1 &= (U^{-1}r_{11})_+, \qquad q^2 = (U^{-1}r_{21})_+, \\
p &= \begin{pmatrix} p^1 \\ p^2 \end{pmatrix}, \qquad q = \begin{pmatrix} q^1 \\ q^2 \end{pmatrix}.
\end{aligned} \tag{5.2}
$$

Then p is determined by the W-H equation

$$E_+(R * p - q) = 0 \tag{5.3}$$

in the case $n = +\infty$, and by

$$E_n(R * p^n - q^n) = 0, \quad p^n, q^n \in M_n, \tag{5.4}$$

in the case $n < \infty$.

We showed in the last section (Propositions 4.1–4.2) that (5.3) and (5.4) have unique solutions. Since R and q depend on c, the solutions can be considered as functions of c. We define maps f_+ and f_n on G to $l^+ \times l^+$ by setting

$$\begin{aligned} f_+(c) &= p, \\ f_n(c) &= p^n, \end{aligned} \tag{5.5}$$

where p and p^n are the solutions of (5.3) and (5.4) corresponding to c. We also define a map g on $l^+ \times l^+$ to l^+ by setting

$$g(p) = \alpha * p^1 * (e - \alpha p^2)^{-1}, \quad \alpha = -p_0^2/\rho. \tag{5.6}$$

In terms of these maps, we can restate Theorem 3.2 in the following way.

COROLLARY 5.1. *A control law $c \in G$ solves R.P. if and only if there exists a p satisfying the fixed-point condition*:

(i) $f_+(g(p)) = p$,

(ii) $f_n(g(p)) = p$, $p \in M_n$,

in the cases $n = +\infty$ and $n < \infty$, respectively. The solution c is then given by $c = g(p)$.

6. Case $n = +\infty$. The R.P. for the case $n = +\infty$ is solved completely in this section. The distinctive feature of R.P. for $n = +\infty$ is that the prediction formula is independent of the control law c as shown by the following lemma.

LEMMA 6.1. *For all $c \in G$, the optimal predictor is given by*

$$f_+(c) = \begin{pmatrix} p^1 \\ p^2 \end{pmatrix} = p,$$

where $p^1 = b^{-1} * (U^{-1}b)_+$ and $p^2 = b_0, b^{-1} * (U^{-1}a)_+$.

Proof. Let $\lambda = \begin{pmatrix} e \\ 0 \end{pmatrix}$. Then from (5.2) and (4.1),

$$q = (U^{-1}R * \lambda)_+ = (U^{-1}A^* * A' * \lambda)_+,$$

and from Proposition 4.1,

$$f_+(c) = p = (A'^{-1}) * (A^{*-1} * (U^{-1}A^* * A'\lambda)_+)_+.$$

Since $(A^{*-1})_i = 0$ for $i > 0$, note that for any y we have the identity

$$(A^{*-1} * y)_+ = (A^{*-1} * (y)_+)_+ .$$

Applying this result in the expression above for p, we get

$$p = (A'^{-1}) * (U^{-1}A' * \lambda)_+ .$$

Then using (5.2) and the formula for A'^{-1} in Proposition 3.1, we obtain, after a straightforward computation, the result stated in the lemma.

The following lemma is a partial converse of Lemma 6.1 and gives the conditions under which $c \in G$.

LEMMA 6.2. *Let p be given by Lemma* 6.1. *Then*

(i) *the control law*

$$c = g(p) = \alpha p^1 * (e - \alpha p^2)^{-1}, \quad \alpha = -p_0^2/\rho,$$

is stable if and only if $1/\alpha \notin \sigma(p^2)$;

(ii) *if c is stable, then $e - a * c$ is a unit if and only if* $1/\alpha \notin \sigma((U^{-1}a)_+)$.

Thus $c \in G$ if and only if both (i) *and* (ii) *hold.*

Proof. (i) The first statement is obvious from the definition of the spectrum $\sigma(p^2)$.

(ii) If c is stable then $e - \alpha p^2$ is a unit, and $e - a * c$ is a unit if and only if $w = (e - a * c) * (e - \alpha p^2) = e - \alpha p^2 - \alpha p^1$ is a unit. Substituting for p^2 and p^1 from Lemma 6.1, and taking z-transforms, we get after some algebra

$$\hat{w}(z) = 1 - \alpha z^{-1}\hat{a}(z);$$

and therefore, since $a_0 = 0$,

$$w = e - \alpha(U^{-1}a)_+ ,$$

from which the result follows.

Combining Lemmas 6.1-6.2 with Corollary 5.1, we have the solution to R.P. with $n = +\infty$.

THEOREM 6.3. *Let*

$$\begin{aligned} p^1 &= b^{-1} * (U^{-1}b)_+ , \\ p^2 &= b_0 b^{-1} * (U^{-1}a)_+ , \\ \alpha &= -p_0^2/\rho. \end{aligned} \tag{6.1}$$

Then a solution to R.P. for $n = +\infty$ exists if and only if

$$1/\alpha \notin \sigma(p^2), \quad 1/\alpha \notin \sigma((U^{-1}a)_+). \tag{6.2}$$

If the conditions (6.2) *hold, then*

$$c = \alpha p^1 * (e - \alpha p^2)^{-1} \tag{6.3}$$

is the required solution.

Proof. If c solves R.P., then by Corollary 5.1, $c = g(p)$ for some p satisfying the equation

$$f_+(g(p)) = f_+(c) = p.$$

But by Lemma 6.1, $f_+(c) = p$, p given by (6.1); and therefore c must be given by (6.3). By Lemma 6.2, $c \in G$ only if the conditions (6.2) are satisfied. This shows the necessity of (6.2).

Conversely, if the conditions (6.2) are satisfied, the control law c defined by (6.1) belongs to G. By Lemma 6.1, $f_+(c) = p$, p given by (6.1), and therefore,

$$f_+(g(p)) = p,$$

which by Corollary 5.1 shows that c solves R.P.

7. Case $n < \infty$. In this section, we show that R.P. has solutions for sufficiently large n. We show, in addition, how the solutions may be computed by an iterative algorithm.

In what follows we assume that conditions (6.2) of Theorem 6.3 are satisfied. Then R.P. for $n = +\infty$ has a unique solution denoted by c^+. Let p^+ be the predictor defined by

$$p^+ = f_+(c^+). \tag{7.1}$$

For $r > 0$, let

$$\begin{aligned} U_r &= \{c: \| c - c^+ \| \leqq r;\ c, c^+ \in l^+\}, \\ V_r &= \{p: \| p - p^+ \| \leqq r;\ p, p^+ \in l^+\}. \end{aligned} \tag{7.2}$$

For each c, let $\Gamma_n(c)$ be the map Γ_n corresponding to the covariance matrix R which is obtained with the control law c.

LEMMA 7.1. *There exists an $r_0 > 0$ and a constant M such that*

$$\sup_{c \in U_{r_0}} \sup_n \| \Gamma_n^{-1}(c) \| \leqq M.$$

Proof. For each n and c consider the map $\Gamma_n(c): M_n \to M_n$, and define

$$\Delta_n(c) = \Gamma_n(c) - \Gamma_n(c^+).$$

We have

$$\| \Delta_n(c) x^n \| = \| E_n[(R(c) - R(c^+)) * x^n] \| \leqq \| R(c) - R(c^+) \| \, \| x^n \|.$$

Since the map $c \to R(c) = A(c) * A'(c)$, where $A(c)$ is given by (5.2), is continuous at $c = c^+$, we can find for each $M > 0$, an $r_0 > 0$ such that

$$\sup_{c\in U_{r_0}} \| R(c) - R(c^+) \| \leqq \frac{1}{4M}.$$

Thus we have for such r_0,

$$\sup_n \sup_{c\in U_{r_0}} \| \Delta_n(c) \| \leqq \frac{1}{4M}. \tag{7.3}$$

By Lemma 4.3 there is an M such that

$$\sup_n \| \Gamma_n^{-1}(c^+) \| \leqq M/2. \tag{7.4}$$

Thus the series

$$\Gamma_n^{-1}(c) = \left(\sum_{i=0}^{\infty} (-\Gamma_n^{-1}(c^+)\Delta_n(c))^i\right)\cdot \Gamma_n^{-1}(c^+), \qquad c \in U_{r_0}, \tag{7.5}$$

is absolutely convergent, and we obtain from (7.5) the desired inequality

$$\sup_{c\in U_{r_0}} \sup_n \| \Gamma_n^{-1}(c) \| \leqq M.$$

LEMMA 7.2. *Let r_0 be chosen according to Lemma 7.1. Then for each $0 < r \leqq r_0$ there is an n_r such that*

$$f_n(U_{r_0}) \subset V_r \quad \textit{for} \quad n \geqq n_r.$$

Proof. By Lemma 6.1 we have for all $c \in U_{r_0}$,

$$E_+(R(c) * p^+ - q(c)) = 0.$$

On the other hand,

$$E_n(R(c) * f_n(c) - q^n(c)) = 0.$$

Writing $p^+ = E_n p^+ + (E_+ - E_n)p^+$, we have from the last two equations,

$$E_n(R(c) * (f_n(c) - E_n p^+) - R(c) * (E_+ - E_n)p^+) = 0.$$

Considering this expression as a Wiener-Hopf equation for $f_n(c) - E_n p^+$, we have by Lemma 7.1,

$$\begin{aligned} \| f_n(c) - p^+ \| &= \| f_n(c) - E_n p^+ \| + \| (E_+ - E_n)p^+ \| \\ &\leqq (1 + M \sup_{c\in U_{r_0}} \| R(c) \|) \| (E_+ - E_n)p^+ \|. \end{aligned}$$

The right-hand side of the last equation converges uniformly to 0 as $n \to \infty$, thus proving the lemma.

THEOREM 7.3. *There exists a positive integer n_0 such that for all $n > n_0$, R.P. has a solution.*

Proof. In view of Corollary 5.1, it suffices to show that the maps

$p \to f_n(g(p)) = T_n(p)$ have fixed points for sufficiently large n. For $r > 0$ let

$$V_{r,n} = \{p: \| p - p^+ \| \leqq r, p \in M_n\},$$
$$U_{r,n} = \{c: c = g(p), p \in V_{r,n}\}.$$

Choose r_0 according to Lemma 7.2. Since g is continuous at p^+ and $g(p^+) = c^+$, there exists an $r > 0$ such that $g(V_r) \subset U_{r_0}$, and since $V_r \supset V_{r,n}$, $g(V_{r,n}) \subset U_{r_0}$.

By Lemma 7.2, there exists an n_0 such that $f_n(U_{r_0}) \subset V_r$ for $n > n_0$, and since $f_n(U_{r_0}) \subset M_n$, $f_n(U_{r_0}) \subset V_{r,n}$. By definition, $g(v_{r,n}) = U_{r,n}$ and we have $T_n(V_{r,n}) = f_n(g(V_{r,n})) \subset V_{r,n}$ for $n > n_0$. Finally, noting that the sets $V_{r,n}$ are convex and compact and that the maps T_n are continuous, we see that the existence of fixed points for $n > n_0$ follows from Schauder's fixed-point theorem.

In proving Theorem 7.3 we showed that the maps $p \to f_n(g(p)) = T_n(p)$ map $V_{r,n}$ into itself for n greater than some n_0. We want to show that there is an n_1 such that T_n is contractive for $n > n_1$. For such n, the fixed point of T_n in $V_{r,n}$ is then unique and may be computed by the formula

$$p = \lim_{m \to \infty} T_n{}^m(p^0), \qquad p^0 \in V_{r,n}, \tag{7.6}$$

and the solution to R.P. is then $g(p)$.

In order to show that T_n is a contraction, we need a few facts about derivatives. Let f be a continuous map of l into l. We recall that f is differentiable at x_0 if there exists a linear map $Df(x_0)$, called the derivative of f at x_0, which satisfies the equation

$$f(x_0 + x) = f(x_0) + Df(x_0)x + o(x),$$
$$\lim_{\substack{x \to 0 \\ x \neq 0}} \frac{\| o(x) \|}{\| x \|} = 0. \tag{7.7}$$

We shall need the following properties of derivatives, the proofs of which are immediate from the definitions (see [10]).

PROPOSITION 7.4. *Let $f: l \to l$ be differentiable at x_0 and $A: l \to l$ be a (constant) linear map. Then the composite map $A \circ f(x) = A(f(x))$, $x \in l$, is differentiable at x_0 and*

$$D(A \circ f)(x_0) = A \circ Df(x_0).$$

PROPOSITION 7.5. *Let f and g be maps on l into l, and suppose g is differentiable at x_0 and f differentiable at $y_0 = g(x_0)$. Then the composite map $f \circ g$ is differentiable at x_0,*

$$D(f \circ g)(x_0) = Df(y_0) \circ Dg(x_0).$$

PROPOSITION 7.6. *Let f and g be maps on l into l, and suppose f and g are differentiable at x_0. Then the map $h = f * g$ defined by $h(x) = f(x) * g(x)$, $x \in l$, is differentiable at x_0 and its derivative is defined by the equation*

$$Dh(x_0)\cdot x = (Df(x_0)\cdot x) * g(x_0) + f(x_0) * (Dg(x_0)\cdot x), \qquad x \in l.$$

COROLLARY 7.7. *Let f, g, h be maps on l into l, and suppose $h = f * g$. Let h and f be differentiable at x_0 and suppose $f(x_0)$ is a unit. Then g is differentiable at x_0 and*

$$Dg(x_0)\cdot x = (f(x_0))^{-1} * (Dh(x_0)\cdot x - Df(x_0)\cdot x) * g(x_0), \qquad x \in l.$$

We now apply the above results to the maps in our problem.

LEMMA 7.8. *There exists an $r > 0$ such that the map $c \to \Delta^{-1}(c) = (e - a*c)^{-1}$ is differentiable for $c \in U_{r_0}$, and*

$$D\Delta^{-1}(c) = \Delta^{-1}(c) * a * \Delta^{-1}(c).$$

Proof. Since the set of units in any Banach algebra is open, we have the identity

$$e = \Delta(c) * \Delta^{-1}(c), \tag{7.8}$$

valid for all c in some neighborhood U_{r_0} of c^+. Applying Corollary 7.7 and Proposition 7.5, we have

$$D\Delta^{-1}(c) = \Delta^{-1}(c) * a * \Delta^{-1}(c).$$

COROLLARY 7.9. *There exists an $r_0 > 0$ such that the maps $c \to R(c)$ and $c \to q(c)$ are differentiable in U_{r_0}. Furthermore,*

$$\sup_{c \in U_{r_0}} \| DR(c) \| < \infty, \quad \sup_{c \in U_{r_0}} \| Dq(c) \| < \infty.$$

Proof. This follows easily from Propositions 7.4–7.5, Lemma 7.8 and the formulas for the maps, (5.2).

LEMMA 7.10. *There exists an $r_0 > 0$ such that the maps $c \to f_n(c)$ are differentiable in U_{r_0}. Furthermore,*

$$\lim_{n\to\infty} \sup_{c \in U_{r_0}} \| Df_n(c) \| = 0.$$

Proof. First we note that $Df_+(c)$, $c \in U_{r_0}$, is the zero map since f_+ is constant on U_{r_0}. By differentiation of the equality

$$E_+(R(c) * f_+(c) - q(c)) = 0$$

(using Propositions 7.4, 7.6 and Corollary 7.9), we have for $x \in l$, $c \in U_{r_0}$,

$$E_+[(DR(c)\cdot x) * f_+(c) + R(c) * (Df_+(c)\cdot x) - Dq(c)\cdot x] = 0.$$

But since $Df_+(c)\cdot x = 0$ for all x,

$$E_+[(DR(c)\cdot x) * f_+(c) - Dq(c)\cdot x] = 0. \tag{7.9}$$

Now we differentiate the expression

$$E_n(R(c) * (f_n(c) - q(c)) = 0$$

to get for $x \in l$, $c \in U_{r_0}$,

$$E_n[(DR(c)\cdot x) * f_n(c) + R(c) * (Df_n(c)\cdot x) - Dq(c)\cdot x] = 0 \tag{7.10}$$

Let $\| x \| = 1$, and let

$$z_n = ((DR(c)\cdot x) * f_n(c) - Dq(c)\cdot x)_+ .$$

In view of (7.9), Lemma 7.2, and Corollary 7.9,

$$\lim_n \sup_{c \in U_{r_0}} \| z_n \| = 0.$$

Solving (7.10) for $Df_n(c)\cdot x$, we have

$$\begin{aligned} \| Df_n(c)\cdot x \| &\leqq \sup_{c \in U_{r_0}} \| \Gamma_n^{-1}(c) \| \cdot \| z_n \| \\ &\leqq M \| z_n \| \to 0, \end{aligned}$$

and the result follows.

THEOREM 7.11. *There exists an $r > 0$ and an integer n_1 such that the mappings*

$$p \to T_n(p) = f_n(g(p)), \qquad p \in V_{r,n},$$

are contractive for all $n > n_1$.

Proof. It suffices to show that $DT_n(p) \to 0$ uniformly for $p \in V_r$. Applying Proposition 7.5 we have $DT_n(p) = Df_n(c) \circ Dg(p)$, $c = g(p)$. Clearly the map

$$p \to g(p) = \alpha p^1 * (e - \alpha p^2)^{-1}$$

is differentiable in some neighborhood V_r of p^+, and $\sup_{p \in V_r} \| Dg(p) \| < \infty$. On the other hand, for r sufficiently small, $c = g(p) \in U_{r_0}$ for all $p \in V_r$. Then applying Lemma 7.10, we have

$$\lim_{n \to \infty} \sup_{p \in V_r} \| DT_n(p) \| \leqq \lim_{n \to \infty} \sup_{c \in V_{r_0}} \| Df_n(c) \| \sup_{p \in V_r} \| Dg(p) \| = 0,$$

which proves the result.

In conclusion, note that $T_n(p)$ can be evaluated for each p with arbitrary precision from a sufficiently long record of input and output: it is only necessary to determine the covariance matrices $R(0), R(1), \cdots, R(n-1)$. Thus the regulation problem can be solved without identifying the vectors a and b, which define the exact characteristics of the system.

REFERENCES

[1] C. GALTIERI, *The problem of identification of discrete time process*, Rep. 63-19, Electronics Research Laboratory, University of California, Berkeley, 1963.
[2] J. EATON, *Identification for control purposes*, IEEE International Convention Record, 15 (1967), Part 3, Automatic Control, pp. 38–52.
[3] G. BOX AND G. JENKINS, *Some statistical aspects of adaptive optimization and control*, J. Roy. Statist. Soc. Ser. B, 24 (1962), pp. 297–343.
[4] E. R. LORCH, *Spectral Theory*, Oxford University Press, New York, 1962.
[5] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1962.
[6] L. ZADEH AND C. DESOER, *Linear Systems Theory. The State Space Approach*, McGraw-Hill, New York, 1963.
[7] M. LOÈVE, *Probability Theory*, Van Nostrand, New York, 1955.
[8] N. WIENER, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, John Wiley, New York, 1949.
[9] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley and Los Angeles, 1958.
[10] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

BOUNDARY CONTROL SYSTEMS*

H. O. FATTORINI†

Introduction. We consider in this paper certain types of control systems, among them those described by a partial differential equation in a domain I of Euclidean space, the equation being of first or second order with respect to time. The control is exerted on the system by means of the boundary conditions and (possibly) by means of parameters distributed all over I, and we seek conditions on the way the control is applied that assure that the system can be steered from an arbitrary initial state to (the vicinity of) an arbitrary final state. These conditions are obtained by replacing the boundary control by distributed controls that have the same effect on the system, and then applying known results for distributed parameter systems (this type of argument has already been used in connection with controllability and optimal problems; see, for instance, [14]). The results are then applied to a number of particular equations in §5 and §6. Sections 1, 3 and 5 are devoted to equations of first order in time, §§2, 4 and 6 to second order equations; since the proof of many facts for these equations are similar to, or depend on, results for first order equations, the even numbered sections are dependent on the rest.

1. The boundary control system (first order equations). If $E = \{u, v, \cdots\}$ is any Banach space we shall denote by $E^* = \{u^*, v^*, \cdots\}$, the dual space of E, and by $\langle u, u^*\rangle$ or $\langle u^*, u\rangle$ the value of the functional $u^* \in E^*$ at $u \in E$. If K is any subset of E, we define $K^\perp = \{u^* \in E^* \mid \langle u^*, u\rangle = 0\}$ for all $u \in K$. Plainly $K^\perp$ is a subspace of E^*; if K itself is a subspace of E, by the Hahn-Banach theorem $ClK = E$ if and only if $K^\perp = \{0\}$.

The space E, the *state-space* of the system, will be a complex Banach space; in the applications we have in mind, E will be a space of (ordinary or generalized, possibly vector-valued) functions defined in a domain I of Euclidean space R^r.

We shall consider a closed linear operator σ with domain $D(\sigma) \subseteq E$ and range in E and a linear operator τ (the boundary operator) with domain $D(\tau) \subseteq E$ and range in some Banach space X. Since σ is closed, $D(\sigma)$ becomes a Banach space if endowed with the usual "graph" norm, i.e.,

* Received by the editors July 31, 1967, and in revised form February 29, 1968.

† Center for Dynamical Systems, Brown University, Providence, Rhode Island. Now with Department of Mathematics, University of California at Los Angeles, Los Angeles, California 90024. This research was supported in part by the National Aeronautics and Space Administration under Grant NGR-40-002-015 and in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under Grant AF-AFOSR-693-67.

$|u|_{D(\sigma)} = |u|_E + |\sigma u|_E$. The operator τ is required to satisfy the following assumption.

ASSUMPTION 1. *$D(\sigma) \subseteq D(\tau)$ and the restriction of τ to $D(\sigma)$ is continuous*[1] (with respect to the graph norm of $D(\sigma)$).

In applications, σ will usually be a suitable linear partial differential operator in E, τ a differential operator acting in S, the boundary of I. Finally, we introduce the space of controls. It will consist of (an appropriate subspace of) $\mathfrak{L} = \mathfrak{L}_I \times \mathfrak{L}_S$, $\mathfrak{L}_I$ (respectively $\mathfrak{L}_S$) the space of all functions $f_I(\cdot)$ (respectively $f_S(\cdot)$) defined and infinitely differentiable in $t \geqq 0$ with values in a complex Banach space F_I (respectively F_S). We shall call F_I, F_S the *control spaces* of the system.

The boundary-distributed parameter control system we shall consider is

$$u'(t) = \sigma u(t) + B_I f_I(t), \tag{1.1}$$

$$\tau u(t) = B_S f_S(t). \tag{1.2}$$

Here B_I (respectively B_S) is a bounded linear operator from F_I to E (respectively from F_S to X). A *solution* of (1.1)-(1.2) in $t \geqq 0$ is an E-valued continuously differentiable function defined in $t \geqq 0$, such that $u(t) \in D(\sigma)$ for all $t \geqq 0$ and (1.1)-(1.2) hold everywhere.

Define

$$D(A) = \{u \in D(\sigma) \mid \tau u = 0\}, \qquad Au = \sigma u \quad \text{in} \quad D(A). \tag{1.3}$$

The operator A so obtained should satisfy the following assumptions.

ASSUMPTION 2. *$D(A)$ is dense in E and $\rho(A)$, the resolvent set of A, is non-void.*

ASSUMPTION 3. *The Cauchy problem for the equation*

$$u'(t) = Au(t) \tag{1.4}$$

is uniformly well-posed in $[0, \infty)$.

Recall [6, §1] that this means the following:

(a) there exists a dense subspace D of E such that if $u \in D$ there exists a solution $u(\cdot)$ of (1.4) with $u(0) = u$;

(b) if $\{u_n(\cdot)\}$ is a sequence of solutions of (1.4) with $u_n(0) \to 0$, then $u_n(\cdot) \to 0$ uniformly on compact subsets of $[0, \infty)$.

We shall not dwell here upon the problem of singling out those A for which Assumptions 2 and 3 hold (see Remark 1.4). We shall use, however, the fact that the subspace D can be assumed to coincide with $D(A)$.

Let the operator-valued function $T(t)$, $t \geqq 0$, be defined by

$$T(t)u = u(t)$$

[1] Continuity of τ will be only occasionally used in what follows.

for $u \in D$ ($u(\cdot)$ the solution of (1.4) with $u(0) = u$) and extended to all of E by continuity. It follows easily from (a), (b) and an approximation argument that $T(\cdot)$ is strongly continuous in $t \geqq 0$. By means of $T(\cdot)$ we can construct solutions of the inhomogeneous equation

$$u'(t) = Au(t) + g(t). \tag{1.5}$$

In fact, if $g(\cdot)$ is continuously differentiable in $t \geqq 0$, then

$$u(t) = \int_0^t T(t - s)g(s)\, ds$$

is a solution of (1.5) with initial data $u(0) = 0$ (see [12, Lemma 6.1]). Consequently, if $u \in D(A)$, then

$$u(t) = T(t)u + \int_0^t T(t - s)g(s)\, ds \tag{1.6}$$

is a solution of (1.5) with $u(0) = u$. Uniqueness of solutions of (1.5) follows immediately from uniqueness for the homogeneous equation, i.e., from (b) following Assumption 3.

All these considerations will allow us to construct solutions of the original problem (1.1)-(1.2); we must, however, impose an additional requirement on B_S.

ASSUMPTION 4. *There exists a bounded operator* $B: F_S \to E$ *such that*:

(i) *if* $f \in F_S$, *then* $Bf \in D(\sigma)$ *and*

$$\tau(Bf) = B_S f; \tag{1.7}$$

(ii) *there exists a constant* C *such that*[2]

$$|Bf|_E \leqq C|B_S f|_X, \qquad f \in F_S.$$

Note that if F_S is m-dimensional, $m < \infty$ and $B_S f = B_S(f_1, \cdots, f_m) = f_1 x_1 + \cdots + f_m x_m$, $x_1, \cdots, x_m$ linearly independent vectors in X, then Assumption 4 reduces to: *there exist* $u_1, \cdots, u_m \in E$ *such that* $\tau u_j = x_j$, $1 \leqq j \leqq m$.

Now let $(f_I, f_S) \in \mathfrak{L}$ and let $v(\cdot)$ be any solution of

$$v'(t) = Av(t) - Bf_S'(t) + \sigma B f_S(t) + B_I f_I(t) \tag{1.8}$$

(by the closed graph theorem, σB is a bounded operator and thus we can construct such a solution by means of the expression (1.6)). Set

$$u(t) = v(t) + Bf_S(t). \tag{1.9}$$

An easy computation shows that $u(\cdot)$ is a solution of (1.1)-(1.2); con-

[2] Part (ii) is not essential and it is only introduced to simplify some later computations.

versely, if $u(\cdot)$ is such a solution and $v(\cdot)$ is given by (1.9), it is not difficult to show that v satisfies (1.8). The initial conditions are of course related by

$$u(0) = v(0) + Bf_S(0).$$

Summarizing, we have the following theorem.

THEOREM 1.1. *Let $u \in D(\sigma)$ be such that $\tau u \in B_S F_S$, $(f_S, f_I) \in \mathfrak{L}$ such that $B_S f_S(0) = \tau u$. Then the (unique) solution of (1.1)-(1.2) with $u(0) = u$ is given by (1.9), where $v(\cdot)$ is the solution of (1.8) with $v(0) = u(0) - Bf_S(0)$.*

The uniqueness part follows from uniqueness of the solution of (1.5) with given initial data, which in turn reduces to the same question for the homogeneous equation (1.4). Observe that if $u(0) = 0$, $B_S f_S(0) = 0$ and thus $Bf_S(0) = 0$.

We close this section with two results concerning (1.4) which are to be used later (see [1, Chap. VIII]).

LEMMA 1.2. *There exist constants K, $\omega < \infty$ such that*

$$|T(t)| \leqq Ke^{\omega t}, \qquad t \geqq 0, \tag{1.10}$$

if $\operatorname{Re}\lambda > \omega$, $\lambda \in \rho(A)$ *and*

$$R(\lambda; A)u = \int_0^\infty e^{-\lambda t} T(t)u \, dt, \qquad u \in E. \tag{1.11}$$

Note that (1.11) implies

$$\lambda R(\lambda; A)u \to u \tag{1.12}$$

as $\operatorname{Re}\lambda \to \infty$, $u \in E$.

LEMMA 1.3. *Let E^* be the dual space of E, A^* the dual of A. If E is reflexive, then the Cauchy problem for*

$$(u^*)' = A^* u^* \tag{1.13}$$

is uniformly well-posed. If $T^*(t)$ is the operator defined from the solution of (1.13) as $T(t)$ is defined from the solutions of (1.4), then

$$T^*(t) = T(t)^*.$$

Remark 1.4. If A satisfies Assumption 2, then Assumption 3 is equivalent to requiring that A should generate a *strongly continuous semigroup* $U(t)$; moreover, $U(\cdot)$ coincides with $T(\cdot)$ (see [1, Chap. VIII]). In all the examples we shall consider in §5, however, H is a Hilbert space and A is *self-adjoint*. Then Assumption 2 is automatically satisfied; as for Assumption 3, it is equivalent to the requirement that A should be *semibounded above*, i.e., that the set $\sigma(A)$ in the real line should be bounded above. The

operator $T(\cdot)$ is given by

$$T(t) = \exp(tA),$$

where the expression on the right-hand side can be computed by means of the usual functional calculus for self-adjoint operators [2, Chap. XII, §2].

2. The boundary control system (second order equations). Here we consider, instead of the system (1.1)-(1.2) of §1, the second order system

$$u''(t) = \sigma u(t) + B_I f_I(t), \tag{2.1}$$

$$\tau u(t) = B_S f_S(t). \tag{2.2}$$

The definition of the solution of (2.1)-(2.2) is similar to that for (1.1)-(1.2). Assumptions 1, 2 and 4 are supposed to hold. The operator A, defined by (1.3), is required to satisfy the following instead of Assumption 3.

ASSUMPTION 3′. *The Cauchy problem for the equation*

$$u'' = Au(t) \tag{2.3}$$

is uniformly well-posed in $[0, \infty)$.

This means now the following (see [6, §1]):

(i) there exists a dense subspace D of E such that if $u_0, u_1 \in D$ there exists a solution of (2.3) with $u(0) = u_0, u'(0) = u_1$;

(ii) if $\{u_n(\cdot)\}$ is a sequence of solutions of (2.3) with $u_n(0) \to 0$, $u_n'(0) \to 0$, then $u_n(\cdot) \to 0$ uniformly on compacts of $[0, \infty)$.

The reader may consult [6, §§5–6] for further properties of operators satisfying these hypotheses; we shall only use the fact that the subspace D may be taken equal to $D(A)$ (see also Remark 2.4).

The role of $T(t)$ in §1 is now played by two strongly continuous operator-valued functions $S(t)$, $T(t)$, $t \geqq 0$, defined as follows: If $u \in D$, $u(\cdot)$ (respectively $v(\cdot)$) is the solution of (2.3) with initial data $u(0) = u$, $u'(0) = 0$ (respectively $v(0) = 0, v'(0) = u$), then define

$$S(t)u = u(t), \quad T(t)u = v(t)$$

and extend S, T to all of E by continuity. Clearly S, T are strongly continous and are related by

$$T(t)u = \int_0^t S(s)u\, ds, \qquad t \geqq 0, \quad u \in E. \tag{2.4}$$

As for first order equations, solutions of the inhomogeneous equation

$$u''(t) = Au(t) + g(t) \tag{2.5}$$

can be constructed by means of $T(\cdot)$. In fact, if $g(\cdot)$ is twice continuously differentiable in $t \geqq 0$, then

$$u(t) = \int_0^t T(t - s)g(s)\, ds \tag{2.6}$$

is a solution of (2.5) with initial data $u(0) = u'(0) = 0$, the proof of this fact being essentially similar to the one for first order equations. Consequently, the solution of (2.5) with initial data $u(0) = u_0$, $u'(0) = u_1$, u_0, $u_1 \in D(A)$, can be written

$$u(t) = S(t)u_0 + T(t)u_1 + \int_0^t T(t - s)g(s)\, ds.$$

Uniqueness of the solutions of (2.5) follows from uniqueness for the solutions of the homogeneous equation (2.3), which, in turn, is a consequence of part (ii) of Assumption 3'. The correspondence between solutions of (2.1)-(2.2) and the solution of the inhomogenous equation (2.5) is now as follows: If $v(\cdot)$ is a solution of

$$v''(t) = Av(t) - Bf_S''(t) + \sigma Bf_S(t) + B_I f_I(t), \tag{2.7}$$

then

$$u(t) = v(t) + Bf_S(t) \tag{2.8}$$

is a solution of (2.1)-(2.2) and vice versa. This leads us to the following theorem.

THEOREM 2.1. *Let u_0, $u_1 \in D(\sigma)$ be such that τu_0, $\tau u_1 \in B_S F_S$, $(f_S, f_I) \in \mathfrak{L}$ such that $B_S f_S(0) = \tau u_0$, $B_S f_S'(0) = \tau u_1$. Then the (unique) solution of (2.1)-(2.2) with $u(0) = u_0$, $u'(0) = u_1$ is given by (2.8), where $v(\cdot)$ is the solution of (2.7) with $v(0) = u_0 - Bf_S(0)$, $v'(0) = u_1 - Bf_S'(0)$.*[3]

The uniqueness part follows as in Theorem 1.1. Note that if $u_0 = u_1 = 0$, then $Bf_S(0) = Bf_S'(0) = 0$.

We close this section with two analogues of Lemmas 1.2 and 1.3.

LEMMA 2.2. *There exist constants K, $\omega > \infty$ such that*

$$|S(t)| \leqq Ke^{\omega t}, \quad |T(t)| \leqq Ke^{\omega t}, \qquad t \geqq 0. \tag{2.9}$$

If $\operatorname{Re} \lambda > \omega$, *then* $\lambda^2 \in \rho(A)$ *and*

$$R(\lambda^2; A)u = \frac{1}{\lambda}\int_0^\infty e^{-\lambda t}S(t)u\, dt = \int_0^\infty e^{-\lambda t}T(t)u\, dt. \tag{2.10}$$

As for first order equations, this implies

$$\lambda R(\lambda; A)u \to u \quad \text{as} \quad \operatorname{Re}\lambda \to \infty \tag{2.11}$$

[3] If $u(t)$ is a solution of (2.1)-(2.2), then it might conceivably happen that $\tau u'(0) = \tau u_1 \notin B_S F_S$ and thus this solution could not be constructed by the method above. We shall not consider here this type of solution, although it is possible to do it by means of slight refinements of the results on (2.3).

for any $u \in E$.

LEMMA 2.3. *Let E^* be the dual space of E, A^* the dual of A. If E is reflexive, then the Cauchy problem for*

$$(u^*)'' = A^*u^* \tag{2.12}$$

is uniformly well-posed. If $S^*(t)$, $T^*(t)$ are the operators defined from the solutions of (2.12) as $S(t)$, $T(t)$ are defined from the solutions of (2.3), then

$$S^*(t) = S(t)^*, \quad T^*(t) = T(t)^*.$$

Remark 2.4. If Assumption 2 is satisfied, then operators satisfying Assumption 3′ can be characterized by a property similar to semigroup generation (see [6, Theorem 5.9]). In all the examples we shall consider, however, A is self-adjoint, and in this case Assumption 3′ is also equivalent to requiring that A should be semibounded above. The operators S, T can be computed by means of the functional calculus for self-adjoint operators as follows:

$$S(t) = f(t; A), \quad T(t) = g(t; A),$$

where $f(t, \lambda) = \cosh(\lambda^{1/2}t)$, $g(t, \lambda) = \lambda^{-1/2}\sinh(\lambda^{1/2}t)$.

3. Controllability (first order equations). Let $\mathcal{E} = \{u \in D(\sigma) \mid \tau u \in B_S F_S\}$ be the subspace of possible initial data for solutions of (1.1)-(1.2). If $u \in \mathcal{E}$, it is clear that any $(f_I, f_S) \in \mathcal{L}$ such that the solution of (1.1)-(1.2) satisfies $u(0) = u$, must satisfy

$$Bf_S(0) = \tau u. \tag{3.1}$$

We shall call $\mathcal{L}_u$ the subset of $\mathcal{L}$ satisfying (3.1).

Now let $E_0 \subseteq E$ be a Banach space with norm $|\cdot|_0$. Call $\mathcal{E}_0$ the set of all $u \in \mathcal{E}$ such that, for any $(f_I, f_S) \in \mathcal{L}_u$, the solution of (1.1)-(1.2) with $u(0) = u$ belongs to E_0 for all $t > 0$. If $\mathcal{E}_0$ is nonvoid we shall say that (1.1)-(1.2) is *completely controllable* (or *E_0-completely controllable*) if, for any $u \in \mathcal{E}_0$, $v \in E_0$, $\epsilon > 0$, there exists $(f_I, f_S) \in \mathcal{L}_u$ such that the solution of (1.1)-(1.2) with $u(0) = u$ satisfies

$$|u(t_0) - v|_0 \leqq \epsilon \tag{3.2}$$

for some t_0 depending in general on u, v, ϵ. If t_0 can be chosen independently of u, v, ϵ, then we shall say that (1.1)-(1.2) is *completely controllable in time t_0*.

If the element u in the conditions above is taken to be the null element of E, then we shall say that (1.1)-(1.2) is *null controllable* (*null controllable at time t_0*).

Let K_{t_0} be the subspace of E_0 consisting of the values at $t = t_0$ of all solutions of (1.1)-(1.2) with $u(0) = 0$, $(f_I, f_S) \in \mathcal{L}_0$, and let K (the

attainable set of (1.1)-(1.2)) be defined as the union of all the K_t for $t > 0$. Then (1.1)-(1.2) is null controllable (null controllable at time t_0) if and only if $ClK = E_0$ ($ClK_{t_0} = E_0$), the closures being taken in the E_0-topology.

For the time being we shall only consider the case $E_0 = E$; we show later (Remark 3.6) how the results can be extended to some spaces E_0 that have some significance in applications. We shall also suppose that E is reflexive; although this assumption can be avoided, it simplifies some proofs. We show in Remark 3.5 how the proofs can be modified for nonreflexive E.

LEMMA 3.1. $u^* \in K^\perp (K_t^\perp)$ *if and only if*

$$\left(B^*(T^*(s) - I) - (\sigma B)^* \int_0^s T^*(r)\, dr\right)u^* = 0, \tag{3.3}$$

$$B_I{}^* T^*(s) u^* = 0 \tag{3.4}$$

for $0 \leqq s$ ($0 \leqq s \leqq t$).

Proof. It follows from Theorem 1.1 and the change of variable $t - s \to s$ in the integral (1.6) representing inhomogenous solutions of (1.5) that $u^* \in K^\perp$ if and only if

$$\langle u^*, \int_0^t T(s)(\sigma B f_S(s) + B f_S{}'(s) + B_I f_I(s))\, ds + B f_S(0)\rangle = 0 \tag{3.5}$$

for all t and all $(f_I, f_S) \in \mathfrak{L}$ with $Bf_S(t) = 0$. Integrating now by parts the first integral on the left-hand side of (3.5), passing to adjoints and exploiting the fact that

$$f_S(0) = -\int_0^t f_S{}'(s)\, ds + f_S(t),$$

we can write (3.5) in the form

$$\begin{aligned} \int_0^t \left\langle \left(B^*T^*(s) - (\sigma B)^* \int_0^s T^*(r)\, dr - B^*\right)u^*, f_S{}'(s)\right\rangle ds \\ + \langle u^*, \int_0^t T(s)\sigma B f_S(t)\, ds\rangle + \langle u^*, B f_S(t)\rangle \\ + \int_0^t \langle B_I{}^* T^*(s) u^*, f_I(s)\rangle\, ds = 0. \end{aligned} \tag{3.6}$$

Taking now $f_S(s) = \eta_S(s) f_S$, $f_I(s) = \eta_I(s) f_I$, $f_S \in F_S$, $f_I \in F_I$, η_S, η_I scalar-valued C^∞ functions such that $\eta_S(t) = 0$, we easily see that the vanishing of (3.6) for all such functions implies the validity of (3.3)-(3.4). Conversely, it is clear that if (3.3)-(3.4) hold, then (3.6) will also hold for any $(f_I, f_S) \in \mathfrak{L}$ such that $Bf_S(t) = 0$. The proof is similar for K_t.

Call $\rho_0(A)$ the connected component of $\rho(A)$ that contains the half-plane $\operatorname{Re}\lambda > \omega$ (ω the constant in Lemma 1.10). We have the following corollary.

COROLLARY 3.2. *$u^* \in K^\perp$ if and only if*

$$[(\sigma B)^* R(\lambda; A^*) - \lambda B^* R(\lambda; A^*) + B^*]u^* = 0, \tag{3.7}$$

$$B_I{}^* R(\lambda; A^*)u^* = 0 \tag{3.8}$$

for $\lambda \in \rho_0(A)$.

Proof. Equalities (3.7)-(3.8) are easily obtained from (3.3)-(3.4) for $\operatorname{Re}\lambda > \omega$ with the help of the Laplace transform expression (1.11) for the resolvent of A, and they follow for any $\lambda \in \rho_0(A)$ by analytic continuation. Conversely, (3.7)-(3.8) for $\operatorname{Re}\lambda > \omega$ imply (3.3)-(3.4) by uniqueness of Laplace transforms (see [1, pp. 626–627]).

In what follows we shall consider the distributed parameter control system

$$u'(t) = Au(t) + (R(\lambda_0; A)\sigma - \sigma R(\lambda_0; A))Bf_S(t) + B_I f_I(t), \tag{3.9}$$

where λ_0 is any element of $\rho_0(A)$. Plainly (3.9) is a system of the type considered in §1 (the boundary control is $\tau u(t) = 0$). Our purpose is now to compare the sets K_t, K with L_t, L, where L_t and L are defined with respect to (3.9) as K_t and K are with respect to (1.1)-(1.2).

THEOREM 3.3. *$ClK = ClL$.*

Proof. By using the identity $\lambda R(\lambda; A) = AR(\lambda; A) + I$ we can write (3.7) as follows:

$$((\sigma B)^* - B^* A^*)R(\lambda; A^*)u^* = 0. \tag{3.10}$$

Writing (3.10) for two different values λ, λ_0 of $\rho_0(A)$, subtracting the two equalities thus obtained and using the first resolvent equation, we obtain

$$((\sigma B)^* - B^* A^*)R(\lambda_0; A^*)R(\lambda; A^*)u^* = 0 \tag{3.11}$$

or

$$((R(\lambda_0; A)\sigma - \sigma R(\lambda_0; A))B)^* R(\lambda; A^*)u^* = 0 \tag{3.12}$$

for $\lambda \in \rho_0(A)$. Conversely, (3.10)—hence (3.7)—can be obtained from (3.11) by multiplying it by λ_0—say, real—letting $\lambda_0 \to \infty$ and using (1.12). But by Corollary 3.2, conditions (3.8)–(3.12) are necessary and sufficient for u^* to belong to $L^\perp$; thus our result follows.

Remark 3.4. The result above does not provide any relation between K_t and L_t for finite t. However, we shall be mainly interested in applying our results to equations of (abstract) parabolic type, and in this case $T^*(\cdot)$ is an *analytic* function of t in $t > 0$. Consequently, the expressions

on the left-hand sides of (3.7), (3.8), (3.12) are also analytic and can vanish in a given interval $[0, t]$, $t > 0$, if and only if they vanish identically. Combining this observation with Theorem 3.3 we have in this case,

$$ClK_t = ClK = ClL \quad \text{for any} \quad t > 0.$$

Remark 3.5. If E is not reflexive, then T^* may not be as smooth as T, A^* may not be densely defined, etc. To avoid these annoying difficulties we only have to formulate—and prove—our results in "weak" form; for instance, Corollary 3.2 would read: $u^* \in K^\perp$ *if and only if* $\langle u^*, (R(\lambda; A)\sigma B - \lambda R(\lambda; A)B + B)f_S\rangle = \langle u^*, R(\lambda; A)B_I f_I\rangle = 0$ *for all* $\lambda \in \rho_0(A), f_S \in F_S$, $f_I \in F_I$. The proofs of this and the other results require only trivial modifications.

Remark 3.6. We show in what follows how ClK can be computed in a topology considerably stronger than the topology of E—the graph topology of $D(\sigma)$. Let $Cl_\sigma BF_S$ be the closure *in* $D(\sigma)$ of the subspace BF_S. If $u \in D(A) \cap Cl_\sigma BF_S$, then $u = \lim Bf_n$, $f_n \in F_S$, $0 = \tau u = \tau(\lim_n Bf_n) = \lim \tau Bf_n = \lim B_S f_n$, and thus by part (ii) of Assumption 4, $u = 0$. Consequently, it makes sense to consider the direct sum $E_0 = D(A) \oplus Cl_\sigma BF_S$, which we shall suppose endowed with the topology that it inherits from $D(\sigma)$. Assume now that $\{u_n + v_n\}$ is a sequence in E_0 converging to some element $w \in D(\sigma)$. Then we can find a sequence $\{f_n\} \subset F_S$ such that $u_n + Bf_n$ also converges to w. But then $\tau(u_n + Bf_n) = \tau Bf_n = B_S f_n \to \tau w$ in X and, again by (ii) of Assumption 4, $\{Bf_n\}$ is a Cauchy sequence in $Cl_\sigma BF_S$. Consequently, so is $\{u_n\}$ and, since $D(A)$ and $Cl_\sigma BF_S$ are both closed in the graph topology, so is E_0. The preceding reasoning with $w = 0$ shows that if $u_n + v_n \to 0$ in E_0, then $u_n \to 0$ in $D(A)$, $v_n \to 0$ in $Cl_\sigma BF_S$; this shows that E_0 can be identified algebraically and topologically with the product $D(A) \times Cl_\sigma BF_S$. As a consequence, E_0^*, the dual space of E_0, can be algebraically and topologically identified with the product $D(A)^* \times (Cl_\sigma BF_S)^*$.

If $\eta(\cdot)$ is any E-valued continuously differentiable function, then the operator $\int_0^t T(s)\eta(s)\, ds$, $t > 0$, maps E into $D(A)$ (this can be easily seen by an integration by parts). It follows from this and from the representation of the solutions of (1.1)-(1.2) given by Theorem 1.1 that any solution of (1.1)-(1.2) with $u(0) = 0$, $(f_I, f_S) \in \mathfrak{L}_0$, belongs to E_0 for any $t > 0$. Then it makes sense to investigate null controllability of (1.1)-(1.2) in E_0.

A reasoning in the line of Lemma 3.1 (a little care should be taken in that the ranges of $B, \sigma B, B_I$ do not necessarily belong to $D(A)$) shows that $(u^*, v^*) \in D(A)^* \times (Cl_\sigma BF_S)^* = E_0^*$ belongs to $K^\perp$ if and only if

$$
\begin{aligned}
&-\langle u^*, \int_0^t \left(\int_0^s (T(r)B - \int_0^r T(q)\sigma B\, dq)\, dr \right) f_S''(s)\, ds\rangle_{D(A)} \\
&+ \langle u^*, \int_0^t \left(T(s)B - \int_0^s T(r)\sigma B\, dr \right) f_S'(t)\, ds\rangle_{D(A)} \\
&+ \langle u^*, \int_0^t T(s)\sigma B f_S(t)\, ds\rangle_{D(A)} \\
&- \langle u^*, \int_0^t \left(\int_0^s T(r)B_I\, dr \right) f_I'(s)\, ds\rangle_{D(A)} \\
&+ \langle u^*, \int_0^t T(s)B_I f_I(t)\, ds\rangle_{D(A)} \\
&+ \langle v^*, Bf_S(t)\rangle_{Cl_\sigma BF_S} - t\langle v^*, Bf_S'(t)\rangle_{Cl_\sigma BF_S} \\
&+ \langle v^*, \int_0^t Bsf_S''(s)\, ds\rangle_{Cl_\sigma BF_S} = 0
\end{aligned}
\tag{3.13}
$$

for all $(f_I, f_S) \in \mathfrak{L}$, $Bf_S(t) = 0$ and all $t > 0$. Choosing now $f_S(s) = \eta_S(s)f_S$, $f_I = \eta_I(s)f_I$, η_S, η_I, C^∞ scalar-valued functions with $\eta_S(t) = \eta_S'(t) = \eta_I(t) = 0$, we easily obtain

$$
\langle u^*, \int_0^t \left(T(s)B - \int_0^s T(r)\sigma B\, dr \right) f_S\, ds\rangle_{D(A)} - t\langle v^*, Bf_S\rangle_{Cl_\sigma BF_S} = 0, \tag{3.14}
$$

$$
\langle u^*, \int_0^t T(s)B_I f_I\, ds\rangle_{D(A)} = 0 \tag{3.15}
$$

for all $t > 0, f_S \in F_S, f_I \in F_I$. Conversely, it is immediate that the above conditions imply that (3.13) vanishes for all $(f_I, f_S) \in \mathfrak{L}$, $Bf_S(t) = 0$, i.e., that $(u^*, v^*) \in K^\perp$.

Observe now that the $D(A)$-valued functions in the left-hand sides of (3.14)-(3.15) are continuous (in the $D(A)$-topology) and bounded by a constant times $e^{\omega t}$, ω the constant in (1.10); both these facts follow easily from the formula

$$
A \int_\alpha^\beta T(s)u\, ds = (T(\beta) - T(\alpha))u.
$$

Consequently, we can multiply (3.14)-(3.15) by $e^{-\lambda t}$, $\operatorname{Re}\lambda > \omega$, and integrate in $(0, \infty)$; using (1.11) we obtain, after some integration by parts,

$$
\langle u^*, (R(\lambda; A)\sigma B - \lambda R(\lambda; A)B)f_S\rangle_{D(A)} + \langle v^*, Bf_S\rangle_{Cl_\sigma BF_S} = 0, \tag{3.16}
$$

$$
\langle u^*, R(\lambda; A)B_I f_I\rangle_{D(A)} = 0 \tag{3.17}
$$

for $\operatorname{Re}\lambda > \omega$ and thus, a fortiori for all $\lambda \in \rho_0(A)$. Conversely, (3.16)-(3.17) imply (3.14)-(3.15) by uniqueness of Laplace transforms.

All the preceding considerations show that (1.1)-(1.2) will be null controllable in E_0 if and only if the validity of (3.16)-(3.17) for $\lambda \in \rho_0(A)$ implies $u^* = v^* = 0$, which amounts to saying that the subspace of E_0 generated by all elements of the form

$$(R(\lambda; A)\sigma B - \lambda R(\lambda; A)B)f_S + Bf_S, \tag{3.18}$$

$$R(\mu; A)B_I f_I, \tag{3.19}$$

$\lambda, \mu \in \rho_0(A)$, $f_I \in F_I$, $f_S \in F_S$, should be dense in E_0. Now, since the graph norm in $D(A)$ is equivalent to the norm $|u| = |(\nu - A)u|$, ν any element of $\rho(A)$, we see that the problem reduces to showing that the subspace of $E \times Cl_\sigma BF_S$ generated by all elements of the form

$$(\nu - A)(R(\lambda; A)\sigma B - \lambda R(\lambda; A)B)f_S + Bf_S,$$

$$(\nu - A)R(\lambda; A)B_I f_I,$$

$\lambda \in \rho_0(A)$, $f_I \in F_I$, $f_S \in F_S$, is dense in $E \times Cl_\sigma BF_S$. Assume this is not the case. Then there exists $(u^*, v^*) \in E^* \times (Cl_\sigma BF_S)^*$ such that

$$\langle u^*, (\nu - A)(R(\lambda; A)\sigma B - \lambda R(\lambda; A)B)f_S\rangle_E + \langle v^*, Bf_S\rangle_{Cl_\sigma BF_S} = 0, \tag{3.20}$$

$$\langle u^*, (\nu - A)R(\lambda; A)B_I f_I\rangle = 0 \tag{3.21}$$

for all $\lambda \in \rho_0(A)$, $f_I \in F_I$, $f_S \in F_S$. Writing now (3.20)-(3.21) for two different values λ, λ_0 in $\rho(A)$, subtracting the equalities thus obtained and using some elementary identities involving the resolvent of an operator, we obtain, by making use of the fact that f_I, f_S are arbitrary,

$$((\sigma B)^* R(\lambda; A^*) - \lambda B^* R(\lambda; A^*) + B^*) \cdot (\nu - A^*)R(\lambda_0; A^*)u^* = 0, \tag{3.22}$$

$$B_I{}^* R(\lambda; A^*)(\nu - A^*)R(\lambda_0; A^*)u^* = 0 \tag{3.23}$$

for all $\lambda \in \rho_0(A)$. If we assume now that the system is null controllable we obtain from Corollary 3.2 that $(\nu - A^*)R(\lambda_0; A^*)u^* = 0$, a fortiori $u^* = 0$; inserting this value of u^* in (3.20) we obtain $v^* = 0$ as well. We have thus proved the following theorem.

THEOREM 3.7. *The system* (1.1)-(1.2) *is null controllable in* E_0 *if and only if it is null controllable in* E.

Better approximation can be obtained, for instance, by using distributed control alone. We content ourselves with stating a result (of which the proof is similar to that of Theorem 3.7).

THEOREM 3.8. *Let* $B_S = 0$ (*i.e., let the control in* (1.1)-(1.2) *be purely distributed*) *and assume*

$$B_I F_I \subseteq D(A^{m-1}), \qquad m \geq 1.$$

Then (1.1)-(1.2) *is null controllable in* E *if and only if it is null controllable in* $D(A^m)$ (A^m endowed as usual with the graph norm).

Remark 3.9. Plainly, complete controllability implies null controllability. The reverse implication may be false; however, we have the next theorem.

THEOREM 3.10. *Assume that* (1.1)-(1.2) *is null controllable in time* t_0. *Then it is completely controllable in time* t_0.

Proof. Let $u \in \mathcal{E}_0$, $v \in E_0$, and let $v(\cdot)$ be the solution of (1.1)-(1.2) with $v(0) = u$ for any $(f_I, f_S) \in \mathcal{L}_u$. Given $\epsilon > 0$, choose $(g_I, g_S) \in \mathcal{L}_0$ such that if $w(\cdot)$ is the solution of (1.1)-(1.2) with $w(0) = 0$, then we have $|w(t_0) - (v - v(t_0))| \leqq \epsilon$. But $u = v + w$ is then a solution of (1.1)-(1.2) with $u(0) = u$, $|u(t_0) - v| \leqq \epsilon$, which completes the proof.

4. Controllability (second order equations). Our definitions here parallel closely those for first order systems. $\mathcal{E} = \{u \in D(\sigma) \mid \tau u \in B_S F_S\}$ is the subspace of possible initial data for solutions of (2.1)-(2.2).[4] If $u_0, u_1 \in \mathcal{E}$, denote by $\mathcal{L}_{(u_0,u_1)}$ the class of all $(f_I, f_S) \in \mathcal{L}$ such that

$$B_S f(0) = \tau u_0, \quad B_S f'(0) = \tau u_1. \tag{4.1}$$

Now let $E_0, E_1 \subseteq E$ be Banach spaces with norms $|\cdot|_0, |\cdot|_1$, respectively. Call $\mathcal{E}_{0,1}$ the set of all pairs (u_0, u_1) in $\mathcal{E}$ such that, for any $(f_I, f_S) \in \mathcal{L}_{(u_0,u_1)}$, the solution of (2.1)-(2.2) with $u(0) = u_0$, $u'(0) = u_1$ satisfies

$$u(t) \in E_0, \quad u'(t) \in E_1$$

for all $t > 0$. If $\mathcal{E}_{0,1}$ is nonvoid, we say that (2.1)-(2.2) is completely controllable if, given $u_0, u_1 \in \mathcal{E}_{0,1}$, $v_0 \in E_0$, $v_1 \in E_1$, $\epsilon > 0$, there exists $(f_I, f_S) \in \mathcal{L}_{(u_0,u_1)}$ such that

$$|u(t_0) - v_0|_0 \leqq \epsilon, \quad |u'(t_0) - v_1|_1 \leqq \epsilon \tag{4.2}$$

for some $t_0 > 0$. The definitions of complete controllability in time t_0, null controllability, etc. are all similar to those for first order equations.

We now define K_{t_0} to be the subspace of the product space $E_0 \times E_1$ consisting of all pairs $(u(t), u'(t))$, $u(t)$ a solution of (2.1)-(2.2) with $u(0) = u'(0) = 0$, $(f, f) \in \mathcal{L}_{(0,0)}$; K, the attainable set of (2.1)-(2.2), is, as before, the union of all the K_t for $t > 0$. The system (2.1)-(2.2) will be null controllable (null controllable in time t_0) if and only if $ClK = E_0 \times E_1$ ($ClK_{t_0} = E_0 \times E_1$). The closures are taken in the product topology of $E_0 \times E_1$.

As in §3 we shall first treat the case $E_0 = E_1 = E$ and assume that E is reflexive; we indicate later (Remarks 4.5-4.6) how to treat the general case.

LEMMA 4.1. $(u_0^*, u_1^*) \in K^\perp (K_t^\perp)$ *if and only if*

$$\left(B^*(T^*(s) - sI) - (\sigma B)^* \int_0^s (s - r) T^*(r)\, dr\right) u_0^* + \left(B^*(S^*(s) - I) - (\sigma B)^* \int_0^s (s - r) S^*(r)\, dr\right) u_1^* = 0, \tag{4.3}$$

[4] See footnote 3.

$$B_I^*(T^*(s)u_0^* + S^*(s)u_1^*) = 0, \qquad 0 \leqq s \quad (0 \leqq s \leqq t). \tag{4.4}$$

Proof. Let us observe first that if $u(\cdot)$ is defined by the integral (2.6), then its first derivative can be written

$$u'(t) = \int_0^t S(t-s)g(s)\,ds.$$

This observation, combined with the representation for solutions of (2.1)-(2.2) furnished by Theorem 2.1, makes it clear that $(u_0^*, u_1^*) \in K^\perp$ if and only if

$$\begin{aligned} &\langle u_0^*, \int_0^t T(s)(\sigma B f_S(s) - Bf_S''(s) + B_I f_I(s))\,ds + Bf_S(0)\rangle \\ &+ \langle u_1^*, \int_0^t S(s)(\sigma B f_S(s) - Bf_S''(s) + B_I f_I(s))\,ds - Bf_S^1(0)\rangle = 0 \end{aligned} \tag{4.5}$$

for all $t > 0$ and all $(f_I, f_S) \in \mathfrak{L}$ with $Bf(t) = Bf^1(t) = 0$. As in the proof of Lemma 3.1, by integration by parts and by observing that

$$f_S(0) = f_S(t) - tf_S'(0) - \int_0^t (t-s)f_S''(s)\,ds,$$

$$f_S'(0) = f_S'(t) - \int_0^t f_S''(s)\,ds,$$

we can transform (4.5) as follows:

$$\begin{aligned} &\int_0^t \Big\langle \Big((\sigma B)^* \int_0^s (s-r)T^*(r)\,dr \\ &\qquad - B^*T^*(s) - (t-s)B^* \Big) u_0^*, f_S''(s) \Big\rangle\,ds \\ &+ \langle u_0^*, \int_0^t T(s)\sigma B f_S(t)\,ds - \int_0^t (t-s)T(s)\sigma B f_S'(t)\,ds\rangle \\ &+ \langle u_0^*, Bf_S(t)\rangle - t\langle u_0^*, Bf_S'(0)\rangle \\ &+ \int_0^t \Big\langle \Big((\sigma B^*) \int_0^s (s-r)S^*(r)\,dr - B^*S^*(s) \\ &\qquad + B^* \Big) u_1^*, f_S''(s) \Big\rangle\,ds \\ &+ \langle u_1^*, \int_0^t S(s)\sigma B f_S(t)\,ds - \int_0^t (t-s)S(s)\sigma B f_S'(t)\,ds\rangle \\ &- \langle u_1^*, Bf_S'(t)\rangle + \int_0^t \langle B_I^*T^*(s)u_0^*, f_I(s)\rangle\,ds \\ &+ \int_0^t \langle B_I^*S^*(s)u_1^*, f_I(s)\rangle\,ds = 0. \end{aligned} \tag{4.6}$$

Now, as customary, we insert in (4.6) functions of the form $f_S = \eta_S f_S$, $f_I = \eta_I f_I$, $f_S \in F_S$, $f_I \in F_I$, η_S, η_I, C^∞ scalar-valued functions. We get immediately (4.4) by setting $\eta_S = 0$. Observing now that (4.6) must vanish when $\eta_I = 0$, $\eta_S(t) = \eta_S'(t) = \eta_S'(0) = 0$, we obtain

$$\left((\sigma B)^* \int_0^s (s-r)T^*(r)\,dr - B^*T^*(s) - (t-s)B^*\right)u_0^* + \left((\sigma B)^* \int_0^s (s-r)T^*(r)\,dr - B^*S^*(s) + B^*\right)u_1^* = u^* \tag{4.7}$$

for all $s \geqq 0$ and some fixed element $u^* \in E^*$. The value of u^* can be easily computed by setting $s = 0$ in (4.7); we get $u^* = -tB^*u_0^*$. Inserting this value of u^* in (4.7) we get (4.3). Conversely, it is easy to see that if (4.7) holds, then (4.6) must vanish for any f_S with $Bf_S'(t) = Bf_S(t) = 0$. The proof is similar for K_t.

COROLLARY 4.2. *$(u_0^*, u_1^*) \in K^\perp$ if and only if*

$$((\sigma B)^*R(\lambda; A^*) - \lambda B^*(\lambda; A^*) + B^*)(u_0^* + \sqrt{\lambda}u_1^*) = 0, \tag{4.8}$$

$$B_I^*R(\lambda; A^*)(u_0^* + \sqrt{\lambda}u_1^*) = 0 \tag{4.9}$$

for all $\lambda \in \rho_0(A)$, where $\sqrt{\lambda}$ denotes the square root of λ satisfying $\arg\sqrt{\lambda} = \frac{1}{2}\arg\lambda$, $-\pi < \arg\lambda \leqq \pi$.

Proof. Equality (4.8) can be obtained (say, for $\lambda > \omega^2$, ω the constant in (2.9)) by multiplying (4.3) by $e^{-\sqrt{\lambda}t}$, integrating in $(0, \infty)$ and applying (2.10). It follows for the rest of $\rho_0(A)$ by analytic continuation. Conversely, (4.8) for $\lambda > \omega^2$ implies (4.3) by uniqueness of Laplace transforms. The same considerations apply to (4.9).

We shall consider in what follows operators A with spectrum satisfying the following assumption.

ASSUMPTION 5. *There exists a simple closed curve C entirely contained in $\rho_0(A)$ and such that the origin is contained in the (bounded) region of which C is boundary.*

Under this condition it is easy to see that ClK has a rather simple structure, as shown in the following theorem.

THEOREM 4.3. *Suppose Assumption 5 holds. Then $ClK = \{(u_0, u_1) \in E \times E \mid u_0, u_1 \in ClL\}$, where L is the attainable set of the system (3.9).*[5]

Proof. Obviously, the result we wish to prove is equivalent to

$$K^\perp = \{(u_0^*, u_1^*) \in E^* \times E^* \mid u_0^*, u_1^* \in L^\perp\}.$$

Assume then $(u_0^*, u_1^*) \in K^\perp$. Write (4.8) for any $\lambda \in C$. As L turns once along C around the origin, (4.8) continues to hold by analytic continuation, but $\sqrt{\lambda}$ changes sign; adding up the two versions of (4.8) thus obtained we get

[5] Here we are implicitly using the fact that if A satisfies Assumption 3', then it also satisfies Assumption 3 (see [6, §5] for a proof).

$$((\sigma B)^*R(\lambda;\, A^*) - \lambda B^*R(\lambda;\, A^*) + B^*)u_0^* = 0, \tag{4.10}$$

$$((\sigma B)^*R(\lambda;\, A^*) - \lambda B^*R(\lambda;\, A^*) + B^*)u_1^* = 0 \tag{4.11}$$

for $\lambda \in C$ and thus, by analytic continuation, for all $\lambda \in \rho_0(A)$. In a similar way, we can obtain from (4.9) the two equalities

$$B_I^*R(\lambda;\, A^*)u_0^* = 0, \tag{4.12}$$

$$B_I^*R(\lambda;\, A^*)u_1^* = 0 \tag{4.13}$$

for $\lambda \in \rho_0(A)$. But (4.10)–(4.12) (respectively (4.11)–(4.13)) are just necessary and sufficient conditions, by virtue of Corollary 3.2 and Theorem 3.3, for u_0^* (respectively u_1^*) to belong to $L^\perp$. Conversely, it is clear that if u_0^*, u_1^* belong to $L^\perp$, then (u_0^*, u_1^*) belongs to $K^\perp$.

Remark 4.4. As for first order equations, Theorem 4.3 does not provide any information on K_t for finite t. We shall see (§6, Example 2), however, that in some cases of importance in applications we can assert $ClK_t = ClK$ for all t greater than or equal to a certain finite t_0.

Remark 4.5. If E is not reflexive, then all the results can be stated and proved in their "weak" form (see Remark 3.5).

Remark 4.6. We consider a situation in which ClK can be computed in a topology stronger than the product topology of $E \times E$. The notations and definitions are the same as the ones in Remark 3.6.

Let $E_0 = D(A) \oplus Cl_\sigma BF_S$ (same topology as in Remark 3.6), $E_1 = E$. If $\eta(s)$ is any E-valued continuously differentiable function, then the operator $\int_0^t T(s)\eta(s)\, ds$, $t > 0$, maps E into $D(A)$, and from this it follows in the same way as for first order equations that any solution of (2.1)-(2.2) belongs to E_0 for all $t > 0$, i.e., satisfies the assumptions in the definition of complete controllability (see §2).

Now let $(u_0^*, v^*, u_1^*) \in (E_0 \times E_1)^* = D(A)^* \times (Cl_\sigma BF_S)^* \times E^*$. A (somewhat tedious) computation similar to the one in Remark 3.6 shows that $(u_0^*, v^*, u_1^*) \in K^\perp$ if and only if

$$\begin{aligned} &\langle u_0^*, \int_0^t \left(T(s)B - \int_0^s (s-r)T(r)\sigma B\, dr \right) f_S\, ds \rangle_{D(A)} - \frac{t^2}{2} \langle v^*, Bf_S \rangle_{Cl_\sigma BF_S} \\ &\quad + \langle u_1^*, \int_0^t \left((S(s) - I)B - \int_0^s (s-r)S(r)\sigma B\, dr \right) f_S\, ds \rangle_E = 0, \end{aligned} \tag{4.14}$$

$$\langle u_0^*, \int_0^t T(s)B_I f_I\, ds \rangle_{D(A)} + \langle u_1^*, \int_0^t S(s)B_I f_I\, ds \rangle_E = 0 \tag{4.15}$$

for all $t > 0$, $f_S \in F_S$, $f_I \in F_I$. Making use now of the fact that if $u \in E$, then

$$A \int_\alpha^\beta T(s)u\, ds = (S(\beta) - S(\alpha))u,$$

and of arguments similar to those in Remark 3.6, we see that we can multiply (4.14), (4.15) by $e^{\sqrt{\lambda}t}$, $\lambda > \omega^2$ (ω the constant in (2.9)), and integrate the resulting expression in $(0, \infty)$. Making use of (2.10) we obtain

$$\begin{aligned}(4.16)\qquad &\langle u_0^*, (R(\lambda; A)\sigma B - \lambda R(\lambda; A)B)f_S\rangle_{D(A)} + \langle v^*, Bf_S\rangle_{Cl_\sigma BF_S} \\ &\quad + \sqrt{\lambda}\,\langle u_1^*, (R(\lambda; A)\sigma B - \lambda R(\lambda; A)B + B)f_S\rangle_E = 0,\end{aligned}$$

$$(4.17)\qquad \langle u_0^*, R(\lambda; A)B_I f_I\rangle_{D(A)} + \sqrt{\lambda}\,\langle u_1^*, R(\lambda; A)B_I f_I\rangle_E = 0$$

for all such λ and thus for all $\lambda \in \rho_0(A)$, $f_S \in F_S$, $f_I \in F_I$. By using now the same type of reasoning employed in Theorem 4.3 we see that if A satisfies Assumption 5, then

$$\begin{aligned}(4.18)\qquad &\langle u_0^*, (R(\lambda; A)\sigma B - \lambda R(\lambda; A)B)f_S\rangle_{D(A)} + \langle v^*, Bf_S\rangle_{Cl_\sigma BF_S} \\ &\quad = \langle u_1^*, (R(\lambda; A)\sigma B - \lambda R(\lambda; A)B + B)f_S\rangle_E \\ &\quad = \langle u_0^*, R(\lambda; A)B_I f_I\rangle_{D(A)} \\ &\quad = \langle u_1^*, R(\lambda; A)B_I f_I\rangle_E = 0\end{aligned}$$

for all $\lambda \in \rho_0(A)$, $f_S \in F_S$, $f_I \in F_I$. We are now in the same situation encountered at the end of Remark 3.6 (see (3.17)–(3.18)), and thus our reasoning ends in exactly the same way. We can formulate the following theorem.

THEOREM 4.7. *Let Assumption* 5 *hold for* $\sigma(A)$. *Then the system* (2.1)-(2.2) *is null controllable in* $E_0 \times E$ *if and only if it is null controllable in* $E \times E$.

We can improve somewhat this result. In fact, it can be shown [6, §6] that under additional conditions on A, $T(t)$, there exist square roots $(\lambda_0 - A)^{1/2}$ of $\lambda_0 - A$, where $\operatorname{Re}\lambda_0 > \omega$, and that $\int_0^t S(s)\eta(s)\,ds$ maps E into $D((\lambda_0 - A)^{1/2})$ for any E-valued smooth η and is strongly continuous as a $D((\lambda_0 - A)^{1/2})$-valued function. Then we may consider null controllability in $E_0 \times E_1$, $E_1 = D((\lambda_0 - A)^{1/2}) + BF_S$. The result is the same as Theorem 4.7, and the proofs are very similar.

If we wish better approximation we may obtain it by using distributed control alone, as for first order equations. We can show that the following theorem holds (see also [5, Remark 2.6]).

THEOREM 4.8. *Let* $B_S = 0$ (*i.e., let the control applied to* (2.1)-(2.2) *be purely distributed*). *Assume that*

$$B_I F_I \subseteq D(A^{m-1}), \qquad m \geqq 1.$$

Assume further that A *satisfies Assumption* 5 *and that* (2.1)-(2.2) *is null controllable in* $E \times E$. *Then it is null controllable in* $D(A^m) \times D(A^{m-1})$.

The proof is entirely similar to that of Theorem 4.7.

Remark 4.9. For second order systems, the concepts of complete and null controllability at time t_0 are also equivalent; in fact, we have the following result.

THEOREM 4.10. *Assume that* (2.1)-(2.2) *is null controllable in time* t_0. *Then it is completely controllable in time* t_0.

The proof is almost identical to that of Theorem 3.10. To obtain a similar result for null controllability we have to introduce a special assumption.

ASSUMPTION 6. Let $\epsilon > 0$. Then there exists $\delta > 0$ such that if $(u_0, u_1) \in \mathcal{E}_{0,1} \cap (E_0 \times E_1)$, $|u_0|_0, |u_1|_0 \leqq \delta$, then there exists an $(f_I, f_S) \in \mathcal{L}_{(u_0,u_1)}$ such that the solution of (2.1)-(2.2) with $u(0) = u_0$, $u'(0) = u_1$ satisfies

$$|u(t)|_0 \leqq \epsilon, \quad |u'(t)|_1 \leqq \epsilon, \qquad t \geqq 0.$$

THEOREM 4.11. *Assume that Assumption* 6 *holds and that* (2.1)-(2.2) *is null controllable. Then it is completely controllable.*

The proof is an immediate consequence of the fact that the solutions of (2.1)-(2.2) can be "run backwards" in time and of Assumption 6. For, if we can steer the system from (0, 0) to the vicinity of $(u_0, -u_1)$, then we can, making use of Assumption 6, steer it from (u_0, u_1) to the vicinity of (0, 0) and from there, making use again of null controllability and of Assumption 6, to the vicinity of an arbitrary pair (v_0, v_1) in $E_0 \times E_1$.

5. Examples (first order equations).

5.1. Example 1. We begin by considering the equation

$$u_t(x; t) = (\sigma u(\cdot\,; t))(x), \tag{5.1}$$

where σ is the formally symmetric formal differential operator with real coefficients [2, Chap. XIII]:

$$\sigma = (-1)^{n-1} \sum_{k=0}^{n} \left(\frac{d}{dx}\right)^k \left(a_k(x) \left(\frac{d}{dx}\right)^k\right), \qquad n \geqq 1, \tag{5.2}$$

with coefficients smooth in $[0, \infty)$, $a_n(x) > 0$ for $x \geqq 0$. We take $E = L^2(0, \infty)$; $D(\sigma)$, the domain of σ, will be the set of all $u \in E$ such that (5.2)—understood in the sense of distributions—belongs to E. It is not difficult to verify that σ is closed [2, Chap. XIII, §2.10]. We wish to exert control with a finite number of parameters, and only in the boundary of I; in order to apply the results of §3 we shall have to examine what kind of boundary conditions can be imposed in order to have A satisfy Assumption 3. We shall assume that σ has no boundary values at ∞ [2, Chap. XIII, §2], i.e., that a self-adjoint restriction of σ can be obtained by imposing boundary conditions only at 0.

Recall that any boundary condition at 0 can be written [2, Chap. XIII, Corollary 2.23]

$$\tau_j u = \sum_{\alpha=0}^{2n-1} \tau_{j\alpha} u^{(\alpha)}(0) = 0, \tag{5.3}$$

and let q, $1 \leqq q \leqq 2n-1$, be the number of (linearly independent) boundary conditions necessary to obtain a self-adjoint restriction of σ; we may suppose they are given by (5.3), i.e., $1 \leqq j \leqq q$ there. Thus we may take $X = \mathfrak{C}^q$ ($\mathfrak{C}$ = complex numbers),

$$\tau u = (\tau_1 u, \cdots, \tau_q u). \tag{5.4}$$

The control space F_S will be $\mathfrak{C}^p$, for some $p \geqq 1$ as yet unspecified, and thus the operator B_S will have the expression

$$(B_S(f_1, \cdots, f_p))_j = \sum_{\nu=1}^{p} b_{j\nu} f_\nu, \qquad 1 \leqq j \leqq q, \tag{5.5}$$

for suitable coefficients $b_{j\nu}$, $1 \leqq j \leqq q$, $1 \leqq \nu \leqq p$. All these considerations show that the control will be exerted as follows:

$$\tau_j u(\cdot\,; t) = \sum_{\nu=1}^{p} b_{j\nu} f_\nu, \qquad 1 \leqq j \leqq q. \tag{5.6}$$

The auxiliary operator B will be given by

$$B(f_1, \cdots, f_p) = \sum_{\nu=1}^{p} b_\nu(\cdot) f_\nu$$

with $b_1, \cdots, b_p$ functions in $D(\sigma)$; to satisfy Assumption 4 we must have[6]

$$\sum_{\alpha=0}^{2n-1} \tau_{j\alpha} \left(\sum_{\nu=1}^{p} b_\nu^{(\alpha)}(0) f_\nu \right) = \sum_{\nu=1}^{p} b_{j\nu} f_\nu, \qquad 1 \leqq j \leqq q;$$

thus,

$$\sum_{\alpha=0}^{2n-1} \tau_{j\alpha} b_\nu^{(\alpha)}(0) = b_{j\nu}, \quad 1 \leqq j \leqq q, \quad 1 \leqq \nu \leqq p. \tag{5.7}$$

It is not difficult to see that functions b_ν satisfying (5.7) can be constructed[7]; in fact, this is a consequence of the following auxiliary lemma.

AUXILIARY LEMMA. *If $a_0, a_1, \cdots, a_r$ are arbitrary complex numbers,*

[6] If the columns of the matrix $\{b_{j\nu}\}$, $1 \leqq \nu \leqq p$, $1 \leqq j \leqq q$, are linearly dependent, then B might not satisfy part (ii) of Assumption 4; however, this will not be employed in what follows.

[7] The system (5.7) can always be solved for the $b_\nu^{(\alpha)}(0)$; this follows from the fact that the boundary conditions are independent; thus rank $\{\tau_{j\alpha}\}$, $1 \leqq j \leqq q$, $0 \leqq \alpha \leqq 2n-1$, equals q.

then there exists $f \in C^\infty[0, \infty)$ *with compact support such that* $f^{(\alpha)}(0) = a_\alpha$, $0 \leqq \alpha \leqq r$.

We have at this stage verified, explicitly or implicitly, all the necessary assumptions with the exception of Assumption 3. Since A is self-adjoint, it reduces to (see Remark 1.4) the requirement that A should be semi-bounded above.

We shall later need some more information about the number and nature of the boundary conditions (5.3). To this end, we shall use the well-known Green's formula [2, Chap. XIII, §2]: if $u(\cdot)$ and $v(\cdot)$, say, belong to $C^\infty[0, \infty)$ and one of them vanishes for sufficiently large x, then

$$\int_0^\infty (u\overline{\sigma v} - \bar{v}\sigma u)\, dx = \sum_{\alpha,\beta=0}^{2n-1} A_{\alpha\beta} u^{(\alpha)}(0)\bar{v}^{(\beta)}(0), \tag{5.8}$$

where the matrix $\mathfrak{A} = \{A_{\alpha\beta}\}_{0\leqq\alpha,\beta\leqq 2n-1}$ is a skew-symmetric, nonsingular matrix whose coefficients depend only on σ. We shall also introduce some vector notations: τ_j, $1 \leqq j \leqq q$, will denote the vector $(\tau_{j\alpha})_{0\leqq\alpha\leqq 2n-1}$ in the unitary space $\mathfrak{C}^{2n}$, while arbitrary elements of $\mathfrak{C}^{2n}$ will be designated by ξ, η, etc. The space N will be the subspace of $\mathfrak{C}^{2n}$ consisting of all ξ that satisfy

$$(\tau_1, \bar{\xi}) = \cdots = (\tau_q, \bar{\xi}) = 0,$$

i.e., all ξ that "satisfy the boundary conditions (5.3)". It is easy to see from the fact that A is symmetric, from Green's formula, (5.4) and from the fact that, by virtue of the auxiliary lemma, any vector $\xi \in \mathfrak{C}^{2n}$ can be the "boundary values" of a $u \in C^\infty[0, \infty)$ with compact support, that

(i) if $\xi, \eta \in N$, then $(\mathfrak{A}\xi, \eta) = 0$;

similarly, the fact that A is self-adjoint implies that

(ii) if $(\mathfrak{A}\xi, \eta) = 0$ for all $\eta \in N$, then $\xi \in N$.

These two properties of N can be summarized as

$$\mathfrak{A}N = N^\perp. \tag{5.9}$$

Since $\mathfrak{A}$ is nonsingular, (5.9) implies $\dim N = \dim N^\perp = n$; since the boundary conditions (5.3) are independent, $q = n$.

The next step for application of the results in §3 to our problem will be the concept of *ordered representation* [2, Chap. XIII] of a Hilbert space (in our case, E) with respect to the self-adjoint operator A. Recall that there exists a measure μ in the real line vanishing outside $\sigma(A)$, an integer m (the *multiplicity* of A), μ-measurable sets $e_m \subseteq e_{m-1} \subseteq \cdots \subseteq e_1 = \sigma(A)$ of positive measure and a set of functions (kernels) $W_1(x, \lambda), \cdots, W_m(x, \lambda)$, defined in $[0, \infty) \times (-\infty, \infty)$, each W_i vanishing for λ in the complement of e_i, measurable with respect to the product measure $\nu \otimes \mu$ (ν = Lebesgue measure), infinitely differentiable in $x \geqq 0$ for each fixed λ and such that

the map

$$(5.10) \qquad f_i(\lambda) = (Uu)_i(\lambda) = \text{l.i.m.}_{N\to\infty} \int_0^N u(x)\bar{W}_i(x,\lambda)\,dx, \quad 1 \leqq i \leqq m,$$

defines an isometric isomorphism from $L^2[0, \infty)$ onto $K = L^2(e_1, \mu) \oplus \cdots \oplus L^2(e_m, \mu)$ carrying the operator A into the operator of multiplication by λ in K. For each fixed x the kernels W_i are square-summable on Borel sets with compact closure contained in e_1, the transformation U^{-1}, inverse to U, being given by

$$(5.11) \qquad u(x) = \text{l.i.m.}_{N\to\infty} \sum_{i=1}^{m} \int_{-N}^{N} f_i(\lambda)W_i(x,\lambda)\mu\,(d\lambda)$$

(see [2, Chap. XIII, especially Theorem 5.1 and Corollary 5.2]).

We shall need a little more information about the kernels. If $(\sigma_1(x, \lambda), \cdots, \sigma_{2n}(x, \lambda))$ is a basis for the space of solutions of $\sigma u = \lambda u$ in $[0, \infty)$ (such a basis can be obtained, for instance, by means of the boundary conditions $\sigma_\beta^{(\alpha)}(0) = \delta_{\beta-1,\alpha}$, $0 \leqq \beta - 1 \leqq 2n - 1$, in which case the σ_β are C^∞ in both variables), then there exist measurable functions $a_{i\beta}(\lambda)$, $1 \leqq i \leqq m$, $1 \leqq \beta \leqq 2n$, such that

$$(5.12) \qquad W_i(x, \lambda) = \sum_{\beta=1}^{2n} a_{i\beta}(\lambda)\sigma_\beta(x, \lambda), \qquad 1 \leqq i \leqq m,$$

the functions $a_{i\beta}(\cdot)$ being μ-square summable on any Borel set with compact closure contained in $\sigma(A)$. It follows from (5.12) that $\sigma W_i = \lambda W_i$ for all λ; it can also be shown that for μ-almost all λ the kernels $W_1, \cdots, W_k$ are linearly independent in e_k, $1 \leqq k \leqq m$, (as functions of x) and satisfy all the boundary conditions defining A. But, since there are n boundary conditions, it follows that $m \leqq n$.

Consider now the linear map

$$(5.13) \qquad u \to (R(\mu; A)u)^{(\alpha)}(0)$$

from E into $\mathfrak{C}$, μ any fixed element in $\rho(A)$, α a fixed integer, $0 \leqq \alpha \leqq 2n - 1$. It is plain that (5.13) is continuous; then, so is the map

$$\{f_i\} \to (R(\mu; A)U^{-1}\{f_i\})^{(\alpha)}(0)$$

from K into $\mathfrak{C}$. Consequently, there exists an element $\{g_{\alpha 1}, \cdots, g_{\alpha m}\}$ in K such that

$$(5.14) \qquad (R(\mu; A)U^{-1}\{f_i\})^{(\alpha)}(0) = \sum_{i=1}^{m} \int_{e_i} f_i(\lambda)\bar{g}_{\alpha i}(\lambda)\mu\,(d\lambda).$$

It is not difficult to identify the $g_{\alpha i}$. Making use of the fact that U carries A into the operator of multiplication by λ, of the form of the map U^{-1}

and of the representation (5.12) for the kernels, we obtain from well-known theorems on differentiation under the integral sign that if, say, all the elements of $\{f_i\}$ have compact support, then (5.14) holds with $(\bar{\mu} - \lambda)^{-1}\bar{W}_i^{(\alpha)}(0, \lambda)$ instead of $g_{\alpha i}$; since the set of all such $\{f_i\}$ is dense in K, we have

$$(5.15)\quad g_{i\alpha}(\lambda) = (\bar{\mu} - \lambda)^{-1}\bar{W}_i^{(\alpha)}(0; \lambda), \quad 0 \leqq \alpha \leqq 2n - 1, \quad 1 \leqq i \leqq m.$$

We can now apply Theorem 3.3. Assume $u \in K^{\perp}$. Then (3.10) vanishes for all $\mu \in \rho(A)$. Observing that for any $v \in E$,

$$B^*v = ((v, b_1), \cdots, (v, b_p)),$$
$$(\sigma B)^*v = ((v, \sigma b_1), \cdots, (v, \sigma b_p)),$$

making use of the identities $A^* = A$, $AR(\mu; A) = \sigma R(\mu; A)$ and of Green's formula, we can write (3.10) as follows:

$$\begin{aligned}((B^*A^* - (\sigma B)^*)R(\mu; A)v)_\nu &= \int_0^\infty (\sigma R(\mu; A)u\bar{b}_\nu - R(\mu; A)u\sigma\bar{b}_\nu)\, dx \\ &= -\sum_{\alpha,\beta=0}^{2n-1} A_{\alpha\beta}(R(\mu; A)u)^{(\alpha)}(0)\bar{b}_\nu^{(\beta)}(0) = 0, \\ &\qquad \nu = 1, 2, \cdots, \quad \mu \in \rho_0(A).\end{aligned} \tag{5.16}$$

Making use of (5.14) and (5.15) we can write (5.16) in the form

$$(5.17)\quad \int_{e_1} (\bar{\mu} - \lambda)^{-1}F_\nu(\lambda)\mu\,(d\lambda) = 0, \qquad \nu = 1, 2, \cdots, p,$$

where

$$(5.18)\quad F_\nu(\lambda) = \sum_{i=1}^{m}\sum_{\alpha,\beta=0}^{2n-1} A_{\alpha\beta}\bar{b}_\nu^{(\beta)}(0)\bar{W}_i^{(\alpha)}(0; \lambda)f_i(\lambda),$$

with $\{f_i\} = Uu$. Observe next that $X_J(\cdot)$, the characteristic function of the open interval $J = (a, b)$, can be written [2, p. 921]

$$(5.19)\quad X_J(\lambda) = \lim_{\delta\to 0}\lim_{\epsilon\to 0}\frac{1}{2\pi i}\int_{a+\delta}^{b-\delta}((\gamma - \epsilon i - \lambda)^{-1} - (\gamma + \epsilon i - \lambda)^{-1})\, d\gamma.$$

Since $\sigma(A)$ is contained in the real axis, $\rho_0(A) = \rho(A)$ contains its complement; using (5.19) for $\mu = \gamma + \epsilon i$ we obtain from (5.19) and (5.17) after an application of Lebesgue's bounded convergence theorem that $F_1, \cdots, F_p$ vanish almost everywhere, i.e.,

$$(5.20)\quad \sum_{i=1}^{k}\left(\sum_{\alpha,\beta=0}^{2n-1} A_{\alpha\beta}\bar{b}_\nu^{(\beta)}W_i^{(\alpha)}(0; \lambda)\right)f_i(\lambda) = 0$$

μ-almost everywhere in $e_k - e_{k+1}$, $k = 1, 2, \cdots, m$ (here we have set $e_{m+1} = \varnothing$), $\nu = 1, 2, \cdots, p$. Consequently, the problem of null controllability of our system reduces to whether or not conditions (5.20) will force $\{f_1, \cdots, f_p\}$ to vanish almost everywhere. It can be proved exactly as in [4, §3] that this will happen if and only if

$$\text{rank}\left\{\sum_{\alpha,\beta=0}^{2n-1} A_{\alpha\beta}\bar{b}_\nu^{(\beta)}(0)W_i^{(\alpha)}(0;\lambda)\right\}_{1\leqq\nu\leqq p,1\leqq i\leqq k} = k, \tag{5.21}$$

$$k = 1, 2, \cdots, m,$$

μ-almost everywhere in $e_k - e_{k+1}$. Thus we have the following theorem.

THEOREM 5.1. *The control system* (5.1)–(5.6) *is null controllable if and only if* (5.21) *holds for the numbers* $b_\nu^{(\alpha)}(0)$ *defined by* (5.7).

Introduce now the vectors

$$b_\nu = \{b_\nu^{(\alpha)}(0)\}_{0\leqq\alpha\leqq 2n-1}, \quad W_i(\lambda) = \{W_i^{(\alpha)}(0;\lambda)\}_{0\leqq\alpha\leqq 2n-1} \text{ in } \mathfrak{C}^{2n}.$$

In vector notation, (5.21) can be written

$$\text{rank}\,\{(\mathcal{A}W_i(\lambda), b_\nu)\}_{1\leqq\nu\leqq p,1\leqq i\leqq k} = k. \tag{5.22}$$

Since each kernel satisfies the boundary conditions (5.3), each $W_i(\lambda)$ belongs to N. It follows from the linear independence of the kernels, from the fact that $\mathcal{A}N = N^\perp$ and from the nonsingularity of $\mathcal{A}$ that the vectors $\mathcal{A}W_1(\lambda), \cdots, \mathcal{A}W_k(\lambda)$ are linearly independent in $N^\perp$ for μ-almost all λ in $e_k - e_{k+1}$, $i \leqq k \leqq m$. But then it is a matter of elementary linear algebra to see that (5.22) will hold almost everywhere in $e_k - e_{k-1}$, $1 \leqq k \leqq m$, whenever the projections in $N^\perp$ of the vectors $b_1, \cdots, b_p$ *generate* $N^\perp$. Since these vectors are defined by the relations

$$(\tau_j, \bar{b}_\nu) = b_{j\nu}, \qquad 1 \leqq j \leqq n, \quad 1 \leqq \nu \leqq p$$

(see (5.7)), this will happen if and only if $p \geqq n$,

$$\text{rank}\,\{b_{j\nu}\}_{1\leqq j\leqq n,1\leqq\nu\leqq p} = n. \tag{5.23}$$

Thus we have the next corollary.[8]

COROLLARY 5.2. *The control system* (5.1)–(5.6) *is null controllable if* (5.23) *holds.*

We do not know at present whether there always exist m vectors $b_1, \cdots, b_m$ such that conditions (5.21) hold, although this seems to happen in every case. If this were true, we could always render (5.1)–(5.6) null controllable by means of only m control parameters (clearly, this could

[8] Actually, the numbers $b_\nu^{(\alpha)}(0)$ are not uniquely defined by (5.7); this, of course, has no effect on the validity of Theorem 5.1. Note, however, that (5.21)—or better, its vector form (5.22)—implies $p \geqq m$ and, since $\mathcal{A}N = N^\perp$ and $W_i(\lambda) \in N$, we may take the components of the $\bar{b}_\nu$ in N equal to zero, i.e., we may take $\bar{b}_\nu \in M^\perp$. Then the $\bar{b}_\nu$ are uniquely determined by $\{b_{j\nu}\}$. The same observation applies to Example 2.

never happen with less parameters). However, the verification of the null controllability conditions would involve, if $m < n$, at least some information about the kernels W_i. On the other hand, Corollary 5.2 gives a sufficient condition for null controllability that, while probably involving more parameters than necessary, does not depend at all on the particular equation under consideration, and it is also of immediate verification.[9]

We examine now a family of examples that show how Corollary 5.2 may or may not be a necessary condition for controllability. We choose σ as follows:

$$\sigma = -\left(\frac{d}{dx}\right)^4 - \rho\left(\frac{d}{dx}\right)^2$$

in $[0, \infty)$, where ρ is a real nonnegative parameter. The boundary control is exerted by means of two independent control parameters f, g as follows:

$$u(0, t) = \alpha f(t), \quad u''(0, t) = \beta g(t),$$

where α, β are two parameters as yet to be determined. It is not difficult to see that the operator A obtained from σ by imposing boundary conditions $u(0, t) = u''(0, t) = 0$ is self-adjoint and semibounded above.

The Fourier sine transform

$$\hat{u}(\sigma) = \underset{N\to\infty}{\text{l.i.m.}} \left(\frac{1}{2}\pi\right)^{-1/2} \int_0^N \sin \sigma x u(x)\, dx$$

is an isometric isomorphism from $L_\kappa^2(0, \infty)$ onto $L_\sigma^2(0, \infty)$ that transforms the operator A into the multiplication operator

$$u(\sigma) \to (-\sigma^4 + \rho\sigma^2)u(\sigma).$$

Making use of this and of elementary changes of variable we can carry out a spectral analysis of A; we have $e_1 = (-\infty, \frac{1}{4}\rho^2]$, $e_2 = [0, \frac{1}{4}\rho^2]$ and, if we define

$$\varphi \pm (\lambda) = (\tfrac{1}{2}\rho \pm (\tfrac{1}{4}\rho^2 - \lambda)^{1/2})^{1/2},$$

then the measure μ and the kernels W_1, W_2 can be given, except for constants, by

$$\mu(d\lambda) = \varphi_+(\lambda)^{-1}(\tfrac{1}{4}\rho^2 - \lambda)^{-1/2}\, d\lambda,$$

$$W_1(x, \lambda) = \sin \varphi_+(\lambda)x, \quad \lambda \in e_1,$$

$$W_2(x, \lambda) = \varphi_+(\lambda)\varphi_-(\lambda)^{-1} \sin \varphi_-(\lambda)\, x, \quad \lambda \in e_2.$$

If $\rho > 0$, $\mu(e_2) > 0$ and A has multiplicity 2, then condition (5.23),

[9] Of course, the simplest way of satisfying (5.23) is to apply independent controls in all the boundary conditions, i.e., $\tau_j u(\cdot, t) = f_j(t)$.

which in our case amounts to

$$\alpha\beta \neq 0, \tag{5.24}$$

is necessary for null controllability of (5.1)–(5.6); this follows immediately from the fact that the vectors $W_1(\lambda)$, $W_2(\lambda)$ generate the space N for each $\lambda \in e_2$. However, condition (5.23) is no longer necessary when $\rho = 0$; for in this case A has multiplicity 1 ($\mu(e_2) = 0$) and

$$W_1(x, \lambda) = \sin(-\lambda)^{1/4}x, \quad \lambda \in e_1.$$

Now, if b is any nonzero vector in C^4, it is immediate from the analyticity of W_1 in λ that

$$(\mathcal{C}W_1(\lambda), b) \neq 0$$

except for a (at most) countable number of points in $(-\infty, 0]$, i.e., μ-almost everywhere. Consequently, the condition for null controllability of (5.1)–(5.6) is now

$$|\alpha| + |\beta| > 0, \tag{5.25}$$

which is weaker than (5.24).

5.2. Example 2. We consider the equation

$$u_t(x, t) = (\sigma u(\cdot\,; t))(x). \tag{5.26}$$

Here σ is again given by the expression (5.2), but now considered in a *compact* interval $[a, b]$. $D(\sigma)$ is defined analogously as in Example 1. Recall [2, Chap. XIII, §1] that any self-adjoint restriction A of σ is obtained by imposition of $2n$ linearly independent boundary conditions

$$\tau_j u = \sum_{\alpha=0}^{2n-1} \tau_{j\alpha}(a)u^{(\alpha)}(a) + \sum_{\alpha=0}^{2n-1} \tau_{j\alpha}(b)u^{(\alpha)}(b) = 0, \quad 1 \leqq j \leqq 2n. \tag{5.27}$$

The spaces and operators are $E = L^2(a, b)$, $B_I = 0$, $X = \mathbb{C}^{2n}$, $\tau u = (\tau_1 u, \cdots, \tau_{2n} u)$, $F_S = \mathbb{C}^p$, $p \geqq 1$,

$$(B_S(f_1, \cdots, f_p))_j = \sum_{\nu=1}^{p} b_{j\nu} f_\nu, \qquad 1 \leqq j \leqq 2n; \tag{5.28}$$

thus the control is applied in the following way:

$$\tau_j u(\cdot\,; t) = \sum_{\nu=1}^{p} b_{j\nu} f_\nu, \qquad 1 \leqq j \leqq 2n. \tag{5.29}$$

The auxiliary operator B in Assumption 4 (see §1) is now given by

$$B(f_1, \cdots, f_p) = \sum_{\nu=1}^{p} b_\nu(\cdot) f_\nu,$$

where the functions $b_1, \cdots, b_p \in D(\sigma)$ and satisfy

$$\sum_{\alpha=0}^{2n-1} \tau_{j\alpha}(a) b_\nu^{(\alpha)}(a) + \sum_{\alpha=0}^{2n-1} \tau_{j\alpha}(b) b_\nu^{(\alpha)}(b) = b_{j\nu} \tag{5.30}$$

for $1 \leqq j \leqq 2n$.[10] The existence of the functions $b_1, \cdots, b_\nu$ is assured by the following analogue of the auxiliary lemma of Example 1.

AUXILIARY LEMMA. *Let $a_0, \cdots, a_r, b_0, \cdots, b_s$ be arbitrary complex numbers. Then there exists $f \in C^\infty[a, b]$ such that $f^{(\alpha)}(a) = a_\alpha, f^{(\beta)}(b) = b_\beta$, $0 \leqq \alpha \leqq r, 0 \leqq \beta \leqq s$.*

We shall also make use here of Green's formula; for the interval $[a, b]$ can be written

$$\begin{aligned} \int_a^b (u\sigma\bar{v} - \bar{v}\sigma u)\, dx = &\sum_{\alpha,\beta=0}^{2n-1} A_{\alpha\beta} u^{(\alpha)}(a) \bar{v}^{(\beta)}(a) \\ &- \sum_{\alpha,\beta=0}^{2n-1} B_{\alpha\beta} u^{(\alpha)}(b) \bar{v}^{(\beta)}(b), \end{aligned} \tag{5.31}$$

where the matrices $\mathfrak{A} = \{A_{\alpha\beta}\}$, $\mathfrak{B} = \{B_{\alpha\beta}\}$, $0 \leqq \alpha, \beta \leqq 2n - 1$, are nonsingular, skew-symmetric and independent of u, v. We shall now find it useful to employ vector notation; write

$$\xi = (\xi(a), \xi(b)) = (\xi_0(a), \cdots, \xi_{2n-1}(a), \xi_0(b), \cdots, \xi_{2n-1}(b)),$$

for vectors in $4n$-dimensional unitary space $\mathfrak{C}^{4n}$, and define the matrix $\mathfrak{S}$ by

$$\mathfrak{S}\xi = \mathfrak{S}(\xi(a), \xi(b)) = \mathfrak{A}\xi(a) - \mathfrak{B}\xi(b).$$

It is clear that $\mathfrak{S}$ is skew-symmetric and nonsingular. If we write $\tau_j u = (\tau_j(a), \tau_j(b)) = (\tau_{j0}(a), \cdots, \tau_{2n-1}(b))$ as vectors in $\mathfrak{C}^{4n}$ and define

$$N = \{\xi \in \mathfrak{C}^{4n} \mid (\tau_j, \bar{\xi}) = 0, 1 \leqq j \leqq 2n\},$$

then it is not difficult to prove, as in Example 1, that

$$\mathfrak{S}N = N^\perp. \tag{5.32}$$

Rather than continuing our reasoning as in Example 1, we shall make use of the (much simpler) structure of A when the basic interval is compact. Recall [2, Chap. XIII, §4.1] that the spectrum of A consists of a sequence of points $\lambda_1 > \lambda_2 > \cdots$ tending to $-\infty$; if $\varphi_{k,1}, \cdots, \varphi_{k,m(k)}$ is a (say, orthonormal) basis for the subspace of eigenfunctions corresponding to the eigenvalue λ_k, $k = 1, 2, \cdots$, then $m(k) \leqq 2n$ for all k and $\{\varphi_{k,j}\}$, $1 \leqq j \leqq m(k)$, $1 \leqq k$, is a basis of E. Now let $\mu \in \rho_0(A) = \rho(A)$, $u \in E$. By

[10] Here again, the existence of the $b_\nu^{(\alpha)}(a), b_\nu^{\alpha}(b)$ is assured by the independence of the boundary conditions.

Green's formula (5.31) we have

$$((B^*A^* - (\sigma B)^*)R(\mu; A^*)u)_\nu$$

$$= -\sum_{\alpha,\beta=0}^{2n-1} A_{\alpha\beta}(R(\mu; A)u)^{(\alpha)}(a)\bar{b}_\nu^{(\beta)}(a) \tag{5.33}$$

$$+ \sum_{\alpha,\beta=0}^{2n-1} B_{\alpha\beta}(R(\mu; A)u)^{(\alpha)}(b)\bar{b}_\nu^{(\beta)}(b).$$

On the other hand, since the left-hand side of (5.33) is a continuous functional of u, there should exist $c_\nu(x) = c_\nu(\mu; x)$, $1 \leqq \nu \leqq p$, in E such that

$$((B^*A^* - (\sigma B)^*)R(\mu; A^*)u)_\nu = \int_a^b u(x)\bar{c}_\nu(x)\,dx. \tag{5.34}$$

Applying now Theorem 3.3 and the results in [3] or [4] on null controllability of distributed parameter systems, we see that the condition (on the c_ν) that will guarantee null controllability of (5.26)–(5.29) is the following:

$$\text{rank } \{(\varphi_{k,j}, c_\nu)\}_{1\leqq j\leqq m(k), 1\leqq \nu\leqq p} = m(k). \tag{5.35}$$

This plainly implies $p \geqq \sup\{m(k), k = 1, 2, \cdots\}$, the multiplicity of A. The scalar products in the matrix in the left-hand side of (5.35) can be easily computed by means of (5.33) and (5.34); setting $u = \varphi_{k,j}$ in (5.33) and using the fact that $R(\mu; A)\varphi_{k,j} = (\mu - \lambda_k)^{-1}\varphi_{kj}$, we obtain

$$(\mu - \lambda_k)^{-1}(\varphi_{k,j}, c_\nu) = -\sum_{\alpha,\beta=0}^{2n-1} (A_{\alpha\beta}\varphi_{k,j}^{(\alpha)}(a)\bar{b}_\nu^{(\beta)}(a) - B_{\alpha\beta}\varphi_{k,j}^{(\alpha)}(b)\bar{b}_\nu^{(\beta)}(b))$$

or, introducing the vectors in $\mathfrak{C}^n$,

$$\psi_{k,j} = (\varphi_{k,j}(a), \cdots, \varphi_{k,j}^{(2n-1)}(a), \varphi_{k,j}(b), \cdots, \varphi_{k,j}^{(2n-1)}(b)),$$

$$b_\nu = (b_\nu(a), \cdots, b_\nu^{(2n-1)}(a), b_\nu(b), \cdots, b_\nu^{(2n-1)}(b)),$$

we can write

$$(\varphi_{k,j}, c_\nu) = (\mu - \lambda_k)(\mathcal{S}\psi_{k,j}, b_\nu).$$

Thus we have the following theorem.

THEOREM 5.3. *The control system* (5.26)–(5.29) *is null controllable if and only if*

$$\text{rank } \{(\mathcal{S}\psi_{k,j}, b_\nu)\}_{1\leqq j\leqq m(k), 1\leqq \nu\leqq p} = m(k), \quad k = 1, 2, \cdots. \tag{5.36}$$

A simple argument based on the category theorem of Baire shows that there exist m vectors $b_1, \cdots, b_m$ such that (5.3) is satisfied for all k, i.e., that the system is now controllable with only m parameters.

As in Example 1 we can obtain a sufficient condition for null controllability that, while perhaps requiring an excessive number of parameters, does not assume any knowledge of the spectral properties of A; in fact, the following corollary holds.

COROLLARY 5.4. *Assume*

$$\text{rank } \{b_{j\nu}\}_{0\leqq j\leqq 2n-1, 1\leqq \nu\leqq p} = 2n. \tag{5.37}$$

Then (5.26)–(5.29) *is null controllable.*

Corollary 5.4 can be somewhat improved for certain special types of boundary conditions. For instance, assume these are

$$\begin{aligned} u(a) = u'(a) = \cdots = u^{(n)}(a) \\ = u(b) = u'(b) = \cdots = u^{(n)}(b) = 0. \end{aligned} \tag{5.38}$$

It follows easily from the theory of ordinary differential equations that in this case the multiplicity of A is $\leqq n$. We can then render the system null controllable by applying n independent control parameters at only one point of the boundary, say, at a, i.e., the conditions

$$b_{j\nu} = 0 \quad \text{for} \quad n < j \leqq 2n, \quad 1 \leqq \nu \leqq n, \tag{5.39}$$

$$\text{rank } \{b_{j\nu}\}_{1\leqq j\leqq n, 1\leqq \nu\leqq n} = n \tag{5.40}$$

are sufficient for null controllability.

Remark 5.5. Let us see briefly how Theorem 3.7 can be applied to our case, beginning with Example 2. Let $L^{2,r} = L^{2,r}(a, b)$, $r \geqq 1$, be the space of all functions u in $L^2(a, b)$ such that $u^{(k)} \in L^2(a, b)$ (the derivative understood in the sense of distributions) normed, say, with

$$|u|_{2,r}^2 = \sum_{k=0}^{r} |u^{(k)}|_2^2,$$

under which $L^{2,r}$ becomes a Hilbert space (observe, incidentally, that $|\cdot|_{2,r}$ is equivalent to the seemingly stronger norm $|\cdot| = \sum_{k=0}^{r-1} \max_{a\leqq x\leqq b} |u^{(k)}(x)| + |u^{(r)}|_2$). If the matrix $\{b_{j\nu}\}$ satisfies (5.37) it is not difficult to see that $D(A) \oplus BF_S = L^{2,2n}$ (see [2, Chap. XIII, §2.16]; since $D(A) \oplus BF_S$ is also a Hilbert space under the graph norm used in Theorem 3.7, by the closed graph theorem, both norms have to be equivalent. Then if (5.26)–(5.29) is null controllable in L^2, the system will also be null controllable in $L^{2,2n}$. If the boundary conditions are, say, (5.38) and $\{b_{j\nu}\}$ satisfies (5.39)–(5.40), we obtain in the same way the result that (5.26)–(5.29) is null controllable in the subspace of $L^{2,2n}$ consisting of all elements u with $u(b) = \cdots = u^{(n)}(b) = 0$. As for Example 1, the space $D(A) \oplus BF_S$ is not so easily characterized but it is easily seen, say, that null controllability of (5.26)–(5.29) implies the possibility of approximating the target function in the norm $L^{2,2n}(0, b)$, $b < \infty$.

To end our treatment of ordinary differential operators we include two samples of the results obtained above.

COROLLARY 5.6. (i) *Let p, q be two C^∞ functions defined in* $[0, \infty)$. *Assume $p > 0$, q bounded above. Then the control system*

$$u_t = (pu_x)_x + qu, \quad 0 \leqq x, \quad u(0, t) = f(t)$$

is null controllable in any time $t_0 > 0$.

(ii) *Let p, q be C^∞ functions defined in $[a, b]$. Assume $p > 0$. Then the control system*

$$u_t = (pu_x)_x + qu, \quad a \leqq x \leqq b, \quad u(a, t) = \alpha f(t), \quad u(b, t) = \beta g(t)$$

is null controllable in any time $t_0 > 0$ if $|\alpha| + |\beta| > 0$.

The proof of (ii) is immediate; as for (i) we only have to verify that the restriction of $\sigma = d(p(d/dx))/dx + q$ obtained by means of the boundary condition $u(0) = 0$ is self-adjoint and semibounded above; this is taken care of by [2, Chap. XIII, §5.5 and §6.14]. The fact that there is null controllability in any finite $t_0 > 0$ follows from analyticity of T'^*; this holds for all the cases treated in this section.

Remark 5.7. All our results in Example 2—in particular, Theorem 5.3, Corollary 5.4, Remark 5.5—remain true for the case of *nonself-adjoint* α, the result on distributed parameter systems necessary to apply Theorem 3.3 being provided by [3, especially Corollary 3.3]. The proofs carry over with some modifications.

5.3. Examples 3 and 4. Our results apply also to some cases in which σ is a *partial* differential operator in a domain I of Euclidean space R^r, $r \geqq 2$. We mention two of them.

Example 3. σ is a negative elliptic, formally self-adjoint operator in a (bounded or unbounded) domain $I \subseteq R^r$ (see [2, Chap. XIV]). If A is a self-adjoint restriction of σ obtained by means of a boundary operator τ, with adequate smoothness assumptions on σ, I, then an ordered representation of E with respect to A can be constructed much in the same way as for ordinary differential operators (see [2, Chap. XIV, §6]). Note, however, that we may have to use an *infinite* number of kernels W_i, and thus we may not be able to obtain null controllability with a finite number of parameters. The reasoning in Example 1 carries over with substantial modifications—the space X is new infinite-dimensional, Green's formula is replaced by n-dimensional Gauss formula, etc. The controllability conditions will now involve integrals over the boundary of certain derivatives of the kernels W_i. Sufficient conditions of the type of Corollary 5.2—i.e., that do not require knowledge of the kernels—may be obtained in some cases (for instance, $F_S = X$, $B = I$) but they are usually far from necessary in that they involve infinite-dimensional control, while the system could be controlled by a finite number of parameters.

Example 4. σ is a negative elliptic not necessarily self-adjoint operator in a bounded domain $I \subseteq R^r$. Under a suitable smoothness assumption on σ, I, we can obtain restrictions A of σ determined by boundary operators τ that satisfy all the necessary assumptions. The reasoning in Example 2 carries over to our case, and our results can then be extended. Note, however, that the multiplicity of the eigenvalues of A may be no longer bounded, and then we may not obtain null controllability with a finite number of parameters. The observation on sufficient conditions in Example 3 applies also here.

We close this section with an example that falls between Examples 3 and 4. Let I be the rectangle $0 \leqq x_1 \leqq \pi/a_1$, $0 \leqq x_2 \leqq \pi/a_2$, $a_1, a_2 > 0$ in R^2, and let

$$\sigma = (\partial/\partial x_1)^2 + (\partial/\partial x_2)^2$$

be the Laplace operator. E is $L^2(I)$, X is (say) $L^2(S)$, S the boundary of I and τ assigns to each $u \in D(\sigma)$ its boundary values. Plainly A is self-adjoint, with eigenvalues $-(a_1^2 n_1^2 + a_2^2 n_2^2)$, $n_1, n_2 = 1, 2, \cdots$, corresponding to (nonnormalized) eigenfunctions $\sin a_1 n_1 x_1 \sin a_2 n_2 x_2$. Assume a_1^2, a_2^2 are linearly independent over the integers. Then each eigenvalue of A has multiplicity 1 and thus, by the analogue of Theorem 5.3 (and the observation following it) for partial differential operators, we should be able to control

$$u_t = \sigma u, \quad \tau u = B_S f \tag{5.41}$$

by means of only one parameter. Let then $B_S f = b_S f$, $Bf = bf$, where b_S is a function in $X = L^2(S)$, b a function in $D(\sigma)$ having b_S as boundary values. We have, for $u \in E$, $\mu \in \rho(A)$,

$$(B^*A^* - (\sigma B)^*)R(\mu; A)u = \int_I (\bar{b}\sigma R(\mu; A) - \sigma\bar{b}R(\mu; A))u \, dx_1 \, dx_2$$

$$= \int_S \left(\bar{b}_S \frac{\partial}{\partial n} R(\mu; A)u - R(\mu; A)u \frac{\partial}{\partial n} \bar{b} \right) dS,$$

dS being the line element in S and $\partial/\partial n$ = normal derivative at S. Assuming (as we may) that $\partial\bar{b}/\partial n$ vanishes on S we obtain, reasoning as in Example 2, that

$$(B^*A^* - (\sigma B)^*)R(\mu; A)u = \int_S \bar{b}_S \frac{\partial}{\partial n} R(\mu; A)u \, dS = \int_I cu \, dx_1 \, dx_2,$$

where c is some element in E. Finally, we use Theorem 4.3; setting u = eigenvectors of A and applying the result in [4, §4], we see that (5.41) will be null controllable if and only if

$$\int_S b_S \frac{\partial}{\partial n} (\sin a_1 n_1 x_1 \sin a_2 n_2 x_2) \, dS \neq 0$$

for $n_1, n_2 = 0, 1, \cdots$, or

$$
\begin{aligned}
&- a_2 n_2 \int_0^{\pi/a_1} b_S(x_1, 0) \sin a_1 n_1 x_1 \, dx_1 \\
&+ (-1)^{n_1} a_1 n_1 \int_0^{\pi/a_2} b_S\left(\frac{\pi}{a_1}, x_2\right) \sin a_2 n_2 x_2 \, dx_2 \\
&- (-1)^{n_2} a_2 n_2 \int_0^{\pi/a_1} b_S\left(x_1, \frac{\pi}{a_2}\right) \sin a_1 n_1 x_1 \, dx_1 \\
&+ a_1 n_1 \int_0^{\pi/a_2} b_S(0, x_2) \sin n_2 a_2 x_2 \, dx_2 \neq 0,
\end{aligned} \tag{5.42}
$$

$$n_1, n_2 = 0, 1, \cdots .$$

Observe that if we drop the condition of linear independence of a_1, a_2 over the integers, our result becomes false; in fact, if, say, $a_1 = a_2$, then A has eigenvalues of arbitrary high multiplicity. Then $u_t = Au + B_I f_I$ cannot be null controllable when $\dim F_I < \infty$ (see [4, §4]) and, in view of Theorem 4.3, neither can the system (5.41).

6. Examples (second order equations).

6.1. Example 1. Since Theorem 4.3 reduces the problem of null controllability of second order control systems to the corresponding one for first order systems, all the results obtained for these systems in §5 will apply as long as A satisfies Assumption 5. For self-adjoint A, Assumption 5 reduces to the requirement that $\rho(A)$ $(= \rho_0(A))$ should intersect the negative real axis; this is always true for the system examined in Example 2 of §5 but not necessarily for the ones in Example 1. We shall see in the following example that if Assumption 5 fails to hold, then the conclusion of Theorem 4.3 may become false.

We shall consider the boundary control system

$$u_{tt} = u_{xx} + \rho u, \quad x \geqq 0, \quad u(0, t) = f(t) \tag{6.1}$$

in $E = L^2(0, \infty)$. The operator A is

$$Au = u_{xx} + \rho u$$

with boundary condition $u(0) = 0$. We have $\sigma(A) = (-\infty, \rho]$; thus if $\rho \geqq 0$, Assumption 5 is not satisfied. It is not difficult to see that if $u(x, t)$ is a solution of (6.1), then its Fourier sine transform (in x) $\hat{u}(\sigma, t)$ satisfies

$$\hat{u}_{tt}(\sigma, t) = (-\sigma^2 + \rho)\hat{u}(\sigma, t) + \sigma f(t), \qquad \sigma \geqq 0. \tag{6.2}$$

Now (6.2) is a "distributed parameter" system that would be of the type considered in [5] were it not for the fact that the function $u(\sigma) = \sigma$ that

multiplies the control in (6.3) does not belong to L^2; however, we only have to modify slightly a similar example in [5]. A simple computation shows that the solution of (6.2) with $u(\sigma, 0) = u_t(\sigma, 0) = 0$ for $\sigma \geqq 0$ can be written[11] for $|\sigma| \geqq \rho^{1/2}$,

$$\begin{aligned} u(\sigma, t) &= \sigma(\sigma^2 - \rho)^{-1} f(t) \\ &\quad - \sigma(\sigma^2 - \rho)^{-3/2} \int_0^t \sin\,(\sigma^2 - \rho)^{1/2}(t - s) f''(s)\, ds. \end{aligned} \tag{6.3}$$

To show that (6.2)—hence (6.1)—is not null controllable, we only have to exhibit a pair $(u_0(\cdot), u_1(\cdot))$ of functions in L^2, not both zero and orthogonal to all pairs $(u(\cdot, t), u_t(\cdot, t))$, $t > 0$, $u(\cdot, t)$ given by (6.3), f an arbitrary C^∞ function with $f(0) = f'(0) = 0$. (We shall construct u_0, u_1 zero for $\sigma \leqq \rho^{1/2}$.) A simple computation shows that u_0, u_1 have to satisfy

$$\begin{aligned} &\int_{\rho^{1/2}}^\infty (t - s - (\sigma^2 - \rho)^{-1/2} \sin\,(\sigma^2 - \rho)^{1/2}(t - s))\sigma(\sigma^2 - \rho)^{-1} u_0(\sigma)\, d\sigma \\ &\quad + \int_{\rho^{1/2}}^\infty (1 - \cos\,(\sigma^2 - \rho)^{1/2}(t - s))\sigma(\sigma^2 - \rho)^{-1} u_1(\sigma)\, d\sigma = 0 \end{aligned} \tag{6.4}$$

for $0 \leqq s \leqq t$, $t \geqq 0$. Assume now $\sigma(\sigma^2 - \rho)^{-1} u_i(\sigma)$, $u_i(\sigma)$, $i = 0, 1$, are summable in $(\rho^{1/2}, \infty)$. By means of the change of variable $(\sigma^2 - \rho)^{1/2} = \eta$ and of elementary trigonometric identities, (6.4) can be written

$$(t - s) \int_0^\infty \tilde{u}_0(\eta)\, d\eta + \int_0^\infty \tilde{u}_1(\eta)\, d\eta + \frac{1}{2} \int_{-\infty}^\infty e^{i\eta(t-s)} h(\eta)\, d\eta = 0, \tag{6.5}$$

where $\tilde{u}_i(\eta) = \eta^{-1} u_i((\eta^2 + \rho)^{1/2})$, $i = 0, 1$, and

$$h(\eta) = (i\eta^{-1}\tilde{u}_0(|\eta|) - \tilde{u}_1(|\eta|)) \operatorname{sgn} \eta.$$

In view of the preceding considerations and of the Paley-Wiener theorem [8, Chap. 8], the third integral in (6.5) will vanish identically for $s \leqq t$ if h belongs to the space H^2 of the upper half-plane, i.e., if h consists of the boundary values of a function $h(\eta + i\phi)$, holomorphic in the upper half-plane and such that

$$\sup_{\phi > 0} \int_{-\infty}^\infty |\, h(\eta + i\phi)\,|^2\, d\eta < \infty.$$

Thus, to construct u_0, u_1 we only have to find a function $h(\eta)$ in H^2 such that $\eta^2 h(\eta)$ is in $L^2(-\infty, \infty)$ and such that if

$$\begin{aligned} \tilde{u}_0(\eta) &= -\tfrac{1}{2} i\eta(h(-\eta) + h(\eta)), \\ \tilde{u}_1(\eta) &= \tfrac{1}{2}(h(-\eta) - h(\eta)), \end{aligned}$$

[11] If $f(0) = f'(0) = 0$; see §4.

then both $\tilde{u}_0$, $\tilde{u}_1$ have null integrals over $(0, \infty)$. This can be achieved by requiring the integrals of $h(\eta)$, $\eta h(\eta)$ to vanish in $(-\infty, 0)$ and $(0, \infty)$, and it is a simple matter to check that such an h, $h \not\equiv 0$, can in fact be constructed. Thus the system (6.1) is *not* null controllable for $\rho > 0$.

It should be noted that the above result can be proved in a much easier way in the particular case $\rho = 0$; for the solution of the wave equation $u_{tt} = u_{xx}$ in $x \geqq 0$ satisfying $u(x, 0) = u_t(x, 0) = 0$, $u(0, t) = f(t)$ with $f(0) = f'(0) = 0$ is given by

$$u(x, t) = \begin{cases} f(t - x) & \text{if } x \leqq t, \\ 0 & \text{if } x \geqq t. \end{cases}$$

Thus the pairs $(u(\cdot, t), u_t(\cdot, t))$ are $(f(t - x), f'(t - x))$ if $x < t$, $(0, 0)$ if $x \geqq t$, and it is clear that they will only approximate pairs of the form $(g(x), g'(x))$.

6.2. Example 2. For some boundary-distributed parameter control systems that are important in applications it is possible to show that null controllability is equivalent to null controllability at some time $t_0 > 0$. We limit ourselves to an example. Consider the control system (in $\boldsymbol{L}^2(a, b)$):

$$u_{tt} = (pu_x)_x + qu, \quad a \leqq x \leqq b, \quad u(a, t) = \alpha f_0(t), \quad u(b, t) = \beta f_1(t), \tag{6.6}$$

where p and q are in $C^\infty[a, b]$, $p > 0$. The operator A obtained from σ by imposing null boundary conditions is self-adjoint and with pure point spectrum, each eigenvalue being of multiplicity 1; thus it can be written

$$Au = \sum_{n=1}^{\infty} \lambda_n (\varphi_n, u)\varphi_n, \qquad u \in E,$$

$\lambda_1 > \lambda_2 > \cdots > \lambda_n \to -\infty$, φ_n the (normalized) eigenfunction corresponding to the eigenvalue λ_n. Without loss of generality we can assume $\lambda_1 < 0$; then the operators S, T of §2 can be written

$$S(t)u = \sum_{n=1}^{\infty} \cos \mu_n t (\varphi_n, u)\varphi_n, \quad T(t) = \sum_{n=1}^{\infty} \mu_n^{-1} \sin \mu_n t (\varphi_n, u)\varphi_n,$$

where we have set $\mu_n = (-\lambda_n)^{1/2}$. This expression and a few simple computations show that the equality (4.3) in Lemma 4.1 can be written in our case as follows:

$$\sum_{n=0}^{\infty} (\mu_n^{-1} \sin \mu_n s - s)(\varphi_n, u_0)(B^* - (\sigma B)^* A^{-1})\varphi_n + \sum_{n=0}^{\infty} (\cos \mu_n s - 1)(\varphi_n, u_1)(B^* - (\sigma B)^* A^{-1})\varphi_n = 0, \tag{6.7}$$

where

$$(B^* - (\sigma B)^* A^{-1})\varphi_n = \mu_n^{-2}(\alpha\varphi_n'(a), \beta\varphi_n'(b)). \tag{6.8}$$

We shall make use in what follows of some well-known asymptotic estimates for the eigenvalues and eigenfunctions of A (see [18, Chap. IV]). We have

$$\mu_n = Kn + O(1/n), \tag{6.9}$$

where $K = \pi\left(\int_a^b p(x)^{-1/2}\,dx\right)^{-1}$, and we have

$$|\varphi_n'(a)| = Cn + O(1), \quad |\varphi_n'(b)| = Cn + O(1), \tag{6.10}$$

where C is a positive constant. Now let N be a positive integer and let us define two new (doubly infinite) sequences $\{\mu_n'\}$, $\{\mu_n''\}$, $-\infty < n < \infty$, as follows: $\mu_n' = \mu_n \operatorname{sgn} n$ for $n \neq 0$, $\mu_0' = 0$, $\mu_n'' = \mu_n'$ for $|n| \geqq N$, $\mu_n'' = n$ for $|n| < N$. Plainly there exists an N such that $|\mu_n'' - Kn| \leqq K\pi^{-1}\log 2$ for all n. Then it follows from [13, Chap. V] that the set $\{e^{i\mu_n''x}\}$ is a basis of $L^2(0, 2\pi K^{-1})$, i.e., (a) $\{e^{i\mu_n''x}\}$ generates $L^2(0, 2\pi K^{-1})$, (b) $\{e^{i\mu_n''x}\}$ fails to generate $L^2(0, 2\pi K^{-1})$ as soon as any of its elements is removed. But then, by [17, Chap. III, §V] we can alter at will a finite number of elements of $\{\mu_n''\}$ without losing the basis property, i.e., $\{e^{i\mu_n'x}\}$ is as well a basis of $L^2(0, 2\pi K^{-1})$. Applying now a result in [17, Chap. I] we see that there exists a biorthogonal system associated with $\{e^{i\mu_n'x}\}$, i.e., a sequence $\{\varphi_n(x)\}$ of functions in $L^2(0, 2\pi K^{-1})$ such that

$$(\varphi_n, e^{i\mu_m(\cdot)}) = \delta_{mn}, \qquad m, n \geqq 1.$$

Return now to the expression (6.7). By (6.8) and the estimate (6.10) it can be differentiated (in the L^2-sense) term by term, and the derivative can be written as a series in the functions $e^{i\mu_n'x}$. Thus, if (6.7) vanishes in $[0, 2\pi K^{-1}]$, so does its derivative; taking scalar products with the φ_n we see that all its coefficients vanish, and thus it is identically zero in $[0, \infty)$. Since (6.7) vanishes for $s = 0$, it also vanishes identically. Thus we have the following theorem.

THEOREM 6.1. *If the system* (6.6) *is null controllable* (*which happens if and only if* $|\alpha| + |\beta| > 0$), *then it is null controllable in time* t_0 *whenever*

$$t_0 \geqq 2\pi K^{-1} = 2\int_a^b p(x)^{-1/2}\,dx. \tag{6.11}$$

This result is best possible, as can easily be verified with the control system (6.6), if we set $p = 1$, $q = 0$. In the same example, however, we find an interesting phenomenon: if *both* α and β are not equal to 0 (i.e., if control is applied at both boundaries), then a result similar to Theorem 6.1 holds

but with $t_0 \geqq \pi K^{-1}$. We do not know whether or not this happens in the general case.[12]

A result similar to Theorem 6.1, but with *strict* inequality in (6.11), could have been obtained with much less information about the asymptotic behavior of the sequence $\{\mu_n\}$. In fact, assume that we are able to compute

$$\Delta(\{\mu_n\}) = \lim_{\eta\to\infty} \limsup_{\xi\to\infty} \eta^{-1} m(\xi, \xi + \eta) = K^{-1}, \qquad 0 \leqq K \leqq \infty, \tag{6.12}$$

where $m(\xi, \xi + \eta)$ is the number of elements of the sequence $\{\mu_n\}$ in $[\xi, \xi + \eta)$, and also assume that

$$\min_{m\neq n} |\mu_n - \mu_m| > 0. \tag{6.13}$$

Then we can apply the theory of pseudoperiodic functions (see [9]). According to this theory, if J_1, J_2 are intervals in the real line, $J_1 \subseteq J_2$, length $(J_1) > 2\pi K^{-1}$, and $H(\{\mu_n\}, J_i)$, $i = 1, 2$, is the subspace generated in $L^2(J_i)$ by $\{e^{i\mu_n x}\}$, then the L^2 norms in $H(\{\mu_n\}, J_i)$, $i = 1, 2$, are comparable; thus the same is true of the subspaces $H(\{\mu_n\}, J_i) + e$, e the function $e(s) = s$. Consequently, if (6.7) vanishes in $[0, \delta]$, $\delta > 2\pi K^{-1}$, it vanishes identically. The result is obviously independent of the particular form of the system and depends only on the fact that A is self-adjoint and has pure point spectrum. The expression (6.12) can be computed, for instance, when A is a self-adjoint partial differential operator by means of Gårding's asymptotic estimates [7]. Note, however, that these estimates take into account multiplicity of the eigenvalues, while we do not, and also they do not guarantee (6.13). Plainly, all the preceding observations hold as well for distributed control.

We shall not treat here approximation in norms other than the L^2 norm; the reader will find no difficulty in establishing an analogue of Remark 5.5 for our case. We end this section with a result that constitutes the analogue of Corollary 5.6 (ii) for second order equations.

COROLLARY 6.2. *Let p, q be C^∞ functions defined in $[0, \infty)$. Assume $p > 0$, q bounded above. Assume, further, that the spectrum of the self-adjoint operator $Au = (pu_x)_x + qu$, $u(0) = 0$ does not contain all the negative real axis. Then the control system*

$$u_{tt} = (pu_x)_x + qu, \quad x \geqq 0, \quad u(0, t) = f(t) \tag{6.14}$$

is null controllable.

Observe that it is in general false that (6.14) should be null controllable

[12] This was observed in [15] for certain symmetric hyperbolic systems. See also [16], where a distributed parameter control problem for equations of the type (6.6) is considered; the emphasis there is in attaining, not only approximating, the target functions.

in any finite time t_0 ; for, if we set $p = 1$, $q = \text{const.} < 0$, the differential equation in (6.14) is the telegraph equation. But then the "disturbance" $f(t)$ propagates along the x-axis with finite speed, and this means that the functions approximated in any fixed finite time t_0 will have support contained in a fixed compact set. On the other hand, we can force A to have pure point spectrum by choosing adequately p, q, and thus we could be in the situation of Theorem 6.1, i.e., we could control the system in finite time.

6.3. Example 3. Our observations for the case in §5 where $\sigma =$ partial differential operators have evident analogues here, via Theorem 4.3. We limit ourselves to a sample result: using the notations and definitions of §5.3, the control system

$$u_{tt} = \sigma u, \quad \tau u = b_S f \tag{6.15}$$

is null controllable if and only if conditions (5.42), n_1 , $n_2 = 1, 2, \cdots$, are satisfied for b_S .

It might be remarked that, in contrast with the case of one space variable, (6.15) is *not* null controllable in any finite time.

REFERENCES

[1] N. Dunford and J. T. Schwartz, *Linear Operators*, vol. I, Interscience, New York, 1958.

[2] ——— , *Linear Operators*, vol. II, Interscience, New York, 1963.

[3] H. O. Fattorini, *Some remarks on complete controllability*, Siam J. Appl. Math., 4 (1966), pp. 686–693.

[4] ——— , *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[5] ——— , *Controllability of higher order linear systems*, Proc. Conference on Mathematical Theory of Control at the University of Southern California, 1967, Academic Press, New York, 1967, pp. 301–311.

[6] ——— , *Ordinary differential equations in linear topological spaces*, J. Differential Equations, to appear.

[7] L. Gårding, *On the asymptotic distribution of the eigenvalues and eigenfunctions of elliptic differential operators*, Math. Scand., 1 (1953), pp. 237–255.

[8] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[9] J. P. Kahane, *Pseudo-periodicité et séries de Fourier lacunaires*, Ann. Sci. École Norm. Sup., 79 (1962), pp. 93–150.

[10] R. Lattès, *Control of boundary conditions for non-stationary problems*, Presented in the Conference on Mathematical Theory of Control at the University of Southern California, 1967.

[11] T. L. Lions, *Control problems and partial differential equations*, Proc. Conference on Mathematical Theory of Control at the University of Southern California, 1967, Academic Press, New York, 1967, pp. 251–271.

[12] R. S. Phillips, *Perturbation theory for semi-groups of linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[13] F. Riesz and B. Sz.-Nagy, *Leçons d'analyse fonctionelle*, 3rd ed. Gauthier-Villars, Paris, 1955.

[14] D. L. RUSSELL, *Optimal regulation of linear symmetric hyperbolic systems with finite-dimensional controls*, Siam J. Appl. Math., 4 (1966), pp. 276–294.

[15] ———, *On boundary value control of linear symmetric hyperbolic systems*, Proc. Conference on the Mathematical Theory of Control at the University of Southern California, 1966, Academic Press, New York, 1967, pp. 312–321.

[16] ———, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967) pp. 542–560.

[17] L. SCHWARTZ, *Etude des sommes d'exponentielles*, 2nd ed., Hermann, Paris, 1959.

[18] F. TRICOMI, *Equazioni differenziali*, Einaudi, Torino, 1948.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

EXISTENCE AND STABILITY THEOREMS FOR EXTERIOR BALLISTICS*

W. R. HASELTINE†

1. Introduction. In [1] and [2] the author studied the equation

$$\ddot{z} + (-i\omega + \epsilon H)\dot{z} - (M + \epsilon m + i\omega\epsilon T)z = G, \tag{1}$$

which is known to describe rather well the yawing motion of symmetrical projectiles. In particular, it was established that if G/M is small and $\epsilon = 1$, the periodic solution predicted heuristically really exists, and that if both G/M and ϵ are small, the heuristic predictions concerning the existence of mixed-mode steady states is also correct.

Further discussion of (1) seems desirable for several reasons.

(a) In flights of real shells there have been observed cases of steady conical yaw, and other cases of steady mixed oscillations, under conditions not covered by some of the restrictions. Numerical solution of the more complete differential equations from which (1) is derived seems to show that these cases can best be explained by a T which depends strongly on $|z|^2$, but with a value of G/M which is not small on the scale of the nonlinearity of Tz.

(b) The basic theorems needed to establish the existence of a steady mixed oscillation, even under the restrictions of [2], were published in [3], which is not readily available to everyone.

(c) In the last sentence of [2] a statement was made about the condition governing stability of the mixed mode. But, so far as the author knows, nothing has been published which would serve as justification for that statement. Nevertheless, as we shall see, the statement is correct.

2. Existence. To be explicit about the symbols in (1) we note that:

(i) z is complex;

(ii) ω and M are real nonzero numbers, and $\omega^2 - 4M > 0$ (the case $\omega^2 \leqq 4M$ is unstable when $\epsilon = 0$, and that together with $\omega = 0$ needs to be handled separately);

(iii) G is a complex number, and we set $-G/M = \gamma e^{i\alpha}$, γ real; H, m, T are real, C^2, functions of $|z|^2$, $d(|z|^2)/dt$ and $|\dot{z}|^2$.

Questions concerning the existence and stability of the singular point $z = \text{const.} = \gamma e^{i\alpha} + O(|\epsilon|)$ are handled adequately in [2]. The general solution for $\epsilon = 0$ of (1) is

$$z = \gamma e^{i\alpha} + r_1 e^{i\varphi_1} + r_2 e^{i\varphi_2},$$

* Received by the editors July 19, 1967, and in revised form February 28, 1968.

† Code 60704, Naval Weapons Center, China Lake, California 93555.

with

$$\dot{\varphi}_i = \nu_i\,,$$

$$\nu_i = \tfrac{1}{2}(\omega \pm \sqrt{\omega^2 - 4M}), \quad r_i \text{ real.}$$

Note that $\nu_i \neq 0$, and $\nu_1 \neq \pm\nu_2$.

To study steady conical yaw, we try

$$\begin{aligned} z &= \gamma e^{i\alpha} + re^{i\varphi} + x, \\ \dot{z} &= i\nu_1(re^{i\varphi} + \sigma x), \\ r &\neq 0, \quad \sigma = \nu_2/\nu_1\,, \end{aligned} \tag{2}$$

insert (2) into (1) and ask whether the resulting system has a solution of the form

$$\text{(A)} \qquad r = r_0 + \epsilon s(\varphi, \epsilon), \qquad x = \epsilon v(\varphi, \epsilon), \qquad \varphi = \eta(l, \epsilon),$$

where s and v have period 2π in φ. Similarly, for the mixed mode we try

$$\begin{aligned} z &= \gamma e^{i\alpha} + r_1 e^{i\varphi_1} + r_2 e^{i\varphi_2}, \\ \dot{z} &= i\nu_1(r_1 e^{i\varphi_1} + \sigma r_2 e^{i\varphi_2}), \\ \sigma &= \nu_2/\nu_1\,, \qquad r_1\,, r_2 \neq 0, \end{aligned} \tag{3}$$

and ask whether the system has a solution of the form

$$\begin{aligned} r_i &= r_{0i} + \epsilon s_i(\varphi_1\,, \varphi_2\,, \epsilon), \\ \varphi_i &= \eta_i(t, \epsilon), \end{aligned} \tag{B}$$

and s_i has period 2π in φ_1 and in φ_2. In neither case has a choice yet been made as to which sign of the radical belongs to ν_1.

Inserting (2) in (1), we find ($x = x_1 + ix_2$)

$$\begin{aligned} \dot{\varphi} &= \nu_1 + \epsilon\Phi_1(\varphi, r, x), \\ \dot{x}_1 &= -\sigma\nu_1 x_2 + \epsilon X_1(\varphi, r, x), \\ \dot{x}_2 &= \sigma\nu_1 x_1 + \epsilon X_2(\varphi, r, x), \\ \dot{r} &= \epsilon R^1(\varphi, r, x), \end{aligned} \tag{4}$$

where Φ_1, X_1, X_2 are periodic in φ, but of form not concerning us at the moment, and

$$R^1 = \frac{1}{\nu_1(1 - \sigma)}\,[(-H\nu_1 + \omega T)r + \omega T\gamma \cos(\alpha - \varphi) + (-H\sigma\nu_1 + \omega T)(x_1 \cos\varphi + x_2 \sin\varphi)].$$

Note that

$$|z|^2 = \{r^2 + x_1^2 + x_2^2 + \gamma^2 + 2rx_1\cos\varphi + 2rx_2\sin\varphi + 2r\gamma\cos(\alpha - \varphi) + 2\nu_1\gamma x_1\cos\alpha + 2\gamma x_2\sin\alpha\},$$

$$\frac{d}{dt}|z|^2 = 2\nu_1[-(1-\sigma)rx_1\sin\varphi + (1-\sigma)rx_2\cos\varphi + \gamma r\sin(\alpha - \varphi) + \sigma\gamma x_1\sin\alpha - \sigma\gamma x_2\cos\alpha],$$

$$|\dot{z}|^2 = \nu_1^2[r^2 + \sigma^2 x_1^2 + \sigma^2 x_2^2 + 2\sigma rx_1\cos\varphi + 2\sigma rx_2\sin\varphi].$$

By using the methods of [4], we find that to have solutions of form (2) a necessary condition is that

$$\int_0^{2\pi} R^1(\varphi, r_0, 0)\, d\varphi = 0,$$

and, given this, the following conditions are sufficient:

$$\int_0^{2\pi} \left[\frac{\partial}{\partial r} R^1(\varphi, r, 0)\right]_{r=r_0} d\varphi \neq 0,$$

$$\det(W(2\pi) - I) \neq 0,$$

with

$$W(\varphi) = \begin{pmatrix} \cos\sigma\varphi & \sin\sigma\varphi \\ -\sin\sigma\varphi & \cos\sigma\varphi \end{pmatrix}_0.$$

But $\det(W(2\pi) - I) = 2(1 - \cos 2\pi\sigma)$, so that unless σ is an integer, the existence of our desired family of solutions depends only on the properties of the function R^1. For a simple example, consider:

$$m \equiv 0, \qquad H = \text{const.}, \qquad T = T_1(|z|^2 - \delta_0^2), \quad T_1 \neq 0.$$

Then

$$(\nu_1 - \nu_2)\int_0^{2\pi} R^1(\varphi, r, 0)\, d\varphi = 2\pi[-H\nu_1 + \omega T_1\{r^2 + 2\gamma^2 - \delta_0^2\}]r.$$

Here our conditions are all satisfied if

$$0 = -H\nu_1 + \omega T_1(r_0^2 + 2\gamma^2 - \delta_0^2), \qquad 2\omega T_1 r_0^2 \neq 0,$$

and σ is not an integer.

Now whether or not σ is integral, we can try the transformation

$$x = ye^{i\sigma\varphi}. \tag{5}$$

We then have the system

$$
\begin{aligned}
\dot{\varphi} &= \nu_1 + \epsilon\Phi, \\
\psi &= \sigma\varphi, \\
\dot{y}_1 &= \epsilon[Y_1 + \sigma\Phi y_2], \\
\dot{y}_2 &= \epsilon[Y_2 - \sigma\Phi y_1], \\
\dot{r} &= \epsilon R,
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
\Phi = \frac{1}{r\nu_1(1-\sigma)}\,[&-mr + \omega T\gamma \sin(\alpha-\varphi) - m\{y_1\cos(\psi-\varphi) \\
&- y_2 \sin(\psi-\varphi)\} - m\gamma\cos(\alpha-\varphi) \\
&+ (-H\sigma\nu_1 + \omega T)\{y_1\sin(\psi-\varphi) + y_2\cos(\psi-\varphi)\}],
\end{aligned}
$$

$$
\begin{aligned}
R = \frac{1}{\nu_1(1-\sigma)}\,[&(-H\nu_1 + \omega T)r + \omega T\gamma\cos(\alpha-\varphi) + m\{y_1\sin(\psi-\varphi) \\
&+ y_2\cos(\psi-\varphi)\} + m\gamma\sin(\alpha-\varphi) \\
&+ (-H\sigma\nu_1 + \omega T)\{y_1\cos(\psi-\varphi) - y_2\sin(\psi-\varphi)\}],
\end{aligned}
$$

$$
\begin{aligned}
Y_1 = \frac{1}{\nu_1(\sigma-1)}\,[&(-H\sigma\nu_1 + \omega T)y_1 + (-H\nu_1 + \omega T)r\cos(\varphi-\psi) \\
&+ \omega T\gamma\cos(\alpha-\psi) + my_2 \\
&+ mr\sin(\varphi-\psi) + m\gamma\sin(\alpha-\psi)],
\end{aligned}
$$

$$
\begin{aligned}
Y_2 = \frac{1}{\nu_1(\sigma-1)}\,[&(-H\sigma\nu_1 + \omega T)y_2 + (-H\nu_1 + \omega T)r\sin(\varphi-\psi) \\
&+ \omega T\gamma\sin(\alpha-\psi) - my_1 \\
&- mr\cos(\varphi-\psi) - m\gamma\cos(\alpha-\psi)].
\end{aligned}
$$

Here

$$
\begin{aligned}
|z|^2 = [&r^2 + y_1^2 + y_2^2 + \gamma^2 + 2ry_1\cos(\varphi-\psi) + 2ry_2\sin(\varphi-\psi) \\
&+ 2r\gamma\cos(\alpha-\psi) + 2\gamma y_1\cos(\alpha-\psi) + 2\gamma y_2\sin(\alpha-\psi)]
\end{aligned}
$$

with correspondingly altered expressions for $d(|z|^2)/dt$ and $|\dot{z}|^2$. Now system (6) is a special case of the system obtained by replacing $\psi = \sigma\varphi$ by $\dot{\psi} = \sigma\nu_1 + \epsilon\sigma\Phi$. This more general system we label (6a).

Though φ, ψ, r, y_1, y_2 are not uniquely determined by z and $\dot{z}$, as are φ, r, x_1, x_2 of (4), nevertheless any solution of (6a) will, together with transformation (5), give a solution of (4).

Before studying system (6a), we shall examine the results of transformation (3). We obtain

$$
\begin{aligned}
\dot{\varphi}_1 &= \nu_1 + \epsilon\Phi_1, \\
\dot{\varphi}_2 &= \sigma\nu_1 + \epsilon\Phi_2, \\
\dot{r}_1 &= \epsilon R_1, \\
\dot{r}_2 &= \epsilon R_2,
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
\Phi_1 = \frac{1}{r_1\nu_1(1-\sigma)}\,[&-mr_1 - mr_2\cos(\varphi_2-\varphi_1) - m\gamma\cos(\alpha-\varphi_1) \\
&+ (-H\sigma\nu_1 + \omega T)r_2\sin(\varphi_2-\varphi_1) + \omega T\gamma\sin(\alpha-\varphi_1)],
\end{aligned}
$$

$$
\begin{aligned}
R_1 = \frac{1}{\nu_1(1-\sigma)}\,[&(-H\nu_1 + \omega T)r_1 + (-H\sigma\nu_1 + \omega T)r_2\cos(\varphi_2-\varphi_1) \\
&+ \omega T\gamma\cos(\alpha-\varphi_1) \\
&+ mr_2\sin(\varphi_2-\varphi_1) + m\gamma\sin(\alpha-\varphi_1)].
\end{aligned}
$$

Φ_2 and R_2 are obtained from Φ_1 and R_1, respectively, by interchanging φ_1 with φ_2, r_1 with r_2, ν_1 with $(\sigma\nu_1)$ and $\sigma\nu_1$ with ν_1, and m with $-m$ in R_2.

Both (6a) and (7) have the form

$$
\begin{aligned}
\dot{\theta} &= \omega + \epsilon\Theta(\theta, z, \epsilon), \\
\dot{z} &= \epsilon Z(\theta, z, \epsilon),
\end{aligned} \tag{8}
$$

where θ, ω, Θ are k-vectors, z and Z are l-vectors, and Θ and Z have period 2π in each θ_j. For any function $f(\theta) = f(\theta_1, \cdots, \theta_k)$ we define

$$
(f(\theta))^\dagger = \lim_{T\to\infty}\frac{1}{T}\int_0^T f(\theta_1 + \omega_1 t, \cdots, \theta_k + \omega_k t)\,dt.
$$

Note that $(f(\theta))^\dagger$ depends only on

$$
\left(\theta_j - \frac{\omega_j}{\omega_1}\theta_1\right), \quad 1 < j \leqq k.
$$

The pertinent theorem of [4] can be stated as follows.

THEOREM. *Given system* (8) *with* Θ *and* Z, C^2 *in* θ, z, ϵ, *if a vector* z_0 *can be found such that* $(Z(\theta, z_0, 0))^\dagger \equiv 0$, *and the eigenvalues of the matrix* $B^\dagger = \partial(Z_j(\theta, z, 0))^\dagger/\partial z_i$ *taken at* $z = z_0$ *all have nonzero real parts and one of the following holds*:

(a) $B^\dagger$ *is constant and*

$$
\frac{\partial C^\dagger}{\partial\theta} = \frac{\partial}{\partial\theta_j}(\Theta(\theta, z_0, 0))^\dagger \equiv 0,
$$

or

(b) $B^\dagger$ *is diagonal, all the eigenvalues of* $B^\dagger$ *are distinct for all* θ, *and* $\partial C^\dagger/\partial\theta \equiv 0$, *or*

(c) *all the eigenvalues of* $B^\dagger$ *are distinct for all* θ, *and* $C^\dagger \equiv 0$;

then there exists a continuous ϵ*-family of periodic* k*-surfaces of* (7). *That is, there is an* $\epsilon_1 > 0$, *and there is a unique* l*-vector function* $S(\theta, \epsilon)$, *of period* 2π *in each* θ_j, S *is* C^1 *in* θ, *and* C *in* (θ, ϵ), *and a* k*-vector function* $\eta(t, \epsilon)$, C^1 *in* t *and* C *in* (t, ϵ); $S(\theta, 0) \equiv z_0$; *and*

$$\text{(C)} \qquad \begin{aligned} \theta &= \eta(t, \epsilon), \\ z &= S(\eta(t, \epsilon), \epsilon) \end{aligned}$$

forms a solution of (8) *for each* ϵ, $0 < |\epsilon| < \epsilon_1$. *If*

$$\int_0^T Z(\theta, z_0, 0)\, dt$$

is bounded, then $S(\theta, \epsilon)$ *is* C^1 *in* (ϵ, θ).

In applying this theorem to (6a), with σ an integer, we must examine the structure of $R^\dagger$, $Y_1^\dagger$, $Y_2^\dagger$, and possibly $\Phi^\dagger$. For the simple example given earlier,

$$\Phi^\dagger = 0 + O(|y|),$$

$$R^\dagger = \frac{1}{\nu_1(1-\sigma)}\left[-H\nu_1 + \omega T_1(r^2 + 2\gamma^2 - \delta_0^2)\right] r,$$

$$Y_1^\dagger = \frac{1}{\nu_1(\sigma-1)}\left[-H\sigma\nu_1 + \omega T_1(2r^2 + 2\gamma^2 - \delta_0^2)\right] y_1 + O(|y|^2),$$

$$Y_2^\dagger = \frac{1}{\nu_1(\sigma-1)}\left[-H\sigma\nu_1 + \omega T_1(2r^2 + 2\gamma^2 - \delta_0^2)\right] y_2 + O(|y|^2),$$

unless $\sigma = 2$ or $\frac{1}{2}$. Thus in this special case we need, for σ integral and not equal to 2,

$$0 = -H\nu_1 + \omega T_1(r_0^2 + 2\gamma^2 - \delta_0^2)$$

as before, and

$$0 \neq -H\sigma\nu_1 + \omega T_1(2r_0^2 + 2\gamma^2 - \delta_0^2)$$

or

$$r_0^2 = H\nu_1/\omega T_1 + \delta_0^2 - 2\gamma^2$$

and

$$H\nu_1(2-\sigma) + \omega T_1(\delta_0^2 - 2\gamma^2) \neq 0.$$

Obviously this only makes sense if r_0^2 comes out greater than zero. Note that we have not required that δ^2 be greater than zero, only that it be real.

If the conditions of the theorem are met, then there are functions $S(\varphi, \psi, \epsilon)$, $U(\varphi, \psi, \epsilon)$, periodic in φ and ψ, both vanishing as ϵ approaches zero, such that $r = r_0 + S$, $x = e^{i(\sigma\varphi+\psi_0)}U$, with $\varphi = \eta(t, \varphi_0, \epsilon)$, $\varphi(0, \varphi_0, \epsilon) = \varphi_0$, form a solution of (5) (here we have set $\psi - \sigma\varphi = \psi_0$). Furthermore, S, $U \to 0$ as $\epsilon \to 0$. Since σ is an integer, S and $e^{i(\sigma\varphi+\psi_0)}U$ have period 2π in φ. The author has not been able to determine, in the general case, whether or not this formal family in ψ_0 reduces to a unique member. But, if the functions on the right-hand sides of (6) are well enough behaved so that S and U can be expressed as power series in ϵ, it is then possible to show that there is indeed only one member of the family. It is also clear that when the family exists *and is stable*, then it must reduce to a single member.

Now the theorem applies equally to cases with σ not an integer. We find then, whether σ is rational or not, $R^\dagger(\theta, \psi, r, 0) = Y^\dagger(\theta, \psi, r, 0) = 0$, so that these, which are the only *necessary* conditions, are met, and we have no contradiction with the direct approach on questions of existence. For σ rational the question of the multiplicity of the ψ_0 family of solutions again arises, and the known answers are exactly the same as for σ integral. On the other hand, for σ irrational, it is easy to show that the ψ_0 family has only a single member. Also

$$\frac{\partial R^\dagger}{\partial y_j} \equiv 0 \quad \text{and} \quad \frac{\partial Y^\dagger}{\partial y} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix},$$

a and b being real constants, so that even the sufficiency conditions fail only if either $\partial R^\dagger/\partial r = 0$ or $a = 0$.

We now examine system (7). If σ is irrational, neither the $\Psi_j^\dagger$ nor the $R_j^\dagger$ depend on φ_1 or φ_2. If σ is rational, the structure of $R_j^\dagger$ and $\Psi_j^\dagger$ must be examined in detail. In our simple example for any value of σ except 0, ± 1, $\frac{1}{2}$, or 2,

$$R_1^\dagger = \frac{r_1}{\nu(1-\sigma)}\left[-H\nu_1 + \omega T_1(r_1^2 + 2r_2^2 + 2\gamma^2 - \delta_0^2)\right],$$

$$R_2^\dagger = \frac{r_2}{\nu(\sigma-1)}\left[-H\nu_1\sigma + \omega T_1(2r_1^2 + r_2^2 + 2\gamma^2 - \delta_0^2)\right],$$

and we do not need to examine the Φ_j. The question of existence reduces to that of finding a pair r_1^2, r_2^2, both positive, such that $R_1^\dagger = R_2^\dagger = 0$, for if we can, then

$$\frac{\partial R^\dagger}{\partial r} = \frac{2\omega T_1}{\nu(1-\sigma)}\begin{pmatrix} r_1^2 & 2r_1 r_2 \\ -2r_1 r_2 & -r_2^2 \end{pmatrix}$$

and an eigenvalue of this matrix can have zero real part only if at least one of the pair r_1, r_2 is zero, or H is zero.

Let us now assemble the results obtained for the example. In doing so, it is convenient to choose the indexing so that $\sigma < 1$ (then $\nu_1/\omega > 0$) and to set $a = H\nu_1/\omega T_1$, $b = \delta_0^2 - 2\gamma^2$. Then:

(i) the singular point $z = \gamma e^{i\alpha} + O(|\epsilon|)$ is always a solution (see [2]);

and if $\sigma \neq 0, \frac{1}{2}, -1, 2$,

(iia) a periodic solution of frequency $\nu_1 + O(|\epsilon|)$ exists if $a + b > 0$ (and $a(2 - \sigma) + b \neq 0$ if σ is integral);

(iib) by symmetry, a periodic solution of frequency $\nu_2 + O(|\epsilon|) = \sigma\nu_1 + O(|\epsilon|)$ exists if $b + \sigma a > 0$ (and $b + (2\sigma - 1)a \neq 0$ if σ is integral);

(iii) a periodic 2-surface, i.e., a mixed mode, exists if we can find r_1^2 and r_2^2, both greater than zero, such that

$$\omega T_1(r_1^2 + 2r_2^2 + 2\gamma^2 - \delta_0^2) - H\nu_1 = 0$$

and

$$\omega T_1(r_2^2 + 2r_1^2 + 2\gamma^2 - \delta_0^2) - H\sigma\nu_1 = 0.$$

If $a > 0$, we can do this if $b > (1 - 2\sigma)a$, obtaining

$$r_2^2 - r_1^2 = (1 - \sigma)a > 0.$$

If $a < 0$ we can do it if $b > (2 - \sigma)|a|$, obtaining

$$r_1^2 - r_2^2 = (\sigma - 1)a > 0.$$

3. Stability. By stability the author means here what is generally known as *asymptotic orbital stability*. That is, suppose there is a fixed geometrical object in the phase space of the problem, such as a simple closed curve, or a periodic 2-surface, such that if at time zero the representative point of the system lies on the object, then it continues to do so for all time. If there is a neighborhood of this object such that the representative point will, for sufficiently large t, come and remain arbitrarily close to the object, provided it was lying in the neighborhood at $t = 0$, then the system is called stable.

Now system (4) is not well adapted to the study of stability of the periodic solution. In fact, a naïve approach can give misleading results. On the other hand, (6a) is well adapted. System (7) needs no change, and both (6a) and (7) are of form (8).

As stated in §1, there appear to be no published theorems on the stability of system (8). Very recently, S. P. Diliberto has provided tools (see Appendix) with the aid of which it is possible to show at least the following:

If system (7) has a family of solutions of form (C), with $S(\theta, \epsilon)C^1$ in θ, C in θ, ϵ, and if one of the following holds:

(a) $(B(\theta, z_0, 0))^\dagger$ is constant, or

(b) $(B(\theta, z_0, 0))^\dagger$ is diagonal, or

(c) the eigenvalues of $(B(\theta, z_0, 0))^\dagger$ are distinct for all θ, and $(\Phi(\theta, z_0, 0))^\dagger \equiv 0$; then

(i) if all the eigenvalues of $(B(\theta, z_0, 0))^\dagger$ have negative real parts, there is an $\epsilon_2 > 0$ such that for $0 < \epsilon < \epsilon_2$ the periodic surface is stable;

(ii) if at least one of the eigenvalues of $(B(\theta, z_0, 0))^\dagger$ has a positive real part, there is an $\epsilon_3 > 0$ such that for $0 < \epsilon < \epsilon_3$ the periodic surface is unstable.

Modifications of these statements to cover negative ϵ are obvious, and case (a) covers the last sentence of [2].

The application of these results of (6a) or (7) is immediate. In our example it is easily seen that if any of the geometric objects (singular point, periodic solution, periodic surface) is to exist and be stable, we *must* have $H > 0$. Assuming this, and again choosing the indexing so that $\sigma < 1$, and excluding $\sigma > 0$, $\frac{1}{2}$, -1, 2, and for definiteness taking $T_1 > 0$, we find that

(a) the singular point is stable if

$$\frac{-H\nu_1}{\omega T_1} < \delta_0^2 - 2\gamma^2 < \frac{H\nu_1 \sigma}{\omega T_1}$$

and unstable if $\delta_0^2 - 2\gamma^2$ lies outside the indicated closed interval;

(b) the periodic solution of frequency $\nu_1 + O(|\epsilon|)$, which exists for $\delta_0^2 - 2\gamma^2 > -H\nu_1/\omega T_1$, is always unstable;

(c) the periodic solution of frequency $\nu_2 + O(|\epsilon|)$ existing for $\delta_0^2 - 2\gamma^2 > -H\nu_1\sigma/\omega T_1$ is stable if $\delta_0^2 - 2\gamma^2 < H\nu_1(1 - 2\sigma)/\omega T_1$ and unstable for $\delta_0^2 - 2\gamma^2$ above this limit;

(d) the mixed mode, which exists when $\delta_0^2 - 2\gamma^2 < H\nu_1(1 - 2\sigma)/\omega T_1$ is stable there, and $r_2^2 > r_1^2$.

If we had chosen $T_1 < 0$, we would have found an analogous set of circumstances with ν_1 interchanged with ν_2, r_1 with r_2 and σ replaced by $\sigma' = 1/\sigma$.

Our results in this example may be directly compared with one of the cases worked out in detail by C. H. Murphy [5], namely, that of a cubic Magnus moment. His λ_{10} and λ_{20} correspond directly to our $-(H\nu_1 + \omega T\delta_0^2)/\nu_1(1 - \sigma)$ and $-(H\nu_1\sigma + \omega T_1\delta_0^2)/\nu_1(\sigma - 1)$, and his λ_{12} and λ_{22} to $\omega T_1(r_1^2 + 2r_2^2)/\nu_1(1 - \sigma)$ and $\omega T_1(2r_1^2 + r_2^2)/\nu_1(\sigma - 1)$. The only real difference is that we add $2\gamma^2$ to both the different values of "effective mean square yaw": $r_1^2 + 2r_2^2$ and $2r_1^2 + r_2^2$. We have, however, rigorously established that the "reasonable" existence and stability criteria are in fact valid, at least for small ϵ.

4. Comment. In the simple example, $H = \text{const.}$, $m = 0$, $T = T_1(|z|^2 - \delta_0^2)$, the value 2 or $\sigma = \nu_2/\nu_1$ is exceptional. The difficulty is real and does not stem just from limitations of method. There is a true resonance condition, and the periodic solution really does not exist unless we assume $\gamma = 0$ or at least γ is itself small of order ϵ. Also, for $\sigma = 2$ or $\frac{1}{2}$, a periodic 2-surface, if it exists at all, will certainly not be close to that which would be predicted by blindly applying the tests which work for other values of σ. It is easy to show that 0, ± 1, $\frac{1}{2}$, 2 are the only exceptional values of σ as long as H, m and T are linear polynomials in $|z|^2$, $(|z|^2)'$ and $|z|^2$. For more general functions, all rational values of σ are suspect. Moreover, there can be a set of measure zero of irrational values of σ at which the surface will exist and be continuous but not differentiable in ϵ. However, if these three functions should behave very nearly like nth degree polynomials in the three variables, when considered over the range of r_1, r_2, σ which we may really expect to encounter (γ being regarded as fixed and known), then residual variation of these functions may usefully be regarded as of order ϵ, e.g., $T = T_n + \epsilon T_r$, T_n an nth degree polynomial. In such a case there will again be only a finite number of possible reasonance values of σ, and the possible difficulties at a small set of irrational σ values disappear.

Appendix. The following discussion is, except for the part concerning the Q matrix, due in all its essentials to S. P. Diliberto.

The notation $x = \Delta(y)$ will signify

$$\lim_{y \to 0} x = 0.$$

Consider the system

$$\dot{\theta} = \omega + \epsilon[\Theta_0^{(\theta)} + \Theta_1(\theta, z, \epsilon)],$$

$$\dot{z} = \epsilon[A(\theta) + B(\theta)z + C(\theta, \epsilon, z)],$$

where θ is a k-vector and z an l-vector, where all functions on the right-hand side are real and C^1 in (θ, z) and C in (θ, z, ϵ), and are periodic in θ; $\Theta_1 = O(|\epsilon| + |z|)$, $C = O(|\epsilon| + |z|^2)$.

Suppose this possesses the periodic surface $z = S(\theta, \epsilon)$ with $S = \Delta(\epsilon)$.

Then

$$S_\theta\{\omega + \epsilon[\Theta_0(\theta) + \Theta_1(\theta, S, \epsilon)]\} = \epsilon[A(\theta) + B(\theta)S + C(\theta, S, \epsilon)].$$

Set $z = S(\theta, \epsilon) + x$; then

$$\dot{\theta} = \omega + \epsilon[\Theta_0(\theta) + \Theta_2(\theta, x, \epsilon)],$$

$$\dot{x} = \epsilon[B(\theta) + C_1(\theta, x, \epsilon)]x.$$

Θ_2 and C_1 are $O(|x| + \Delta(\epsilon))$.

Let

$$\Theta_0 = \Theta_{0M} + R_{1M},$$
$$B(\theta) = B_M + R_{2M},$$

where Θ_{0M} and B_M are trigonometric polynomials, and $R_{jM} = \Delta(1/M)$. Let

$$\frac{\partial \Psi(\varphi)}{\partial \varphi}\omega = \Theta_{0M}(\varphi) - \Theta^{\dagger}_{0M}(\varphi),$$

with $\Psi(\varphi)$ periodic;

$$\frac{\partial D(\varphi)}{\partial \varphi}\omega = B_M - B_M{}^{\dagger},$$

with D periodic. D exists, since

$$\int_0^T (B_M(\theta - \omega\tau) - B_M{}^{\dagger}(\theta + \omega\tau))\, d\tau$$

is bounded (see [6] and [7]), and similarly for Ψ:

$$\dot{\varphi} = [I + \epsilon\Psi_{\varphi}]^{-1}[\omega + \epsilon(\Theta_0(\varphi + \epsilon\Psi) + \Theta_2(\varphi + \epsilon\Psi, x, \epsilon))],$$
$$\dot{x} = \epsilon[B(\varphi + \epsilon\Psi) + C_1(\varphi + \epsilon\Psi, x, \epsilon)]x,$$
$$\dot{\varphi} = \omega + \epsilon[\Theta_0(\varphi) + \Theta_3(\varphi, x, \epsilon)],$$
$$\dot{x} = \epsilon[B(\varphi) + C_2(\varphi, x, \epsilon) + \Delta(1/M)]x,$$

where Θ_3 and C_2 are $O(|x| + \Delta(\epsilon))$. Let $x = (I + \epsilon D)y$. Then

$$\dot{y} = \epsilon[(B(\varphi))^{\dagger} + C_3(\varphi, y, \epsilon) + \Delta(1/M)]y.$$

We shall show below that in certain cases there exists a real C^1 periodic matrix $T(\varphi)$ such that the transformation $y = Tw$ brings the last equation into

$$\dot{w} = \epsilon[B_3(\varphi) + C_4(\varphi, w) + \gamma H + \Delta(\epsilon) + \Delta(1/M)]w, \quad 0 < \gamma \leqq 1.$$

Here $B_3(\varphi)$ is diagonal or is built of diagonal blocks

$$\begin{pmatrix} \alpha_j & \nu_j \\ -\nu_j & \alpha_j \end{pmatrix}$$

and has the same eigenvalues as $(B(\varphi))^{\dagger}$ and the form

$$B_3 = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix},$$

such that with

$$w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix},$$

$w_1'B_1w > \lambda w_1'w_1$, $\lambda > 0$, $w_2'B_2w \leqq 0$, $C_4 = O(|w|)$, and $\|H\|$ is bounded. The terms $\Delta(1/M)$ may depend on γ, and the terms C_4 and $\Delta(\epsilon)$ on γ and M. H is constant.

Now suppose that all eigenvalues of $(B(\varphi))^\dagger$ have negative real parts for φ. Then $B_3 = B_2$, $w'B_3w < -\mu w'w$ with $\mu > 0$, and

$$\frac{d|w|^2}{dt} = 2w'\epsilon\left[B_3 + \gamma H + C_4 + \Delta\left(\frac{1}{M}\right) + \Delta(\epsilon)\right]w.$$

Choose γ so that $\gamma\|H\| < \mu/6$; then M so that $|\Delta(1/M)| < \mu/6$; then ϵ_0 so that $|\Delta(\epsilon)| < \mu/6$ if $|\epsilon| < \epsilon_0$ and r_0 so that $|C_4| < \mu/6$ if $|w| < r_0$. Then for $|w| < r_0$ and $0 < \epsilon < \epsilon_0$,

$$\frac{d|w|^2}{dt} < -\frac{2\epsilon\mu}{3}|w|^2.$$

The system is stable.

On the other hand, suppose that at least one eigenvalue of $(B(\varphi))^\dagger$ is positive for all φ. Then choosing γ, M, ϵ_0, r_0 as before, but with $\mu/6$ replaced by $\lambda/16$, an argument modeled on that in [8, pp. 317, 318] will show that the system is unstable.

First note that unless $\omega = 0$, $(B(\varphi))^\dagger$ depends on at most $k-1$ independent variables. If, for example, $\omega_1 \neq 0$, then the θ_j appear, if at all, in the combinations $\theta_j - \omega_j\theta_1/\omega_1$, for $1 < j \leqq k$. Call this set of independent variables ψ_j, $1 \leqq j \leqq k^1$. (If $\omega = 0$, we may have $k^1 = k$.) We can now list some important cases in which one can find an appropriate $T(\varphi)$:

(a) $(B(\varphi))^\dagger$ is diagonal. This is obvious.

(b) $(B(\varphi))^\dagger$ is constant. For then there is a constant T_1 such that $B_4 = T^{-1}B^\dagger T_1$ is in Jordan normal form. If B_4 is not diagonal, we can, using a well-known trick, construct a matrix $P(\gamma)$ such that with $T_2 = T_1P$, $T_2^{-1}B^\dagger/T_2 = B_5 + \gamma H$, B_5 diagonal, and the only nonzero elements of H are 1's on the superdiagonal. If $0 < \gamma \leqq 1$, and $y = Tw$, there is an r, $0 \leqq r$, $\gamma^r|w| \leqq |y| \leqq |w|$. If the eigenvalues of $B^\dagger$ are all real, set $T = T_2$. If any are complex, T_2 will be also, but there will be a constant matrix T_3, an obvious generalization of

$$\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix},$$

such that $T = T_2T_3$ is real and $T^{-1}B^\dagger T = B_6 + \gamma H_1$, B_6 being made up of

diagonal blocks, either single elements or elements of form

$$\begin{pmatrix} \alpha_j & \nu_j \\ -\nu_j & \alpha_j \end{pmatrix}.$$

(c) The eigenvalues of $(B(\varphi))^\dagger$ are distinct for all φ, their real parts change sign nowhere, and $(\Theta_0(\varphi))^\dagger \equiv 0$.

It will be shown below that there is a matrix $Q(\varphi)$, C^1 in φ, nonsingular, and periodic in φ (though the periods may be double those of $(B(\varphi))^\dagger$) such that $B_7 = Q^{-1}B^\dagger Q$ is diagonal. Furthermore, $Q_\varphi \omega = 0$. Then with $y = Qu$,

$$\dot{u} + Q^{-1}Q_\varphi[\omega + \epsilon_0(\Theta_0^\dagger + \Theta_3)]u = \epsilon[B_7 + Q^{-1}[C_3 + \Delta(1/M)]Q]u,$$

$$\dot{u} = \epsilon[B_7 + C_5(u) + \Delta(\epsilon) + \Delta(1/M)]u,$$

$$C_5 = O(|w|).$$

The remarks above on the reality of T_2 apply also to the reality of Q. Finally, set $T = Q$. Let $(B(\varphi))^\dagger = \beta(\psi)$, $\psi = (\psi_1, \cdots, \psi_{k^1})$, and consider the ψ_j rescaled so that the period of β is unity in each ψ_j. In the ψ-space consider the sets of closed intervals $\{r_j^N\}$, where each r_j^N is of the form $n_{j,i} \leqq 2^N\psi_i \leqq n_{j,i} + 1$, $1 \leqq i \leqq k^1$ and $0 \leqq n_{j,i} \leqq 2^N - 1$. The index j is taken to be that of the particular enumeration of the 2^{Nk^1} intervals such that $j^1 > j$ if $n_{j^1,1} > n_{j,1}$, and if $n_{j^1,m} = n_{j,m}$ for $1 \leqq m \leqq k^1$, $n_{j^1,m+1} > n_{j,m+1}$. For any N the union of the $\{r_j^N\}$ covers the fundamental interval $0 \leqq \psi_j \leqq 1$. Now the distinctness of the eigenvalues of β implies first that each of the eigenvalues is everywhere distinguishable, that if one of them is complex, then it is so everywhere, and that for each eigenvalue λ_n, there is at each point of ψ-space at least one principal subdeterminant of $\beta - \lambda_n I$ which is not zero. Furthermore, λ_n is C^1. An argument similar to that of a well-known proof of the Heine-Borel theorem shows that there exists an N such that for each r_j^N there is at least one of these subdeterminants, say, Δ_{kj}, which vanishes nowhere on r_j^N, and hence nowhere on some open interval including r_j^N. On this open interval we can construct an eigenvector $\xi_{r,j}$ of β, belonging to λ_n, by taking Δ_{kj} as its kjth component and solving for the other components. If λ_n is real, so will $\xi_{n,j}$ be. If λ_n is complex, then $\xi_{n,j}^*$ will be an eigenvector belonging to $\lambda_n{}^*$. We normalize $\xi_{n,j}$ by dividing each component $\xi_{n,j,l}$ by

$$\{\sum_l |\xi_{n,j,l}|^2\}^{1/2},$$

taking the positive root. Call the resulting vector $\alpha_{n,j}$. It is C^1 on our open interval. At each point a normalized eigenvector belonging to λ_n is unique to within a scalar factor of unit magnitude. Thus, if r_j^N and $r_{j^1}^N$, $j^1 \neq j$,

have points in common, there is an open set including their intersection on which $\alpha_{n,j} = \alpha_{n,j}e^{i\gamma(\psi)}$, where γ is real and C^1. If λ_n is real, so are $\alpha_{n,j}$ and α_{n,j^1}, so that $e^{i\gamma(\psi)}$ can be only ± 1, and must be the same on the whole of the overlap of regions of definition. It is now clear that we can in this case define the eigenvector η_n to be C^1 over the whole fundamental interval by setting $\eta_n = \alpha_{n,1}$ on $r_1{}^N$, and then extending the definition to other $r_j{}^N$ by steps at each of which j increases by one. Then, for any j, on the side $\psi_j = 1$ of the fundamental interval, η_n will be the same function of the other ψ's as on the side $\psi_j = 0$, apart from a possible constant factor of -1.

The case of complex λ_n is a little more subtle. But it can be shown by actual construction that if a function $g(x)$, $x = (x_1, \cdots, x_m)$, is defined and C^1 on an open set including the union of one or more S_j, where S_j is the intersection of the manifold $x_j = 0$ with the interval $0 \leqq x_k \leqq 1$, $1 \leqq k \leqq m$, then there is a function $f(x)$ which is C^1 on an open interval including the unit closed one, and which is equal to and has the same first partial derivatives as $g(x)$ on the union of the S_j in question. Again we can define $\eta_n = \alpha_{n,1}$ on $r_1{}^N$, extend the definition to $r_2{}^N$ by multiplying $\alpha_{n,2}$ by $e^{i\gamma(\psi)}$ with $\gamma(\psi)$ an appropriate C^1 function, and continue in the same fashion increasing j by one at each step, obtaining η_n, C^1 over the whole fundamental interval, and, in fact, on an open set including it. If as may happen, we do not have $\eta_n(1, \psi, \cdots, \psi_{k^1}) = \eta_n(0, \psi_2, \cdots, \psi_{k^1})$ and $\eta_{n\psi_1}(1, \psi_2, \cdots, \psi_{k^1}) = \eta_{n\psi_1}(0, \psi_2, \cdots, \psi_1)$, we still have $\eta_n(\psi_1, \psi_2, \cdots, \psi_{k^1}) = \eta_n(\psi_1 - 1, \psi_2, \cdots, \psi_{k^1})e^{i\gamma_1(\psi_1\cdots\psi_{k^1})}$ near $\psi_1 = 1$. Then, if we take $\delta_1 = -\psi_1{}^2(a + b\psi_1)$, with

$$a = 3\gamma_1(1, \psi_2, \cdots, \psi_{k^1}) - \gamma_{1\psi_1}(1, \psi_2, \cdots, \psi_{k^1}),$$

$$b = -2\gamma_1(1, \psi_2, \cdots, \psi_{k^1}) + \gamma_{1\psi_1}(1, \psi_2, \cdots, \psi_{k^1}),$$

and $\mu_n = \eta_n e^{i\delta_1}$,

$$\mu_n(1, \psi_2, \cdots, \psi_{k^1}) = \mu_n(0, \psi_2, \cdots, \psi_{k^1}),$$

$$\mu_{n\psi_1}(1, \psi_2, \cdots, \psi_{k^1}) = \mu_{n\psi_1}(0, \psi_2, \cdots, \psi_{k^1}).$$

We can deal with nonagreement on other pairs of sides similarly, and then rename the eigenvector η_n.

In either the real or complex case we can now extend η_n to be C^1 in all of ψ-space and to be periodic, though perhaps of period 2. A matrix made up from the l eigenvectors η_n of β will have the properties desired of Q.

REFERENCES

[1] W. R. Haseltine, *Existence theorems for nonlinear ballistics*, this Journal, 11 (1963), pp. 553-563.

[2] ———, *Existence theorems for nonlinear ballistics, II: Effect of gravity*, this Journal, 11 (1963), pp. 1071–1077.

[3] S. P. DILIBERTO, *Perturbation theorems for periodic surfaces*, Rend. Circ. Math. Palermo, 1 (1960), pp. 2–9.

[4] S. P. DILIBERTO AND G. HUFFORD, *Perturbation theorems for nonlinear ordinary differential equations*, Ann. of Math. Studies no. 3, S. Lefschetz, ed., Princeton University Press, Princeton, 1956, pp. 207–236.

[5] C. H. MURPHY, *Prediction of the motion of missiles acted on by nonlinear forces and moments*, Aero-Space Sci., 24 (1957), pp. 473–479.

[6] S. P. DILIBERTO, *Perturbation theory of periodic surfaces, II*, Tech. Rep. 9, Contract Nonr-222(37), University of California, Berkeley, 1957.

[7] ———, *Perturbation theory of periodic surfaces, III*, Tech. Rep. 10, Contract Nonr-222(37), University of California, Berkeley, 1967.

[8] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

THE OPTIMAL CONTROL OF A TRANSVERSE VIBRATION OF A BEAM*

V. KOMKOV†

Summary. This paper deals with the specific problem of a transversely vibrating beam, whose ends are either free, freely supported, or built in. Sections 1–4 state the basic equations, and define the optimum control problems. In §5 we prove the existence and a form of uniqueness of an optimal control. Sections 6 and 7 state Pontryagin's principle and give a possible example of its application. Section 8 introduces a limiting process for the fixed time interval optimal controls, which results in an optimal control, called the instantly optimal control. It is shown that this control is unique and independent of the limiting process used in deriving it. Moreover, Pontryagin's principle for the instantly optimal control makes use of the displacement vector which could be computed without the knowledge of the finite state. Finally we offer a brief sketch of an argument showing that the instantly optimal controls are different in general from either the fixed time interval optimal controls, or from the time optimal controls.

Introductory remarks. This paper deals directly with the problem of optimal control of a vibrating beam and derives the basic results concerning the existence, uniqueness, and the appropriate form of Pontryagin's principle directly from the differential equation governing the behavior of the beam, together with the physically motivated boundary and initial conditions. The technique of proof parallels the classical arguments of Pontryagin, Boltyanskii, Gamkrelidze and Mishchenko [6], further developed in [2] and [9]. The theoretical development given by Russell in [7] was inapplicable, since his arguments were developed for hyperbolic systems. There is, of course, a close parallel between our results and the discussion by Russell of the optimal control of a vibrating string. In a larger framework our control problem could be regarded as the problem of the calculus of variations, where we wish to minimize the functional equal to the total energy of the beam, whose motion is subject to the constraints imposed by the initial condition, and by the abovestated boundary conditions.

This point of view has been advanced by Gamkrelidze and by Cesari. The work of Cesari [12] develops the important existence theorems for the Lagrange problems with constraints, when only weak solutions are assumed (which is exactly the case in our problem). However, the specialized results of this paper do not turn out to be straightforward applications of the existing general theory at the present time.

* Received by the editors October 26, 1967, and in revised form February 28, 1968.

† Department of Mathematics, The Florida State University, Tallahassee, Florida 32306. This research was supported by the National Science Foundation under Grant GP 7457.

1. The basic equation and the related hypothesis. Under the usual simplifying assumptions the equation of transverse vibration of a beam is

$$L(w) = \rho(x)\cdot A(x)\cdot \frac{\partial^2 w(x,t)}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left(E(x)\cdot I(x)\cdot \frac{\partial^2 w(x,t)}{\partial x^2}\right) = f(x,t), \tag{1}$$

$$-l/2 \leqq x \leqq +l/2, \quad t \geqq 0.$$

The corresponding homogeneous equation is

$$L(w) = 0. \tag{1a}$$

(See, for example, [3, Chap. 14, pp. 218–220] for the discussion of the hypothesis of the "simple theory" of bending of beams.)

The physical meanings of the symbols are given below.

$\rho(x)$: the material density;
$A(x)$: the cross-sectional area;
$w(x, t)$: the transverse displacement;
$E(x)$: Young's modulus;
$I(x)$: the moment of inertia of the cross-sectional area about the neutral axis;
$f(x, t)$: the applied load.

Because of the physical meaning of (1) we must assume that the displacement $w(x, t)$ is a continuously differentiable function of x and t. On the interval $[-l/2, +l/2]$, ρ, A, E, I are uniformly bounded, piecewise smooth, positive functions of x. An additional assumption of material homogeneity would imply that ρ and E are constant. ($A(x)$, $I(x)$ may still vary along the length of the beam.) We assume the correctness of Hooke's law and equate the strain energy with the complementary energy of the beam.

In the remainder of this paper we shall consider only the class of weak solutions of (1) obeying the following conditions:

(a) $w(x, t)$ is a continuously differentiable function of x and of t on $\Omega = [-l/2, +l/2] \times [0, T]$ (i.e., $\partial w(x,t)/\partial t$ and $\partial w(x,t)/\partial x$ are continuous functions of x and t on Ω);

(b) $\sqrt{\rho(x)\,A(x)}\ (\partial w(x,t)/\partial t)$, and $\sqrt{E(x)I(x)}\ (\partial^2 w(x,t)/\partial x^2)$ are square integrable functions of x on $[-l/2, +l/2]$, and the energy functions

$$K(t) = \frac{1}{2}\int_{-l/2}^{+l/2} \rho(x)A(x)\left[\frac{\partial w(x,t)}{\partial t}\right]^2 dx, \tag{2}$$

$$V(t) = \frac{1}{2}\int_{-l/2}^{+l/2} E(x)I(x)\left[\frac{\partial^2 w(x,t)}{\partial x^2}\right]^2 dx \tag{3}$$

are continuous functions of t, uniformly bounded on the interval $[0, T]$;

(c) $w(x, t)$ obeys one of the three conditions (4a), (4b) or (4c) listed below at each boundary point $x = -l/2, x = +l/2$:

$$w(\pm\, l/2, t) = 0, \qquad \frac{\partial w(\pm\, l/2, t)}{\partial x} = 0 \tag{4a}$$

(a built-in end);

$$w(\pm\, l/2, t) = 0, \qquad EI\,\frac{\partial^2 w(\pm\, l/2, t)}{\partial x^2} = 0 \tag{4b}$$

(a freely supported end);

$$EI\,\frac{\partial^2 w(\pm\, l/2, t)}{\partial x^2} = 0, \qquad \frac{\partial}{\partial x}\left[EI\,\frac{\partial^2 w(\pm\, l/2, t)}{\partial x^2}\right] = 0 \tag{4c}$$

(a free end);

(d) $w(x, t)$ obeys given initial conditions of the form

$$w(x, 0) = \psi(x), \tag{5a}$$

$$\frac{\partial w(x,0)}{\partial t} = \eta(x), \tag{5b}$$

$$\psi(x), \eta(x) \in C^1[-l/2, +\, l/2].$$

We also assume that the inhomogeneous term $f(x, t)$ satisfies either of the conditions:

(i) $f(x, t)$ is a square integrable (hence absolutely integrable) function of the variable x in the interval $[-l/2, +\, l/2]$, and $\int_{-l/2}^{+l/2} |f(x, t)|\, dx$ is a measurable and uniformly bounded function of t in the interval $[0, T]$. Also $f(x, t)$ is assumed to obey the inequality

$$\|f(x, t)\|_{(x)}^2 = \left[\int_{-l/2}^{+l/2} |f(x, t)|\, dx\right]^2 \leqq 1. \tag{6}$$

Note. There is no additional generality in assuming that $\|f(x, t)\|_{(x)}^2 \leqq C$ for some $C > 0$.

These conditions imply that the total energy of the beam $\mathcal{E}(t) = K(t) + V(t)$ is also uniformly bounded for all $t \in [0, T]$, where $K(t)$ and $V(t)$ are defined by (2) and (3) respectively.

The case (i) will be called the case of a distributed load, and the control function $f(x, t)$ satisfying the condition (i) will be called a distributed load control.

(ii) $f(x, t)$ is assumed to be of the form

$$f(x, t) = \sum_{i=1}^{N} \delta(x - \xi_i(t))\phi_i(t) + \sigma(x, t), \tag{7}$$

where $\sigma(x, t)$ is a distributed load control; $\delta(x - \xi_i(t))$ is the "shifted" Dirac delta function, regarded as a generalized function (see [4, pp. 3, 4]). The functions $\phi_i(t)$ are measurable functions of the variable t, obeying the condition

$$\sum_{i=1}^{N} |\phi_i(t)| + \int_{-l/2}^{+l/2} |\sigma(x, t)|\, dx \leqq 1 \tag{7a}$$

for all $t \in [0, T]$.

The total energy of the beam is uniformly bounded above by some constant. (We could, of course, incorporate point couples in expression (7a):

$$f(x, t) = \sum_{i=1}^{N} \delta(x - \xi_i(t))\phi_i(t) + \sum^{M} \delta'(x - \zeta_i(t))\mu_i(t) + \sigma(x, t), \tag{7b}$$

where $'$ denotes $\partial/\partial x$.)

For the sake of simplicity no point couple controls will be used in this paper, except as possible limits of sequences considered in the last paragraph.

The functions $\xi_i(t)$ are measurable functions whose domain includes the interval $[0, T]$ and whose range lies in the interval $[-l/2, +l/2]$. In case (ii) we shall denote by $\| f(x, t) \|_{(x)}$ the quantity

$$\| f(x, t)\|_{(x)} = \sum |\phi_i(t)| + \int_{-l/2}^{+l/2} |\sigma(x, t)|\, dx.$$

As before we assume $\| f(x, t) \|_{(x)} \leqq 1$. The first term on the left-hand side of (7a) will be called the point load controls. A generalized function $f(x, t)$ obeying either of the conditions (i) or (ii) will be called an admissible control.

2. Remarks. The control functions are regarded as generalized functions over the space of test functions satisfying the conditions (a), (b), (c) and (d). It is clear that the products $(\delta(x), w(x, t))$ $(\delta(x), \partial w(x, t)/\partial t)$ are defined for any test function satisfying the condition (a).

The problem of existence and uniqueness of solutions of the mixed boundary and initial value problem (MBVP) posed by (1) with conditions (4)

and (5) will not be considered in this paper. In the case when the control function $f(x, t)$ is both an absolutely integrable and a square integrable function of both x and t, proofs can be found in the literature. In the more usual case (ii) the author has not been able to find a published proof. However, it is easy to check that the classical proofs can be extended to cover the case when $f(x, t)$ obeys (7) by the use of suitable delta-convergent sequences (see [4, vol. 1, §2.5] for explanation of the procedure). The existence and uniqueness of solutions of the MBVP will be assumed in the subsequent discussion. The important problem of controllability will also be neglected in this paper.

For purposes of convenience we shall denote the usual bilinear products $(\phi(x), f(x))$ arising in the generalized function theory by the symbol $\int \phi(x)\, f(x)\, dx$ even though $\phi(x)$ may be a generalized function which is *not* locally integrable. (This follows the usual practice in physics and engineering, and avoids the inconvenience of double notation, and of discussion of separate cases.) The solution of the inhomogeneous equation (1) subject to conditions (4) and (5), with a control function $\phi(x, t)$ obeying the conditions (i), is known to obey Duhamel's principle

$$w(x, t) = w_H(x, t) + \int_0^t \int_{-l/2}^{+l/2} G(x, \xi, t, \tau) \cdot \phi(\xi, \tau)\, d\xi\, d\tau, \tag{8}$$

where $w_H(x, t)$ is the solution of the homogeneous equation, while $G(x, \xi, t, \tau)$ depends only on the coefficients ρ, A, E and I, and on the boundary conditions (4), but does not depend on either $\phi(x, t)$ or on the initial value functions $\psi(x)$ and $\eta(x)$.

Again an elementary argument concerning delta-convergent sequences shows that this statement may be extended to cover the case (ii). We observe that the admissible controls form a convex set, i.e., if $f_1(x, t)$ and $f_2(x, t)$ are admissible controls, then $\Lambda f_1 + (1 - \Lambda) f_2$ is also an admissible control for any $0 \leqq \Lambda \leqq 1$.

3. The energy terms. The kinetic energy of the beam is given by

$$K = \frac{1}{2} \int_{-l/2}^{+l/2} \rho(x) A(x) \left(\frac{\partial w(x, t)}{\partial t} \right)^2 dx, \tag{9}$$

and the strain energy by

$$V = \frac{1}{2} \int_{-l/2}^{+l/2} E(x) I(x) \left(\frac{\partial^2 w(x, t)}{\partial x^2} \right)^2 dx. \tag{10}$$

The total energy $\mathcal{E}(t)$ is the sum of the kinetic energy and the strain energy

$$\mathcal{E} = K + V = \frac{1}{2} \int_{-l/2}^{+l/2} \rho(x) A(x) \left[\frac{\partial w(x, t)}{\partial t} \right]^2 + E(x) I(x) \left[\frac{\partial^2 w}{\partial x^2} \right]^2 dx. \tag{11}$$

The physical interpretation of (9) and (10) implies that $\partial w/\partial t$ and $\partial^2 w/\partial x^2$ have to be square integrable on the interval $[-l/2, +l/2]$. We also assume that they are square integrable on $[0, T]$. We introduce the following product of two functions $u(x, t)$, $v(x, t)$, whose derivatives $\partial u/\partial t$, $\partial^2 u/\partial x^2$, $\partial v/\partial t$, $\partial^2 v/\partial x^2$ are square integrable functions in the interval $-l/2 \leqq x \leqq +l/2$:

$$\langle u, v\rangle = \frac{1}{2}\int_{-l/2}^{+l/2}\left[\rho(x)\cdot A(x)\,\frac{\partial u}{\partial t}\,\frac{\partial v}{\partial t} + E(x)I(x)\,\frac{\partial^2 u}{\partial x^2}\,\frac{\partial^2 v}{\partial x^2}\right]dx. \tag{12}$$

$\langle u, v\rangle$ is clearly a function of t only. If $u = v$, then $\langle u, v\rangle$ is the total energy, as defined by the formulas (9), (10) and (11). A property of this product proved in Lemma 1 below will be of importance in the subsequent development of Pontryagin's principle.

LEMMA 1. *Let $u(x, t)$, $v(x, t)$ be two solutions of the MBVP with corresponding controls $f(x, t)$, $g(x, t)$. Then*

$$\frac{d}{dt}\langle u, v\rangle = \frac{1}{2}\int_{-l/2}^{+l/2}\left[f(x, t)\,\frac{\partial v}{\partial t} + g(x, t)\,\frac{\partial u}{\partial t}\right]dx. \tag{13}$$

Before we prove this lemma, we emphasize that $u(x, t)$ is the solution corresponding to the control $f(x, t)$, and $v(x, t)$ corresponds to $g(x, t)$. $f(x, t)$ and $g(x, t)$ could be integrable and square integrable on $[-l/2, +l/2]$, or they could be Dirac delta functions. Despite the fact that the Dirac delta function is *not* a locally integrable function, we shall retain the commonly accepted use of the integral sign and interpret the resulting product as the usual linear map (see, for example, [4, pp. 1–4]). No other changes will be necessary.

Proof of Lemma 1. We shall make use of any of the formulas (4a), (4b) or (4c) in integrating by parts. Thus

$$\begin{aligned}\frac{d}{dt}\langle u, v\rangle &= \frac{1}{2}\int_{-l/2}^{+l/2}\left\{\rho(x)A(x)\left[\frac{\partial^2 u}{\partial t^2}\,\frac{\partial v}{\partial t} + \frac{\partial u}{\partial t}\,\frac{\partial^2 v}{\partial t^2}\right.\right.\\ &\qquad\left.\left. + \frac{\partial}{\partial t}\left[E(x)I(x)\,\frac{\partial^2 u}{\partial x^2}\,\frac{\partial^2 v}{\partial x^2}\right]\right]\right\}dx\\ &= \frac{1}{2}\int_{-l/2}^{+l/2}\left\{\left[f(x, t) - \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\,\frac{\partial^2 u}{\partial x^2}\right)\right]\cdot\frac{\partial v}{\partial t}\right.\\ &\qquad + \left[g(x, t) - \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\,\frac{\partial^2 v}{\partial x^2}\right)\right]\cdot\frac{\partial u}{\partial t}\\ &\qquad + \frac{\partial v}{\partial t}\cdot\left[\frac{\partial^2}{\partial x^2}\left(E(x)I(x)\,\frac{\partial^2 u}{\partial x^2}\right)\right] + v\cdot\frac{\partial}{\partial t}\left[\frac{\partial^2}{\partial x^2}\left(E(x)I(x)\,\frac{\partial^2 u}{\partial x}\right)\right]\Bigg\}dx\\ &= \frac{1}{2}\int_{-l/2}^{+l/2}\left[f(x, t)\,\frac{\partial v}{\partial t} + g(x, t)\,\frac{\partial u}{\partial t}\right]dx\end{aligned}$$

after integration by parts and interchange of the order of differentiation.

COROLLARY 1. *If $v(x, t)$ is the solution of the homogeneous equation* (1a), *then*

$$\langle u, v\rangle_{t=\tau} = \langle u, v\rangle_{t=0} + \int_0^\tau \int_{-l/2}^{+l/2} \left[f(x, t) \frac{\partial v}{\partial t} \right] dx, \qquad 0 \leqq \tau \leqq T. \tag{14}$$

Proof. Since $v(x, t)$ is the solution of a homogeneous equation, $g(x, t) \equiv 0$, and

$$\langle u, v\rangle_{t=\tau} - \langle u, v\rangle_{t=0} = \int_0^\tau \int_{-l/2}^{+l/2} \left[f(x, t) \frac{\partial v}{\partial t} \right] dx$$

or

$$\langle u, v\rangle_{t=\tau} = \langle u, v\rangle_{t=0} + \int_0^\tau \int_{-l/2}^{+l/2} \left[f(x, t) \frac{\partial v}{\partial t} \right] dx$$

as required.

COROLLARY 2. *If $f(x, t) \equiv g(x, t) \equiv 0$, then $\langle u, v\rangle \equiv$ const.*

We observe that in the particular case when $u(x, t) = v(x, t)$ and $f(x, t) = g(x, t) \equiv 0$, this corollary reduces to the trivial statement that the total energy is conserved if the beam vibrates freely.

LEMMA 2 (The Cauchy-Schwarz inequality).

$$\langle u, v\rangle^2_{t=\text{const.}=\tau} \leqq \langle u, u\rangle \cdot \langle v, v\rangle \,|_{t=\tau},$$

and equality holds only if $\partial u/\partial t = c\partial v/\partial t$, $\partial^2 u/\partial x^2 = c\partial^2 v/\partial x^2$ for some constant c.

Proof. It is sufficient to observe that $\langle u, v\rangle$ does satisfy all requirements of a scalar product for the "vectors" $[\partial u/\partial t, \partial^2 u/\partial x^2]$ $[\partial v/\partial t, \partial^2 v/\partial x^2]$, and that $\langle u, u\rangle$ is a norm (see, for example, [1, p. 5] for a classical proof).

4. Statement of the control problems. Given the initial conditions (5a) and (5b) and one of the boundary conditions (4a), (4b) or (4c) at each boundary point $x = -l/2, x = +l/2$ and given $T > 0$, find an admissible control $\tilde{\phi}(x, t)$ such that the total energy of the beam obeys the inequality

$$\mathcal{E}(\tilde{\phi}(x, t), t = T) \leqq \mathcal{E}(\phi(x, t), t = T), \tag{15}$$

where $\phi(x, t)$ is any other admissible control. The control $\tilde{\phi}(x, t)$ will be called an optimal control for the interval $[0, T]$.

The control problem stated above will be called the fixed interval control problem. Closely related to it is the minimal time control problem. Given the same initial and boundary conditions and given a nonnegative number E, such that $E < \mathcal{E}(t = 0)$, find the control $\hat{\phi}(x, t)$ which reduces the total energy of the beam to the value E in the shortest possible time.

5. The existence and uniqueness theorems for the fixed interval control problem.

THEOREM 1 (Existence of an optimal control). *Let the MBVP be posed*

for an interval $[0, T]$. *Then there exists at least one admissible optimal control* $\tilde{\phi}(x, t)$ *for the* $[0, T]$ *fixed interval control problem.*

Proof. Let E be the greatest lower bound on the total energy attainable at the time T through the use of admissible controls. (Clearly such a number exists.) Since $\mathcal{E}(T)$ depends continuously on the function $F(t) = \int_{-l/2}^{+l/2} \phi(x, t)\, dx$, where $\phi(x, t)$ is the control, we can choose a sequence of admissible controls $\phi_i(x, t)$ such that $\lim_{i\to\infty} \mathcal{E}(\phi_i(x, t), t = T) = E$. The generalized functions $\Phi_i(x, t)$ obey the inequality

$$\| \phi_i(x, t)\|_{\Omega}^2 = \int_0^T (\|\phi_i(x, t)\|_x)^2\, dt \leqq 2T.$$

Hence the hypotheses of the lemma in Appendix 1 are satisfied, and we can assert the existence of an admissible control $\phi(x, t) = \lim_{i\to\infty} \phi_i(x, t)$. It is easy to check that $\mathcal{E}(\phi(x, T), T) = E$, using (8) together with (9), (10) and (11).

DEFINITION 1. The set of all functions $w(x, t)$, of the class C^1 in $\Omega = [-l/2, + l/2] \times [0, T]$, which are solutions of the MBVP, for which the inhomogeneous term is an admissible control, will be called *an attainable set of displacements.* The corresponding functions $E(x)I(x)\ (\partial^2 w(x, t)/\partial x^2)$ will be called *the attainable bending moments* and will be denoted by $M(x, t)$.

For convenience we introduce a new notation. We shall denote by $\mathbf{W}(x, t)$ the vector

$$\left[\frac{\partial w(x, t)}{\partial t}\ ,\ \frac{\partial^2 w(x, t)}{\partial x^2}\right].$$

We make the observation that the attainable set of displacements $\mathbf{w}$ forms a convex subset of $L_2\{[- l/2, + l/2] \times [0, T]\}$. This follows immediately from the linear dependence of displacements $\mathbf{w}(x, t)$ upon the controls (see (8)) and from the convexity of the admissible controls.

THEOREM 2 (Uniqueness of the finite state). *Let* $\tilde{\phi}_1(x, t)$, $\tilde{\phi}_2(x, t)$ *be two admissible controls which are optimal controls for the* $[0, T]$ *fixed interval. Then the corresponding displacement and velocity functions coincide at the time* $t = T$, *i.e.*,

$$\begin{aligned} w_1(\tilde{\phi}_1\ , x, T) &= w_2(\tilde{\phi}_2\ , x, T), \\ \frac{\partial w_1}{\partial t}\ (\tilde{\phi}_1\ , x, T) &= \frac{\partial w_2}{\partial t}\ (\tilde{\phi}_2\ , x, T). \end{aligned} \tag{16}$$

Proof. Let us assume to the contrary that there exist two optimal controls $\tilde{\phi}_1(x, t)$ and $\tilde{\phi}_2(x, t)$ such that

$$\mathbf{W}_1(\tilde{\phi}_1\ , x, T) \not\equiv \mathbf{W}_2(\tilde{\phi}_2\ , x, T).$$

By the convexity of the attainable displacement set we conclude that $\frac{1}{2}(\mathbf{W}_1 + \mathbf{W}_2) = \mathbf{W}(x, t)$ is also an attainable displacement. The corresponding total energy at the time T is

$$\begin{aligned}\mathcal{E}(\mathbf{W}(x, T)) &= \frac{1}{2}\left\{\int_{-l/2}^{+l/2} \rho\,(x)A(x)\left[\frac{\partial(w_1 + w_2)}{\partial t}\right]^2 \right. \\ &\quad \left. + \frac{E(x)I(x)}{4}\left[\frac{\partial^2(w_1 + w_2)}{\partial x^2}\right]^2 dx\right\}\Bigg|_{t=T} \\ &= \tfrac{1}{4}\,\mathcal{E}(w_1(x, T)) + \tfrac{1}{4}\,\mathcal{E}(w_2(x, T)) + \tfrac{1}{2}\langle w_1\,,\, w_2\rangle_{t=T} \\ &= \tfrac{1}{2}\,E + \tfrac{1}{2}\,\langle w_1\,,\, w_2\rangle_{t=T}\,,\end{aligned}$$

where E is the energy attainable at the time T by an optimal control.

By Lemma 2,

$$\langle w_1\,,\, w_2\rangle^2 \leqq \langle w_1\,,\, w_1\rangle \cdot \langle w_2\,,\, w_2\rangle = E^2.$$

Since E is the minimum total energy attainable, we must have $\langle w_1\,,\, w_2\rangle^2 = E^2$. (Otherwise $\mathcal{E}(w, T) < E$, which is a contradiction.) But the strict equality

$$\langle w_1\,,\, w_2\rangle^2 = \langle w_1\,,\, w_1\rangle \cdot \langle w_2\,,\, w_2\rangle$$

implies that there exists a constant C such that

$$\frac{\partial w_1}{\partial t} = C\,\frac{\partial w_2}{\partial t}\,, \qquad \frac{\partial^2 w_1}{\partial x^2} = C\,\frac{\partial^2 w_2}{\partial x^2}\,.$$

Since $\mathcal{E}(w_1\,,\, T) = \mathcal{E}(w_2\,,\, T) = E$, upon substitution into (11) we find that $C = 1$, and that

$$\frac{\partial w_1}{\partial t} = \frac{\partial w_2}{\partial t}\Bigg|_{t=T} \quad \text{and} \quad \frac{\partial^2 w_1}{\partial x^2}\Bigg|_{t=T} = \frac{\partial^2 w_2}{\partial x^2}\Bigg|_{t=T}$$

for almost all $x \in [-\,l/2, +\,l/2]$.

From the assumption of piecewise smoothness of $\partial^2 w/\partial x^2$, we conclude that

$$w_2(x, T) = w_1(x, T) + C_1(x) + C_2\,,$$

where C_1 and C_2 are constants. (An identical conclusion would follow a more general hypothesis that $\partial^2 w/\partial x^2$ could be a step function with a finite number of steps in $[-\,l/2, +\,l/2]$, and the formula is clearly valid when $\partial^2 w/\partial x^2$ is a sum of such a step function and of a piecewise smooth function. This problem does not arise with $\partial w/\partial t$ which must be continuous in $[-\,l/2, +\,l/2]$.)

Assuming that either conditions (4b) or (4c) are applicable at one end

of the beam, and either (4a), (4b) or (4c) at the other end, we see that $C_1 = C_2 = 0$. In the remaining case, when conditions (4a) are applied at both ends of the beam, we also arrive at the conclusion that $C_1 = C_2 = 0$. Then in all possible cases $C_1 = C_2 = 0$, and $w_1(x, T) \equiv w_2(x, T)$. This completes the proof.

6. Pontryagin's principle. The principle stated in Theorem 3 below is in complete agreement with the maximum principle of Pontryagin (see [6]), and also with the results of [2] and [7]. The proof parallels the proof of Russell [7] for the symmetric hyperbolic systems with a few important differences.

THEOREM 3. *Let $\tilde{\phi}(x, t)$ be an optimal control for the fixed interval control of the MBVP, as stated in the preceding section. We assume that $\mathcal{E}(\tilde{\phi}(x, t), t) > 0$ if $t \in [0, T]$. Let $\tilde{w}(x, t)$ be the solution of the MBVP corresponding to $\tilde{\phi}(x, t)$. Let $v(x, t)$ be a solution of the homogeneous equation (1) satisfying the same boundary conditions, and such that $\mathbf{v}(x, T) = \tilde{\mathbf{w}}(x, T)$, i.e.*

$$v(x, T) = \tilde{w}(x, T),$$

$$\frac{\partial v(x, T)}{\partial t} = \frac{\partial \tilde{w}(x, T)}{\partial t}.$$

Then

$$(17) \qquad \int_{-l/2}^{+l/2} \left[-\tilde{\phi}(x, t) \frac{\partial v(x, t)}{\partial t} \right] dx = \max \int_{-l/2}^{+l/2} \left[-\phi(x, t) \frac{\partial v(x, t)}{\partial t} \right] dx$$

for all admissible controls $\phi(x, t)$.

Proof. If we can prove the theorem under the additional assumption that all controls including $\tilde{\phi}(x, t)$ are piecewise continuous (piecewise smooth), the general statement will follow as an easy corollary. Let $t = \tau$ be a point of continuity of the optimal control $\tilde{\phi}(x, t)$, $\tau \in (0, T)$. Then there exists $\sigma > 0$, such that $\tilde{\phi}(x, t)$ is continuous in the interval $I_\sigma = [\tau - \sigma, \tau + \sigma]$, and the interval I_σ is contained in $[0, T]$. Let $\psi(x, t)$ be an admissible control such that for a sufficiently small number $\epsilon > 0$, $\tilde{\phi} + \epsilon\psi$ is an admissible control. (Clearly, if no such $\psi(x, t)$ can be found, $\tilde{\phi}(x, t)$ is the unique admissible control in I_σ, and there is nothing to prove.)

Let us consider the control

$$\phi'(x, t) = \begin{cases} \tilde{\phi}(x, t) & \text{for } t \in [0, T] - I_\sigma, \\ \tilde{\phi}(x, t) + \epsilon\psi(x, t) & \text{for } t \in I_\sigma. \end{cases}$$

$\phi'(x, t)$ is clearly an admissible control.

Let ψ_σ be the control:

$$\psi_\sigma(x, t) = \begin{cases} 0 & \text{for } t \notin I_\sigma, \\ \psi(x, t) & \text{for } t \in I_\sigma. \end{cases}$$

σ can be chosen so that ψ_σ is smooth in I_σ. (All controls are piecewise smooth functions of time.)

Let $w(\phi, x, t)$ be the solution of the MBVP corresponding to the control $\phi(x, t)$, and $\tilde{w}_\sigma(\psi_\sigma, x, t)$ be the solution of the MBVP corresponding to the control ψ_σ with the same boundary conditions but with *zero initial conditions*. (We do not wish to add w_H to both terms of the right-hand side in (18) below.)

The solution $w'(x, t)$ corresponding to the control $\phi'(x, t)$ is

$$w'(x, t) = \tilde{w}(x, t) + \epsilon w_\sigma(x, t). \tag{18}$$

The expression for total energy is

$$\mathcal{E}(w'(x, t), t) = \mathcal{E}(\tilde{w}(x, t), t) + 2\epsilon\langle\tilde{w}, w_0\rangle + \epsilon^2\mathcal{E}(w_\sigma(x, t), t). \tag{19}$$

Let w_σ be such that w' is not an optimal displacement:

$$\mathcal{E}(w'(x, t), T) > \mathcal{E}(\tilde{w}(x, t), T). \tag{20}$$

Then $\langle\tilde{w}, w_\sigma\rangle_{t=T} \geqq 0$.

Since ϵ was arbitrary, and the total energy is a continuous function of time, there must also exist an interval $[T - \epsilon, T]$ such that

$$\langle\tilde{w}, w_\sigma\rangle \geqq 0 \quad \text{for all} \quad t \in [T - \epsilon, T].$$

By Corollary 2 of Lemma 1, $\langle v, w_\sigma\rangle = \text{const.}$ on the interval $[t + \sigma, T]$ since in this interval $w_\sigma(x, t)$ is a solution of the homogeneous MBVP, while $v(x, t)$ is a solution of the homogeneous MBVP by hypothesis.

Hence,

$$\langle v, w_\sigma\rangle_{t=\tau+\sigma} = \langle v, w_\sigma\rangle_{t=T} = \langle\tilde{w}, w_\sigma\rangle_{t=T} \geqq 0.$$

In the interval $[0, \tau - \sigma]$, we have $\langle\tilde{w}, w_\sigma\rangle \equiv 0$ since $w_\sigma \equiv 0$ in that interval. In the interval I_σ we consider the limit:

$$\lim_{\sigma\to 0} \frac{1}{\sigma}\int_{\tau-\sigma}^{\tau+\sigma} \langle v, w_\sigma\rangle\, dt = \lim \frac{1}{\sigma}\int_{\tau-\sigma}^{\tau+\sigma}\int_{-l/2}^{+l/2}\left[\psi_\sigma \frac{\partial v(x, t)}{\partial t}\right] dx\, dt = 0$$

uniformly, since τ was a point of continuity of ψ_σ, and $\partial v/\partial t$, ψ_σ are smooth functions of time in I_σ. Consequently, $\langle v, w_\sigma\rangle \geqq 0$ in I_σ. Collecting these results, we have $\langle v, w_\sigma\rangle \geqq 0$ in $[0, T]$.

Using (13) and (14) we have

$$\int_{-l/2}^{+l/2}\left[-\tilde{\phi}(x, t)\frac{\partial v(x, t)}{\partial t}\right] dx = \max \int_{-l/2}^{+l/2}\left[-\phi(x, t)\frac{\partial v(x, t)}{\partial t}\right] dx$$

for all admissible controls $\phi(x, t)$ and for all $t \in [0, T]$. The proof is complete since τ, σ were arbitrary and ϕ', $\tilde{\phi}$ were piecewise smooth functions of time, and any admissible control could be obtained by altering $\tilde{\phi}(x, t)$ on a collection of suitable intervals I_σ.

Note. The proof is unchanged if $t = \tau$ is chosen to be a regular point in the sense of Pontryagin, rather than a point of continuity in $[0, T]$. Hence the theorem goes through with weaker hypothesis regarding $\tilde{\phi}(x, t)$, but with more complex arguments.

7. Application of Theorem 3. At first it seems that the result is only of theoretical interest, since in an attempt at direct application we need to know the finite state of the beam following an optimum control to decide if the control was optimal. This situation arises in a number of physical problems and is usually dealt with by some iterative schemes. The immediate value of the maximum principle as stated in Theorem 3 is in providing *negative* answers to the question: Is a proposed control optimal? An easy example of such application is given below.

Consider a beam which is simply supported at both ends, that is, conditions (4b) are satisfied at the points $x = \pm\, l/2$. The initial condition is given by

$$w(x, 0) = \frac{p}{24E \cdot I}\left[l^3\left(x + \frac{l}{2}\right) - 2l\left(x + \frac{l}{2}\right)^2 + \left(x + \frac{l}{2}\right)^4\right],$$

where p, E and I are constant, $\partial w(x, 0)/\partial t \equiv 0$. This represents the case of an initial deflection due to a constant load p (say due to a wind load), with the load being suddenly removed at the time $t = 0$.

The fundamental frequency of the beam is given by

$$n_1 = \frac{\pi}{2}\frac{\sqrt{g}EI}{\rho l^4}.$$

We now propose what would intuitively appear to be a "good" control. Among the piecewise continuous functions $\phi(x, t)$ which have the property $\int_{-l/2}^{+l/2} \phi(x, t)\, dx \leqq p \cdot l$ for all $t \geqq 0$, we select the control

$$\phi(x, t) = -p \cdot \operatorname{sgn} \frac{\partial w_H}{\partial t},$$

where the sign function is given by

$$\operatorname{sgn} y = \begin{cases} -1 & \text{if } y < 0, \\ 0 & \text{if } y = 0, \\ +1 & \text{if } y > 0, \end{cases}$$

$$w_H = w(x, 0) \cos 2\pi n_1 t,$$

and hope that this control is optimal for the interval $[0, n_1/2]$.

This is seen to be incorrect (without even applying Pontryagin's principle), since this control amounts to an immediate restoration of the static

load, and the total energy of the beam will remain constant. Clearly a better control is attained by assuming

$$\phi(x, t) \equiv 0, \qquad 0 \leqq t \leqq \frac{n_1}{4},$$

$$\phi(x, t) = -p \operatorname{sgn} \frac{\partial w_H}{\partial t}, \qquad \frac{n_1}{4} < t \leqq \frac{n_1}{2}.$$

The total energy dissipated by application of this control is given by

$$\mathcal{E}_D = \int_{n_1/4}^{n_1/2} \int_{-l/2}^{+l/2} \phi(x, t) \frac{\partial w_H}{dt}(x, t)\, dx\, dt.$$

The initial energy of the beam is given by

$$\mathcal{E}(0) = \frac{1}{2} \int_{-l/2}^{+l/2} EI \left(\frac{\partial^2 w(x, 0)}{\partial x^2} \right)^2 dx$$

and the final energy by

$$\mathcal{E}\left(\frac{n_1}{2}\right) = \frac{1}{2} \int_{-l/2}^{+l/2} EI \left(\frac{\partial^2 w(x, n_1/2)}{\partial x^2} \right)^2 dx.$$

Then,

$$w\left(x, \frac{n_1}{2}\right) = \frac{p_1}{24EI} \left(l^3 \left(x + \frac{l}{2} \right) - 2l \left(x + \frac{l}{2} \right)^2 + \left(x + \frac{l}{2} \right)^4 \right),$$

where p_1 is computed from the relationship

$$\mathcal{E}(n_1/2) = \mathcal{E}(0) - \mathcal{E}_D .$$

The functions $w_H(x, t)$ and $v(x, t)$ of Theorem 3 are scalar multiples of each other, so that in the interval $[n_1/4, n_1/2]$, $\phi(x, t)$ may well be optimal. However, it is not optimal in the interval $[0, n_1/4]$ where any force $f(x, t)$ equal to $-(p \cdot l \cdot \operatorname{sgn}(\partial v/\partial t))$ on a subinterval $[\tau - \sigma, \tau + \sigma]$ with σ sufficiently small, and equal to zero otherwise, results in

$$\int_{-l/2}^{+l/2} -\left(f(x, t) \cdot \frac{\partial v}{\partial t} \right) > 0.$$

Hence, $\phi(x, t) \equiv 0$ is not optimal on the subinterval $[0, n_1/4]$. We can now easily modify $\phi(x, t)$ to improve the control on the subinterval $[n_1/8, n_1/4]$, etc. It is clear how the maximal principle can be used to effect a gradual improvement of some arbitrarily chosen control. A more systematic approach using nonlinear programming techniques has been discussed in [8]. The inconvenience of having to compute the final state before being able to check the optimality is obvious.

In the next section of this paper we shall introduce a different criterion

of optimality, which results in a form of Pontryagin's principle requiring the knowledge of only the initial and current displacements.

8. Some general discussion. We stress the assumption $\mathcal{E}(\tilde{\phi}(x, t), t) > 0$ for all $t \in [0, T]$ of Theorem 3. If $\mathcal{E}(\tilde{\phi}(x, T), T) = 0$, Theorem 3 is meaningless, because $v(x, t) \equiv 0$. We shall be able to restate the maximal principle in a meaningful way for the case when $\mathcal{E}(\tilde{\phi}(x, t), t) > 0$ when $t \in [0, T)$, allowing $\mathcal{E}(\tilde{\phi}(x, t), T)$ to be equal to zero. However, if $\mathcal{E}(\tilde{\phi}(x, t), t) \equiv$ const. on some subinterval of $[0, T]$, it appears to be impossible to patch up the difficulties.

In particular, we want to avoid the situation demonstrated by Fig. 1. Consider a beam vibrating freely, so that the homogeneous equation (1) has a solution $w_H(x, t)$ such that $\mathbf{w}(x, 0) = \mathbf{w}(x, t_1)$. If we choose the value T large enough, it is possible to apply two optimal controls for the interval $[0, T]$:$\phi_1(x, t)$ and $\phi_2(x, t)$, such that $\phi_2(x, t + t_1) = \phi_1(x, t)$ when $t > t_1$, $\phi_2(x, t) \equiv 0$ when $t < t_1$, and $\mathcal{E}(\phi_1, T) = \mathcal{E}(\phi_2, T) = 0$. We specifically want to avoid optimal controls like $\phi_2(x, t)$. Before making a more precise statement we need to prove the following theorem.

THEOREM 4. *Let $\tilde{\phi}(x, t)$ be a time optimal control reducing the total energy of the beam to the value $E < \mathcal{E}(0)$ in the shortest possible time $T > 0$. Then $\tilde{\phi}(x, t)$ is an optimal control for the fixed interval $[0, T]$ control problem.*

Note. This theorem is known. Its proof is given for convenience.

Proof. Assume to the contrary that $\tilde{\phi}(x, t)$ is not optimal for the interval $[0, T]$. Hence there exists an admissible control $\psi(x, t)$, such that $\mathcal{E}(\psi(x, t), T) < E$. Since $\mathcal{E}(\psi(x, t), 0) = \mathcal{E}(0) > E$, and since $\mathcal{E}(\psi(x, t), t)$ is a continuous function of time, therefore there must be a time $t = \tau$, $0 < \tau < T$, such that $\mathcal{E}(\psi(x, t), \tau) = E$. This contradicts the time optimal property of the control $\tilde{\phi}(x, t)$, thereby proving the theorem.

Remark 1. The converse theorem is clearly false, as can be demonstrated by examples similar to the one illustrated in Fig. 1.

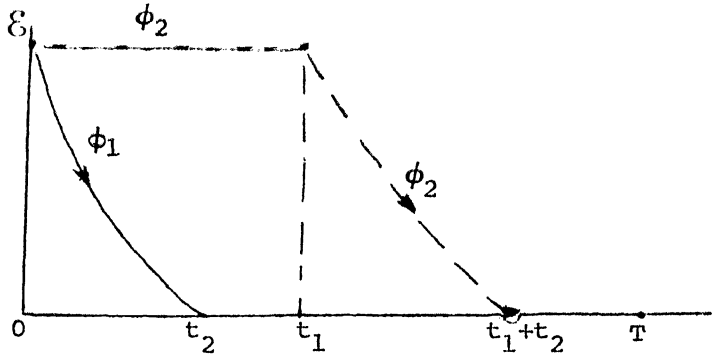


FIG. 1

Remark 2. Theorem 4 did not use any hypothesis of the beam theory and is true in the general case. We shall now define an optimal control, which possesses the property of reducing the total energy at the maximum rate at each instant of time, and show that Pontryagin's principle formulated for a control optimal in the above sense utilizes only the displacement vector, which could at least theoretically be computed without the knowledge of the finite state.

To state our next definition we need to introduce some concepts from the theory of generalized functions. In what follows, the controls $\phi(x, t)$ will be either integrable and square integrable functions on the interval $[-l/2, +l/2]$ for a fixed t, or the $\phi(x, t)$ could be generalized functions (for example, the Dirac delta function) whose supports are a finite number of points in $[-l/2, +l/2]$. By a theorem due to Lidskii, $\phi(x, t)$ is a derivative of a continuous function (see [4, p. 146]). In our discussion $\phi(x, t)$ will be either the Dirac delta function, or its derivative, or else it will be a regular function in $L_2 \cap L_1[-l/2, +l/2]$. The space of test functions $w(x, t)$ or $\partial w(x, t)/\partial t$ can be enlarged to include functions only of the class C^1 in $[-l/2, +l/2]$, since for all such functions the product $(\delta(x - \xi), w(x))$ and $(\delta(x - \xi), \partial w(x, t)/\partial t)$ is defined (for any ξ in the interval $[-l/2, +l/2]$). If $\phi(x, t)$ is a regular function in $L_2[-l/2, +l/2]$, then the product (ϕ, w) is given by

$$\int_{-l/2}^{+l/2} \phi(x, t) \cdot w(x, t)\, dx.$$

We interpret similarly the product $(\phi(x, t), \partial w(x, t)/\partial t)$. Since $w(x, t)$ is integrable in $[-l/2, +l/2]$, we can define

$$\| w(x, t) \| = \int_{-l/2}^{+l/2} | w(x, t) |\, dx_{(t=\text{const.})}.$$

The norm of the functional $\phi(x, t)$ can be defined as

$$\| \phi \| = \sup_{\| w \| = 1} (\phi, w).$$

$\| \phi \| \leqq K$ will define admissibility of a control. We shall say that a sequence of generalized functions ϕ_i converges to a generalized function ϕ if, for every test function ψ, we have $\lim_{i \to \infty} (\phi_i, \psi) = (\phi, \psi)$. We shall use the accepted notation $\int_a^b \phi(x, t)\psi(x, t)\, dx$ instead of $(\phi, \psi)_{[a,b]}$ even though ϕ may not be a locally integrable function. (It is well known that the Dirac delta function, or its derivatives, are *not* locally integrable functions.)

It is not hard to establish some sufficient conditions for the existence of a unique generalized function ϕ, such that $\lim_{i \to \infty} (\phi_i, \psi) = (\phi, \psi)$ for every test function ψ. Examples of sufficient conditions are given below in (i), (ii) and (iii).

(i) ϕ_i are regular functionals (that is, are locally summable) with a bounded support and obey the hypothesis of the Lebesgue theorem on dominated convergence.

(ii) ϕ_i have a bounded support, decrease (increase) monotonically in every neighborhood of any point which lies in their support, and are bounded below (above) by a locally summable function.

(iii) ϕ_i have a bounded support, their norm is bounded by some constant, and the sequence of numbers (ϕ_i, ψ) converges for every test function ψ.

For the case when it is possible to assign a norm to each test function, the proof is given, as previously stated, in the Lemma in Appendix 1. For the proof of this statement, in the case of the test function space K ($\psi \in K$ implies $\psi \in C_0^\infty$) see [4, Appendix A, pp. 368–369].

We shall now introduce a limiting process, which will allow us to resolve some of the difficulties.

Let $\tilde{\phi}_0(x)$ be the time optimal control associated with the energy level E, $E < \mathcal{E}$ $(t = 0)$, reducing $\mathcal{E}(t)$ to value E in the time $t = T_0$. Theorem 4 asserts that $\tilde{\phi}(x)$ is also an optimal interval control for the interval $[0, T_0]$.

We can subdivide this interval selecting points $t_0, t_{1,1}, t_{2,1}, \cdots, t_{n,1}$; $t_0 = 0, t_{n,1} = T_0$, such that the corresponding energy levels are

$$\mathcal{E}(0) = E_0 > E_{1,1} > E_{2,1} > \cdots > E_{n,1} = E.$$

We now introduce a collection of admissible controls $\phi_{i,j}$, such that $\phi_{1,1}$ reduces the energy from E_0 to $E_{1,1}$ in the shortest possible time, subject to initial conditions on $w(0, x)$, $\partial w(0, x)/\partial t$ given in the statement of the problem. The final condition for this subinterval: $[w(t_{1,1}), \partial w(t_{1,1})/\partial t]$ is then uniquely determined by $\phi_{1,1}$. The admissible control $\phi_{2,1}$ reduces the energy to the value $E_{2,1}$ in the shortest possible time, with the initial conditions for $\phi_{2,1}$ coinciding with the final conditions of $\mathrm{w}(\phi_{1,1})$. On each interval we formulate the minimal time control problem and find a solution $\phi_{i,1}$. (These are not necessarily unique.) The collection of controls $\phi_{i,1}$ determines a control ϕ_1 which reduces the total energy to the value E in some time T_1, $T_1 \geqq T_0$. Now we subdivide further the energy interval $[E_0, E]$, that is, we subdivide each of the subintervals $[E_{i,1}, E_{i+1,1}]$, and establish a new control $\phi_2(x, t)$. We carry out a sequence of subdivisions, such that

$$\lim_{j \to \infty} (E_{i-1,j} - E_{i,j}) = 0 \quad \text{and} \quad | t_{i-1,j} - t_{i,j} | \to 0.$$

The corresponding controls ϕ_j satisfy the conditions (iii), and we conclude that there exists a unique generalized function $\hat{\phi}(x, t)$:

$$\hat{\phi}(x, t) = \lim_{j \to \infty} \phi_j(x, t), \quad \| \hat{\phi} \| \leqq \max \| \phi_j \|.$$

From the construction of the generalized function $\hat{\phi}(x, t)$ we can immediately deduce some important properties.

THEOREM 5. *$\hat{\phi}(x, t)$ satisfies the maximum principle:*

$$\int_{-l/2}^{+l/2} \left[-\hat{\phi}(x, t) \frac{\partial w(\hat{\phi}, x, t)}{\partial t} \right] dx = \max \int_{-l/2}^{+l/2} \left[-\phi(x, t) \frac{\partial w(\phi, x, t)}{\partial t} \right] dx$$

for all admissible controls $\phi(x, t)$, and for any t in the domain of the control $\phi(x, t)$, such that $E(t) > E$.

Proof. Given $\epsilon > 0$ there exists an index K_1, such that on each subinterval $[t_{i-1,j}, t_{i,j}]$,

$$\int_{-l/2}^{+l/2} \left[-\phi(x, t) \frac{\partial w(\phi(x, t))}{\partial t} + \phi_j(x, t) \frac{\partial w(\phi_j, x, t_{i,j})}{\partial t} \right] dx < \frac{\epsilon}{2}$$

for all $j > K_1$.

Also there exists an index K_2 such that

$$\int_{-l/2}^{+l/2} \left[\phi_j(x, t) \frac{\partial w(\phi_j(x, t), t_{i,j})}{\partial t} + \phi_j \frac{\partial v(x, t_{i,j})}{\partial t} \right] dx$$

$$= \int_{-l/2}^{+l/2} \phi_j(x, t) \cdot \frac{\partial}{\partial t} \{ v(x, t_{i,j}) - w(x, t_{i,j}) \} \, dx < \frac{\epsilon}{2}$$

for all $j > K_2$ (where $v(x, t)$ denotes as before the solution of the homogeneous equation). This follows from the hypotheses that

$$v(x, t_{i,j}) = w(x, t_{i+1,j}),$$

that v and w are continuously differentiable, and their time derivatives are uniformly bounded.

Choosing $K = \max(K_1, K_2)$, we have for all $j > K$,

$$\int_{-l/2}^{+l/2} \left[-\phi(x, t) \frac{\partial w(\phi(x, t))}{\partial t} + \phi_j \frac{\partial (v(x, t))}{\partial t} \right] dx < \epsilon$$

for all $t \in [0, T]$, where $T = \lim T_i$. (In general we have to allow for the possibility that $T = +\infty$, although in the case of a vibrating beam it can be shown that T always exists.) Since for each jth subdivision, $\phi_j(x, t)$ is the optimal control on each fixed interval $[t_{i-1,j}, t_{i,j}]$,

$$\int_{-l/2}^{+l/2} \left[-\phi_j \frac{\partial v(x, t_{i,j})}{\partial t} \right] dx = \max \int_{-l/2}^{+l/2} \left[-\phi \frac{\partial v(x, t_{i,j})}{\partial t} \right] dx$$

for any admissible control $\phi(x, t)$ having the same final state $w(x, t_{i,j}) = v(x, t_{i,j})$. An obvious argument leads now to the desired conclusion:

$$\int_{-l/2}^{+l/2} \left[-\hat{\phi}(x, t) \frac{\partial w(\hat{\phi}(x, t))}{\partial t} \right] dx$$

$$= \max \int_{-l/2}^{+l/2} - \left[\phi(x, t) \frac{\partial w(\phi(x, t))}{\partial t} \right] dx. \tag{21}$$

Since this integral is equal to

$$\frac{d}{dt}\langle w(\hat{\phi}, x, t), w(\hat{\phi}, x, t)\rangle = \frac{d}{dt} E(w(\hat{\phi}, x, t)),$$

the control $\hat{\phi}(x, t)$ has the property that the total energy decreases at a maximum rate at each point of the time interval $[0, \hat{T}]$.

The principle (21) is valid if $E(w(\hat{\phi}(x, t)) > 0$ for all $t \in [0, \hat{T}]$, regardless of whether $E(T)$ is positive, or equal to zero. The control $\hat{\phi}(x, t)$ shall be called *instantly optimal.*

THEOREM 6. *The instantly optimal control $\hat{\phi}(x, t)$ is unique, i.e., $\hat{\phi}(x, t)$ does not depend on the limiting process chosen, or on the properties of the elements of the sequence of time minimal functions $\{\hat{\phi}_i(x, t)\}$.*

Proof. Assume that the generalized functions $\hat{\phi}_1(x, t)$, $\hat{\phi}_2(x, t)$, both satisfy $\lim_{i\to\infty} \tilde{\phi}_i^{(1)} = \hat{\phi}_1$ and $\lim_{i\to\infty} \tilde{\phi}_i^{(2)} = \hat{\phi}_2$ for two limiting processes, as described above. We use the convexity of the time optimal controls and obtain

$$\lim_{i\to\infty} \tfrac{1}{2}(\tilde{\phi}_i^{(1)} + \tilde{\phi}_i^{(2)}) = \tfrac{1}{2}(\hat{\phi}_1 + \hat{\phi}_2),$$

which is also an instantly optimal control. If $\mathbf{w}_1(x, t) \triangleq [w(x, t), \partial w(x, t)/\partial t]$ is the displacement vector corresponding to $\hat{\phi}_1(x, t)$ and $\mathbf{w}_2(x, t)$ corresponds to $\hat{\phi}_2(x, t)$, then $\frac{1}{2}(\mathbf{w}_1 + \mathbf{w}_2)$ corresponds to $\frac{1}{2}(\hat{\phi}_1 + \hat{\phi}_2)$ by linearity (see (8)). By construction of $\hat{\phi}_1$ and $\hat{\phi}_2$ and using Theorem 2, we have

$$\mathcal{E}(\mathbf{w}_1(\hat{\phi}_1(x, t), t)) = \mathcal{E}(\mathbf{w}_2(\hat{\phi}_2(x, t), t))$$

for all $t \in [0, \hat{T}]$. Hence,

$$\langle w_1 , w_1\rangle = \langle w_2 , w_2\rangle = \tfrac{1}{4}\langle (w_1 + w_2),(w_1 + w_2)\rangle.$$

The use of Lemma 2 (Cauchy-Schwarz inequality) shows that we have in fact equality $\langle w_1 , w_2\rangle^2 = \langle w_1 , w_1\rangle \cdot \langle w_2 , w_2\rangle$, and the equality $\mathbf{w}_1 = \mathbf{w}_2$ easily follows for all $t \in [0, T]$. But this implies that $\hat{\phi}_1(x, t)$ and $\hat{\phi}_2(x, t)$ are equal in the sense of our definition of §1, that is, they belong to the same equivalence class.

Remark. It is easy to see that, given E, $0 < E < \mathcal{E}(0)$, and given the instantly optimal control $\hat{\phi}(x, t)$, such that $\mathcal{E}(w(\hat{\phi}(x, t), \hat{T})) = E$, the time $\hat{T}$ is either greater or equal to the minimal time $\tilde{T}$ corresponding to the minimal time optimum control $\tilde{\phi}(x, t)$. It is important to exhibit cases when a control $\hat{\phi}(x, t)$ exists which is instantly optimal, but is not time optimal, to justify the separate definitions. The possibility of such a situation is demonstrated by a simple argument. Consider the initial

conditions

$$w(x, 0) \equiv 0, \quad x \in \left[-\frac{l}{2}, +\frac{l}{2}\right],$$

$$\frac{\partial w(x, 0)}{\partial t} = \frac{p}{24EI}\left[l^3\left(x + \frac{l}{2}\right) - 2l^2\left(x + \frac{l}{2}\right)^2 + \left(x + \frac{l}{2}\right)^4\right]$$

(corresponding to a uniform load case). The time optimal control $\tilde{\phi}(x, t)$ will satisfy the relationship (17), i.e.,

$$\int_{-l/2}^{+l/2} -\tilde{\phi}(x, t)\,\frac{\partial v(x, t)}{\partial t}\,dx = \max \int_{-l/2}^{+l/2} -\phi(x, t)\,\frac{\partial v(x, t)}{\partial t}\,dx,$$

$$\| \tilde{\phi} \| \leqq K.$$

On the other hand, the instantly optimal control would satisfy

$$\int_{-l/2}^{+l/2} -\hat{\phi}(x, t)\,\frac{\partial w(\hat{\phi}, x, t)}{\partial t}\,dx = \max \int_{-l/2}^{+l/2} -\phi(x, t)\,\frac{\partial w(\phi, x, t)}{\partial t}.$$

It is easy to show that the instantly optimal control at the time $t = 0$ is the Dirac delta function (a point load) of a given norm K (which is sufficiently small) applied at the point $x = 0$.

Consider some time $t = \tau$ during which the beam has been deformed by the application of point loads applied to the points of highest velocity of the beam. Letting the beam vibrate freely *back* to the initial time $t = 0$, we check numerically that in general a completely different configuration of the beam $w(x, 0) \not\equiv 0$ is obtained, and in general the product

$$\int_{-l/2}^{+l/2} \hat{\phi} \cdot \frac{\partial v(x, 0)}{\partial t}\,dx$$

will not coincide with

$$\int_{-l/2}^{+l/2} \hat{\phi}\,\frac{\partial w(\hat{\phi}, x, 0)}{\partial t}\,dx,$$

showing that $\hat{\phi}(x, t)$ was not time optimal for the interval $[0, \tau]$.

Appendix 1. A lemma on the completeness of some spaces of generalized functions. In certain physical problems special types of generalized functions (for example the Dirac delta function) are applied to classes of test functions which are much larger than the classes K, S or Z generally discussed in literature (see, for example, [1] or [4]). For example, the linear map δ:

$$(\delta(x - \xi), \quad \phi(x)) = \phi(\xi)$$

makes sense when δ is applied to any function which is only continuous in some neighborhood of the point $x = \xi$. (Of course, even less restrictive conditions could be proposed for $f(x)$ under which this map would be defined.) In the problems of deflections of beams or thin plates we consider the effects of point loads (the Dirac delta functions), of point moments (the derivative of the Dirac delta function), etc., acting on the space of test functions $w(x)$, or $w(x, y)$ which can only be assumed to be of the class C^1 in a compact, connected region Ω of E^1 or E^2, respectively, with the boundary of $\Omega \subset E^2$ satisfying some additional hypothesis. The admissible deflection functions, that is, the functions defined and differentiabln in Ω vanishing outside Ω and satisfying an appropriate differential equation almost everywhere in Ω, are easily assigned a norm.

For example, we can define

$$\| w \| = \int_\Omega | w | \, d\mu,$$

where μ is the usual Lebesgue measure in Ω. A linear functional f can be assigned a norm

$$\| f \| = \sup_{\| \phi \| = 1} | (f, \phi) |.$$

f is said to be unbounded if $\| f \| = \infty$. It follows from the definition of $\| f \|$ that $|(f, \psi)| \leqq \| f \| \cdot \| \psi \|$. We are now ready to state a theorem which is the analogue of the theorem due to Brodskii proving the completeness of the space K'. (For the statement and proof of this theorem see [4, Appendix A, pp. 308–309].)

LEMMA. *Let Φ be a space of test functions such that Φ is a normed (but not necessarily complete) vector space, and let $f_1, f_2, \cdots, f_n, \cdots$ be a sequence of continuous linear functionals mapping the elements of Φ into R^1, such that the sequence of real numbers (f_i, ϕ) converges for every vector $\phi \in \Phi$, and such that*

$$\| f_i \| < M, \quad i = 1, 2, \cdots,$$

for some $M > 0$.

Then there exists a continuous linear functional $f \in \Phi^*$ (where Φ^* denotes the dual or conjugate space of Φ) such that

$$\lim_{i \to \infty} (f_i, \phi) = (f, \phi) \quad \textit{for every} \quad \phi \in \Phi.$$

Proof. We define the functional f by the formula $(f, \phi) = \lim_{i \to \infty} (f_i, \phi)$. f is linear, since for any real numbers α, β we have

$$\begin{aligned} (f, \alpha\phi_1 + \beta\phi_2) &= \lim_{i \to \infty} (f_i, \alpha\phi_1 + \beta\phi_2) \\ &= \lim_{i \to \infty} \{\alpha(f_i, \phi_1) + \beta(f_i, \phi_2)\} = \alpha(f, \phi_1) + \beta(f, \phi_2). \end{aligned}$$

f is easily shown to be a continuous functional, since $\| f \| \leqq \sup \| f_i \| \leqq M$, and therefore f is bounded in the norm.

REFERENCES

[1] N. I. AKHIEZER AND I. M. GLAZMAN, *The Theory of Linear Operators in Hilbert Space*, vol. 1, Ungar, New York, 1961.

[2] A. G. BUTKOVSKII AND A. YA. LERNER, *Optimal controls with distributed parameters*, Dokl. Akad. Nauk SSSR, 134 (1960), pp. 778–781.

[3] H. FORD, *Advanced Mechanics of Materials*, John Wiley, New York, 1963.

[4] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, vol. 1, Academic Press, New York, 1964.

[5] J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Non-linear Oscillations, vol. V, Ann. of Math. Studies no. 45, Princeton University Press, Princeton, 1960, pp. 1–24.

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

[7] D. L. RUSSELL, *Linear symmetric hyperbolic systems*, this Journal, 4 (1966), pp. 276–294.

[8] ———, *Optimal regulation of linear symmetric hyperbolic systems with finite-dimensional controls*, Rep. 566, Mathematics Research Center, U. S. Army, University of Wisconsin, Madison, 1965.

[9] A. I. EGEROV, *Optimal processes and invariance theory*, this Journal, 4 (1966), pp. 601–661.

[10] G. K. NAMAZOV, *Boundary value problems for a second order parabolic equation with discontinuous coefficients*, Izv. Akad. Nauk Azerbaidzan. SSR Ser. Fiz-Tehn. Mat. Nauk, no. 3, 1961, pp. 39–46.

[11] O. A. OLEINIK, *Boundary value problems for linear elliptic and parabolic equations with discontinuous coefficients*, Izv. Akad. Nauk SSSR Ser. Mat., 25 (1961), pp. 3–20.

[12] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints, I, II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–430.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

APPROXIMATION THEOREMS ON SOME CLASSES OF AUTOMATA*

ANDRE DE KORVIN†

Abstract. This paper considers a machine as a pair (G, M), where G is a group or a semigroup and where M is a state-space. The first part of the paper considers the case where G is a locally compact group and M is any locally compact space. The essential requirement is that $(x, p) \to x(p)$ be continuous where $x \in G$, $p \in M$ and $x(p) \in M$: i.e., we require that the next state function be continuous. The notion of projective limit is discussed and a criterion is given as to when G is the projective limit of some of its quotient groups. Next an infinitesimal element is defined. An identification is then made near the respective identities of G and the set of infinitesimal operations.

The second part of the paper treats the case when G is a so-called amenable semigroup, having a representation of bounded operators on a Hilbert space. In the case in which the representation is an isometry, weakly continuous, a decomposition theorem is given. On a particular subspace the representation turns out to be a direct sum of finite-dimensional operations. Diverse characterizations of that space are given. Next the notion of coordinates of a representation is defined and two orthogonality theorems are stated.

The whole paper might be considered as an attempt at giving approximation theorems on essentially infinite automata.

1. Introduction. A classical way to view a machine is as follows: Let M be any set which represents the status of the machine and let G be a group or semigroup of transformations on M. The pair (G, M) constitutes a machine (see [1], [3], [8], [10]). It is then meaningful to study purely algebraic properties of G (see [2], [5], [6], [11]). In this paper a machine is considered as a pair (G, M), where G is a topological group or semigroup and M is a topological space. It is then meaningful to talk about topological properties of groups and semigroups. This paper brings into focus such relevant properties. The techniques are well known to people working in topological groups (see [4], [7], [9]).

2. "Reversible-state" machines.

2.1 Definitions.

(a) Let M be a set with a topology t defined on it. Assume M is locally compact in this topology. M will be called a *state-space.*

Example 1. Let M be finite, and let t be the discrete topology.

Example 2. Let M be a set of n-tuples with the usual topology.

Example 3. Let M be a set of $n \times n$ matrices with the usual norm topology.

(b) Let G be a group of transformations on the space M. Suppose G

* Received by the editors December 20, 1966, and in revised form January 11, 1968.

† Department of Mathematics, Carnegie-Mellon University, Schenley Park, Pittsburgh, Pennsylvania 15213.

is topologized so as to be locally compact (i.e., G consists of continuous maps of the space M into itself; G has an algebraic group structure; the group operations are continuous on G; moreover, G is locally compact in that topology). G will then be called a *tape group on* M.

(c) A *reversible-state machine* will be a pair (G, M), where M is a state-space and G is a tape group on M. In addition we shall require that the map $(x, p) \to x(p)$, where $x \in G$ and $p \in M$, be continuous.

(d) A reversible-state machine is determined if one knows its set of states and how a series of inputs acts on a particular state. G can be thought of as a set whose elements are a series of inputs with the obvious composition law. Because elements of G are continuous maps on M, this means that *if a series of inputs is slightly modified, then the resulting state is also slightly modified.* The topologies on M and G of course give meaning to the concept of "slight" in G and M. A machine of the above type will be called *transitive* if, given p and q in M, there exists x in G such that $x(p) = q$.

2.2 Quotient state-spaces. 1. Let H be a closed subgroup of G (not necessarily normal). Consider the map π which maps any element x of G into the set xH (all elements of form x multiplied by an element of H). Pick the open sets in G/H to be the ones whose inverse by π are open in G. Then π is a continuous map of G into G/H. Moreover, π maps open sets of G into open sets of G/H. G/H is a locally compact Hausdorff space. All these statements follow by definition.

2. We can state the following theorem.

THEOREM 1. *Let H be a closed subgroup of G; then $(G, G/H)$ is a transitive reversible-state machine.*

Before proving this theorem, let us give a physical interpretation. Starting with a machine which has a given state-space, we are able to construct in general a new state-space on which G is a tape group. Moreover, the new state-space is related to the first state-space as follows: If two tape elements x_1 and x_2 are such that x_1 can be obtained by composition of x_2 with an element of H, then $\pi(x_1)$ and $\pi(x_2)$ are identical states in the new state-space.

Proof. Define $\rho_y(xH) = yxH$. Then, given x_1H and x_2H, $\rho_{x_2x_1^{-1}}(x_1\mathrm{H}) = x_2H$, so the machine is transitive.

Now $(x, y) \to xy$ is continuous and $(xy) \to (xy)H$ is continuous. The map $(x, y) \to (x, yH)$ preserves open sets; hence $(x, yH) \to (xy)H$ is continuous.

3. Consider the machine (G, M). Let p be an element of M. Let G_p be the set of tape elements of G which leave the state p invariant. Clearly, G_p is a closed group. We can then form the state-space G/G_p. Now we can define a natural map $\psi G/G_p \to M$ by $\psi(xG_p) = x(p)$. Let us assume that (G, M) is a transitive machine. What can we say about ψ?

Let us verify first of all that ψ is a well-defined map. If $xG_p = yG_p$, then we have $y^{-1}xG_p$ and $y^{-1}x(p) = p$ so that $x(p) = y(p)$. This also shows that ψ is 1:1. ψ is onto M because we have a transitive machine. Let $\phi(x) = x(p)$ $(x \in G)$. By definition, ϕ is continuous. $\pi: G \to G/G_p$ is an open map; hence ψ is a continuous map. If we know that ϕ is open, then ψ is a homeomorphism (i.e., the spaces M and G/G_p are topologically equivalent). It is natural then to ask the following question: Under what conditions on G and M is G/G_p homeomorphic to M?

THEOREM 2. *Suppose (G, M) is a transitive machine. Assume G can be covered by countably many translates of each neighborhood of the identity.* (This is the case if G is separable.) *Then ψ is a homeomorphism.*

Note that this theorem essentially gives a criterion as to when G/G_p is the same state-space as M. The state-space G/G_p is obtained from G and M as follows. Each $x \in G$ is mapped in such a fashion that tape elements which "cancel" each other out on the state p are mapped into the same element. Then the natural quotient topology is introduced on that new set. Now we proceed to the proof of the theorem.

Proof. We remark that a locally compact space is not a countable union of nowhere dense sets.

Now we shall prove that $x \to \phi(x) = x(p)$ is an open map. Let $x_0 \in V \subseteq G$, where V is open. We shall show that $\phi(V)$ contains an open neighborhood about $\phi(x_0)$. Pick a compact neighborhood U of G such that $U = U^{-1}$ and $x_0U^2 \subseteq V$. By hypothesis there exists a sequence $\{x_n\}$ such that $G = V_nx_nU$ whose image by ϕ is $V_n\phi(x_nU) = M$. By the above remark one of the $\phi(x_nU)$ is not nowhere dense, so contains an open set. Thus $\phi(U)$ has an interior point u_0p $(u_0 \in U)$.

Therefore, $\phi(V)$ has $x_0(p)$ as an interior point.

2.3 Decomposition theorem. 1. Consider the tape group G. Let ρ_α be a continuous, open homomorphism of G onto G_α which is also a tape group. The problem we shall consider is the following: When is G topologically equivalent to the projective limit of the G_α? A criterion of this problem will give us a way to look at the tape group G as a projective limit of homomorphic tape groups, the projective limit being topologically equivalent to the tape group G.

2. Now let A be a directed set (i.e., for α, β in A there exists $\lambda \in A$ such that $\lambda > \alpha$, $\lambda > \beta$). To each α associate a tape group G_α (in this section we are not interested in the state-spaces). To each pair α, β of A such that $\beta > \alpha$ we associate an open homomorphism $\pi_{\beta\alpha}$ of G_β on G_α satisfying the following: If $\alpha < \beta < \lambda$, then $\pi_{\lambda\alpha} = \pi_{\lambda\beta}\pi_{\beta\alpha}$ (the convention here is to read from left to right). Form $\tilde{G} = \prod_{\alpha \in A} G_\alpha$ (i.e., $\tilde{G}$ is the product group with the natural topology on it). Now let $G = \{x \in G | x_\alpha$

$= \pi_{\beta\alpha}(x_\beta)$ whenever $\beta > \alpha\}$, where G is a subgroup of $\tilde{G}$. Let us topologize G by the relative topology respectively to G. It is easy to check the following:

(a) An open basis at $e \in G$ is defined by the sets of the form $\{x = \{x_\beta\} | x_\alpha \in$ open neighborhood of e in $G_\alpha\}$ where α is fixed.

(b) G is closed in $\tilde{G}$.

Now let κ_α be the projection of G in G_α. If κ_α is onto G_α, then G is the projective limit of the tape groups G_α. (Note that κ_α is an open map.)

THEOREM 3. *Let G be a tape group. N_α, $\alpha \in A$, is a collection of normal subgroups of G satisfying the following conditions:*

(i) *If $\alpha, \beta \in A$, then there exists $\gamma \in A$ such that $N_\gamma \leqq N_\alpha \wedge N_\beta$.*

(ii) *If U is any neighborhood of e, there exists $\gamma \in A$ such that $N_\gamma \leqq U$.*

(iii) *At least one of the N_α is compact. Then we can form G' the tape group which is the projective limit of the $G_\alpha = G/N_\alpha$.*

Further, let Φ map G into G' by $\Phi(g) = \{\pi_\alpha(g)\}$. Then Φ is an isomorphism onto G'.

Let us comment on this theorem before giving the proof. We have here a criterion as to "how thin" we must choose a set of tape subgroups such that the tape group G may be reconstituted as a projective limit of its quotient tape groups. G will be topologically equivalent to this projective limit. The physical interpretation of the projective limit is tentatively the following: For each α form the tape group G/N_α. Then consider the state-space which is the Cartesian product of the state-spaces of G/N_α. Then a state in that product is given as a sequence of states. The αth element of the sequence represents a state of the state-space of G/N_α. Then topologically the tape group can be considered as acting on that new state-space. The transformation of such a sequence is done "component-wise", i.e., g is identified with $\{g_\alpha\}$ and

$$g(p_1, p_2, \cdots, p_\alpha, \cdots) = (g_1(p_1), g_2(p_2), \cdots, g_\alpha(p_\alpha), \cdots).$$

Now we prove the theorem.

Proof. First, Φ is 1:1 and continuous. The verification is trivial. Now we have to show that Φ is open. Let $g \in G$ and $g \in V$ which is open in G.

Pick a neighborhood U of e such that $gU^2 \leqq V$. Select $\alpha \in A$ such that $N_\alpha \leqq U$. Consider $\{x \in G | x_\alpha \in \pi_\alpha(gU)\}$. This is an open set in the relative topology for the range of Φ. Indeed, if this set contains $\Phi(h)$ we shall show that $h \in V$.

We have $\pi_\alpha(h) \in \pi_\alpha(gU)$, so $h \in gUN_\alpha \leqq gU^2 \leqq V$.

Now if we show that range $\Phi = G'$, we shall have proved the theorem. Let $x = \{x_\alpha\} \in G'$, $x_\alpha \in G/N_\alpha$, $x_\alpha = s_\alpha N_\alpha$, $s_\alpha \in G$. Consider $\alpha_1, \cdots, \alpha_n$. Pick $\beta > \alpha_i$ for $i = 1, \cdots, n$. Then $N_\beta \leqq N_{\alpha_i}$ so $s_\beta N_\beta \leqq S_{\alpha_i} N_{\alpha_i}$ (because $\pi_{\beta\alpha_i}(x_\beta) = x_{\alpha_i}$). Because N_α is compact for some α, there exists $g \in \pi_\alpha s_\alpha N_\alpha$. Therefore $\pi_\alpha(g) = x_\alpha$.

2.4. Infinitesimal tape elements. In this section V will denote a finite-dimensional vector space over the real numbers. $M(V)$ = all linear transformations which map V into V. $GL(V)$ = all nonsingular linear transformations of $M(V)$. Hence $GL(V)$ is a group for multiplications. Define

$$\exp T = \sum_{k=0}^{\infty} \frac{1}{k!} T^k, \tag{1}$$

where $T \in M(V)$. Then it is easily seen that this exponential function maps topologically a neighborhood of 0 in $M(V)$ onto a neighborhood of I in $GL(V)$. Physically we shall keep in mind that V is the state-space, and $GL(V)$ is a tape group on V. Now let G be a closed subgroup of $GL(V)$. $x \in M(V)$ is called an *infinitesimal operation on* V if there exist a sequence $A_n \in G$, ϵ_n which decreases monotonically to zero, and

$$\frac{A_n - I}{\epsilon_n} \to x.$$

Convergence here is defined in the norm sense of $M(V)$. $L(G)$ is the set of all infinitesimal operations formed with elements of G. (Note that $L(G)$ is nothing else but the Lie algebra of G.) Now if G is a closed subgroup of $GL(V)$ and $x \in L(G)$, then $\exp(x) \in G$. This is a trivial conclusion using the fact that G is closed and that $A_n[1/\epsilon_n] \to \exp(x)$, where $[1/\epsilon_n]$ is the smallest integer above $1/\epsilon_n$.

Now in $L(G)$ introduce a new operation $[A, B] = AB - BA$, where A, $B \in L(G)$. It can be shown then that $L(G)$ is a linear subspace of $M(V)$ closed under [,] operation.

THEOREM 4. *The exponential function defined in* (1) *maps topologically some neighborhood of* 0 *in* $L(G)$ *onto some neighborhood of* I *in* G.

Let us comment now on this theorem. Here we are interested in tape elements which are not necessarily part of a tape group, i.e., the infinitesimal elements. The action of two infinitesimal elements on the state-space results in the usual vector addition. On the other hand, we are interested in the action of a tape group. The theorem says that if the tape elements and infinitesimal operations are close enough to the respective identities, then the infinitesimal operations are topologically equivalent to the tape elements. (This is really a local identification near the identities.)

We prove this theorem in the following manner.

Proof. All we have to show is that the exponential function is an open map. So assume that $\exp[L(G)]$ does not contain a neighborhood of I in G. Let N be the complementary subspace of $L(G)$ in $M(V)$:

$$M(V) = L(G) + N,$$

$$N \wedge L(G) = 0.$$

So there exists $\{A_n\}$ such that:

$$A_n \in G \quad \text{for all} \quad n,$$
$$A_n \to I,$$
$$\log A_n \notin L(G) \quad \text{for all} \quad n,$$
$$\log A_n = x_n + y_n,$$
$$x_n \in L(G),$$
$$y_n \in N,$$
$$y_n \neq 0 \quad \text{for all} \quad n$$

(for if y_n could equal zero then $\log A_n \in L(G)$, which is a contradiction).

$$\log A_n \to 0$$

so that

$$x_n \to 0 \quad \text{and} \quad y_n \to 0.$$

Consider $y_n/\| y_n \|$. The unit sphere is compact, so some subsequence converges to y. Therefore $\| y \| = 1$.

Since $y \in N$, $\exp \lambda y \neq I$ for small λ, because for small values of λ, the exponential is 1:1; and since $\lambda y \neq 0$, $\exp \lambda y \neq I$. If now we show that $y \in L(G)$, we shall have demonstrated the contradiction:

$$\begin{aligned}
\left\| \frac{1}{\| y_n \|} \{A_n - \exp x_n - y_n\} \right\| &= \frac{1}{\| y_n \|} \| \exp (x_n + y_n) - \exp x_n - y_n \| \\
&= \frac{1}{\| y_n \|} \left\| \sum_{k=2}^{\infty} \frac{1}{k!} ((x_n + y_n)^k - x_n{}^k) \right\| \\
&\leqq \frac{1}{\| y_n \|} \sum_{k=2}^{\infty} \frac{1}{k!} [(\| x_n \| + \| y_n \|)^k - \| x_n \|^k] \\
&= \frac{1}{\| y_n \|} [\exp (\| x_n \| + \| y_n \|) \\
&\qquad - \exp \| x_n \| - \| y_n \|] \\
&= \exp \| x_n \| \to 1 \quad \text{as} \quad \frac{\exp \| y_n \| - 1}{\| y_n \|} \to 0;
\end{aligned}$$

$$\begin{aligned}
\lim \frac{y_n}{\| y_n \|} &= \lim \frac{A_n - \exp x_n}{\| y_n \|} \\
&= \lim \exp x_n \left(\frac{\exp (-x_n) A_n - I}{\| y_n \|} \right) \\
&= \lim \frac{\exp (-x_n) A_n - I}{\| y_n \|} \in L(G);
\end{aligned}$$

thus $y \in N \wedge L(G)$, so $y = 0$, which is a contradiction.

THEOREM 5. *Let G_1 and G_2 be closed subgroups of $GL(V_1)$ and $GL(V_2)$. Let ϕ be a continuous homomorphism of G into Q. Then there exists a homomorphism $d\phi$ of $L(G)$ into $L(Q)$ such that:*

$$\begin{array}{ccc} L(G) & \xrightarrow{d\phi} & L(Q) \\ \exp\downarrow & & \downarrow\exp \\ G & \xrightarrow[\phi]{} & Q \end{array}$$

The diagram is commutative and $d\phi$ preserves the [,] operation. The proof of this theorem is a known proof of the theorem which says that a homomorphism between so-called classical groups can be extended to a homomorphism between their Lie algebras.

If the exponential function is physically interpreted as a transformer, i.e., if it changes an infinitesimal operation into a tape element of a group, with little deformations around the identities, then we can define a homomorphism (i.e., functor) between the corresponding infinitesimal operations which extends in a natural fashion the homomorphism of the tape group elements.

Using the classical groups theory we also have the following result: Let $M \in GL(V)$ and $G(M) = \{A \in M(V) | A^t M A = M\}$, i.e., elements of $G(M)$ leave M "invariant". Then

$$L[G(M)] = \{x \in M(V) | x^t M = -Mx\}.$$

In particular, this gives a straightforward way to compute the infinitesimal operations of the orthogonal, unitary and Lorentz group.

This ends what could be described as approximation theorems for "reversible machines". In the next part we shall attempt to describe some properties of "irreversible machines."

3. Representation theorems for "nonreversible machines."

3.1. Definitions.

(a) A *nonreversible machine* will be defined as a pair (G, M), where G is an "amenable" semigroup, i.e., G has the algebraic structure of a monoid (it is closed under composition and the identity is present). Let $C(G)$ be the set of all bounded and continuous functions. G is said to be *left-amenable* if there exists a linear functional defined on $C(G)$ which is positive, left-invariant, and normalized; the same definition applies for right-amenable. G is *amenable* if it has a left and a right mean. (Note that the left mean is not necessarily equal to the right mean.) Formally, a nonreversible machine will then be the pair (G, M), where G is a semigroup with

identity and where λ_1, λ_2 are linear functionals such that:

$$\lambda_i(I) = 1, \quad i = 1, 2,$$

where I is the identity element of $C(G)$,

$$\lambda_1 f(x) = \lambda_1 f(ax) \quad \text{for all} \quad a \in G, \quad f \in C(G), x \in G,$$

$$\lambda_2 f(x) = \lambda_2 f(xb) \quad \text{for all} \quad b \in G,$$

$$\lambda_i f \geqq 0 \quad \text{if} \quad f \geqq 0, \quad i = 1, 2.$$

Example 1. Any Abelian semigroup is amenable.

Example 2. Any compact group is amenable.

(b) We shall define now a so-called "almost convergence". Let $f \in C(G)$ and let α be a number. Then we say f *converges almost to* α if $\lambda f = \alpha$ for all λ where λ is a mean (left or right). We denote this fact as $f \overset{(a)}{\longrightarrow} \alpha$. It can be shown that if $G = \{1, 2, 3, \cdots\}$, in order that $f \in C(G)$ satisfies $f \overset{(a)}{\longrightarrow} \alpha$, it is necessary and sufficient that

$$\frac{1}{n} \sum_{k=1}^{n} f(k + m) \to \alpha \quad \text{uniformly in } m$$

(ergodic theorem). Now we assume that we have a Hilbert space h, and that the tape semigroup G has a representation in $L(h)$ as a set of bounded linear operators on h; i.e., if $x \in G$, then $x \to T_x \in L(h)$ and $T_{xy} = T_x T_y$.

Roughly speaking we assume now that each tape element (which is an element of a semigroup G) can be represented as a linear operator over the Hilbert space h (h in general is infinite-dimensional). We make the hypothesis that the semigroup G is amenable. We want to study the properties of such a representation, and hence, hopefully, have a different way to look at elements of G. A representation T_x is called *bounded* if $\| T_x \| \leqq k$ for all $x \in G$. A representation T_x is called *weakly continuous* if $x \to (T_x \zeta, \eta)$ is a continuous function on G for all $\zeta, \kappa \in h$.

3.2. Bounded, isometric representations.

THEOREM 6. *Let u_x be a bounded, weakly continuous representation of G on h. Then there exists $Q \in L(h)$ such that*

$$(u_x \zeta, \kappa) \overset{(a)}{\longrightarrow} (Q\zeta, \kappa) \quad \textit{for all} \quad \zeta, \kappa \in h.$$

Moreover, if u_x is an isometry, Q = projection on $\{\zeta | u_x \zeta = \zeta$ for all $x \in G\}$.

Proof. Fix λ a left or right mean. Then there exists Q_x such that $\lambda(u_x \zeta, \kappa) = (Q_\lambda \zeta, \kappa)$ for all $\zeta, \kappa \in h$. This is merely the Riesz theorem.

Observe that $Q_\lambda \zeta$ is an element of the closed convex hull of $\{u_x \zeta | x \in G\}$ (use the separation property for convex sets).

Now let λ_1 be a right mean.

CLAIM. $Q_{\lambda_1} u_x = Q_x$.

Indeed,

$$\begin{aligned}(Q_{\lambda_1} u_x\zeta, \kappa) &= \lambda_1(u_y\zeta, \kappa)\\ &= (Q_{\lambda_1}, \zeta, \kappa).\end{aligned}$$

If λ is a left mean, then it is equally trivial to check that $u_x Q_x = Q_x$.

Now we show that $Q_\lambda = Q_{\lambda_1}$. By the same token we shall have shown that Q_λ is independent of λ.

Consider $\zeta \in h$, $x_1, \cdots, x_n \in G$, $\alpha_i, \cdots, \alpha_n$ such that $\sum_1^n \alpha_i = 1$ and

$$\begin{aligned}Q_{\lambda_1}(\sum \alpha_i u_{x_i}(\zeta)) &= \sum \alpha_i Q_{\lambda_1} u_{x_i}\zeta\\ &= \sum \alpha_i Q_{\lambda_1}\zeta\\ &= Q_{\lambda_1}\zeta.\end{aligned}$$

Since $Q_\lambda\zeta$ is an element of the closed convex hull of $\{u_x\zeta | x \in G\}$, then $Q_\lambda\zeta$ or $Q_{\lambda_1}Q_\lambda = Q_\lambda$.

Now

$$\begin{aligned}(Q_{\lambda_1}Q_\lambda\zeta, \kappa) &= \lambda_1(u_x Q_\lambda\zeta, \kappa)\\ &= (Q_\lambda\zeta, \kappa)\end{aligned}$$

so that $Q_\lambda > 0$ and $Q_{\lambda_1} = Q_\lambda$.

Now we prove the second part of the theorem. Let E be the projection described in the theorem. Then

$$\begin{aligned}(Q\zeta, E\kappa) &= \lambda[u_x\zeta, E\kappa]\\ &= \lambda[Eu_x\zeta, \kappa]\\ &= \lambda[u_x E\zeta, \kappa]\\ &= (E\zeta, \kappa)\end{aligned}$$

so $EQ = E$. Now let us check that $EQ = Q$. This will show that $E = Q$.

For all $x \in G$ we have $u_x Q\zeta = Q\zeta$. $Q\zeta$ fixed under u_x implies $Q\zeta$ is an element of the range of E. So $EQ\zeta = Q\zeta$ and therefore $EQ = Q$.

THEOREM 7. *There exists a positive self-adjoint linear map Φ of $G_{h,s}$ into $G_{h,s}$ such that*:

(i) $(Au_x\zeta, u_x\kappa) \to^{(a)} (\Phi(A)\zeta, \kappa)$ *for all* $A \in G_{h,s}$,

(ii) $\Phi(A)u_x = u_x\Phi(A)$,

(iii) $\Phi \geqq 0$ *and is self-adjoint, that is,* $A^* = A, \Phi(A) = \Phi(A^*)$.

Let us explain some of the terminology and conditions involved. First, u_x is assumed to be a weakly continuous isometric representation of G.

$G_{h,s}$ represents the so-called Hilbert-Schmidt class of operators, i.e., $A \in G_{h,s}$ means $A : h \to h$ and if e_α is a complete orthonormal basis for h, then

$$\sum_\alpha \| Ae_\alpha \|^2 < \infty.$$

Here $\sum_\alpha$ means the supremum of the set of all finite sums. It can be shown that $G_{h,s}$ is closed under * (adjoint) and is an ideal in $L(h)$. $G_{h,s}$ is also a Hilbert space in its own right under the inner product

$$[A, B] = \sum_\alpha (Ae_\alpha , Be_\alpha)$$

(here $\sum_\alpha$ makes sense since only a countable number of terms are not equal to 0). Consider $G_{h,s}$ as a Hilbert space (with $[\ ,\ ]$ as an inner product). Consider the map:

$$\begin{aligned} x \to u_x^* A u_x &= T_x(A), \quad A \in G_{h,s}, \\ \| T_x(A) \|^2 &= [u_x^* A u_x , u_x^* A u_x] \\ &= \sum_\alpha \| u_x^* A u_x e_\alpha \|^2 \\ &= \sum_\alpha \| A u_x e_\alpha \|^2 \\ &\leqq \| \sum_\alpha A e_\alpha \|^2 \\ &= [A, A]. \end{aligned}$$

So T_x is an operator on $G_{h,s}$ of norm less than or equal to 1. Now

$$\begin{aligned} T_{xy}(A) &= [(u_{xy})^* A u_{xy}] \\ &= u_y^* (u_x^* A u_x) u_y \\ &= T_y T_x(A). \end{aligned}$$

Hence T_α is a bounded anti-representation of G on $G_{h,s}$.

CLAIM. This anti-representation is weakly continuous.

It suffices to show that $x \to [T_x A, B]$ is continuous for A, and that B is an element of a dense subspace of $G_{h,s}$, i.e., for finite rank. So let

$$\begin{aligned} A\zeta &= (\zeta, \sigma)\sigma, \\ B\zeta &= (\zeta\tau)\tau', \\ \sigma, \sigma', \quad \tau, \tau' &\in h, \\ [T_x A, B] &= \sum_\alpha (u_x^* A u_x e_\alpha , Be_\alpha) \\ &= \sum_\alpha ((u_x e_\alpha , \sigma) u_x^* \sigma', (e_\alpha , \tau)\tau'). \end{aligned}$$

CLAIM. This series converges uniformly in x.

Let $\epsilon > 0$. There exists a finite set, call it f, such that

$$\Big| \sum_{\alpha \notin f} (u_x e_\alpha , \sigma)(\overline{e_\alpha , \tau})(u_x{}^* \sigma', \tau') \Big|^2$$

$$\leqq | (u_x{}^* \sigma', \tau') |^2 \sum_{\alpha \notin f} | (u_x e_\alpha , \sigma) |^2 \sum_{\alpha \notin f} | (\tau_\zeta e_\alpha) |^2,$$

$$| (u_x{}^* \sigma', \tau') |^2 \leqq \| \sigma' \|^2 \| \tau' \|^2,$$

$$\sum_{\alpha \notin f} | (u_x e_\alpha , \sigma) |^2 \leqq \| \sigma \|^2,$$

$$\sum_{\alpha \notin f} | (\tau_\zeta e_\sigma) |^2 \quad \text{can be rendered} \quad \leqq \epsilon / \| \sigma \|^2 \| \sigma' \|^2 \| \tau' \|^2.$$

Given any $\epsilon > 0$, such an f can be chosen. This proves the claim. $[T_x A, B]$ is hence weakly continuous as a uniform limit. By the previous theorem, there exists a bounded operator Φ such that

$$[T_x A, B] \to^{(a)} [\Phi(A), B].$$

Choose B of the form

$$B\rho = (\rho, \zeta)\kappa,$$

$$\| \zeta \| = 1.$$

Imbed ζ in an orthonormal basis with $\zeta = e_{\alpha_0}$:

$$[\Phi(A), B] = \sum_\alpha (\Phi(A) e_\alpha , B e_\alpha)$$

$$= (\Phi(A)\zeta, \kappa).$$

Likewise,

$$[T_x A, B] = (u_x{}^* A u_x \zeta, \kappa).$$

We have

$$(u_x{}^* A u_x \zeta, \kappa) \to^{(a)} (\Phi(A)\zeta, \kappa) \quad \text{for all} \quad \zeta, \eta.$$

This proves Theorem 7(i).

Now let λ be a right mean:

$$(\Phi(A)_y \zeta, u_y \kappa) = \lambda(u_x{}^* A u_x u_y \zeta, u_y \kappa)$$

$$= \lambda(u_x{}^* A u_x \zeta, \kappa)$$

$$= (\Phi(A)\zeta, \kappa)$$

so that

$$u_y{}^* \Phi(A) u_y = \Phi(A).$$

This proves Theorem 7(ii).

Now let $A \geqq 0$. Then

$$(Au_x\zeta, u_x\zeta) \geqq 0 \quad \text{for all} \quad x, \zeta,$$
$$\lambda(u_x^* A u_x\zeta, \zeta) \geqq 0,$$

so that

$$(\Phi(A)\zeta, \zeta) \geqq 0.$$

Thus Φ is a positive map, and Theorem 7(iii) is proved.

3.3 Peter-Weyl theory on amenable semigroups. 1. In this section u_x will be assumed to be an isometric, weakly continuous representation of G in $L(h)$. We shall state the essential results. The representation u_x will be said to be reducible over h if there exists a subspace m of h (nontrivial) such that $u_x(m) \leqq m$ and $u_x(m^\perp) \leqq m^\perp$ for all $x \in G$.

THEOREM 8. *Let* $\kappa = \{\zeta \in h | \Phi(A)\zeta = 0 \text{ for all } A \in G_{h,s}\}$ (Φ *as defined above*). *Then the following statements hold*:

(i) κ *reduces the representation*;

(ii) $u_x \mid_\kappa$ (*restricted to* κ) *has no finite-dimensional subrepresentation*;

(iii) $u_x \mid_{\kappa^\perp}$ *is the direct sum of finite-dimensional subrepresentations.*

This theorem may be regarded as a sort of localization theorem. It states how u_x behaves on a certain subspace of h, and, namely, u_x is a direct sum of finite-dimensional subrepresentations on κ. The theorem ties κ and the Hilbert-Schmidt class of operators.

In the proof we shall make use of the following elementary fact: Let m be a proper closed subspace of the Hilbert space h; then the representation reduces m if an only if u_x for each x commutes with P which is the associated projection on m. By Zorn, let P_α be the maximal collection of nonzero, finite-dimensional, orthogonal projections each commuting with all u_x. (This family could be empty.)

Let $P = \sum_\alpha P\alpha$. Then P commutes with all u_x. Let M be the space on which $I - P$ projects.

CLAIM. u_x restricted to M has no finite-dimensional subrepresentation.

In fact, let $Q \neq 0$ be a finite-dimensional projection which commutes with all u_x.

CLAIM. $Q \leqq P$.

If not $Q(I - P) \neq 0$ so $(I - P)Q(I - P) \neq 0$ and is an element of $G_{h,s}$ since the latter is a two-sided ideal and any operation of finite rank is an element of $G_{h,s}$. This operation is self-adjoint and compact, so by the spectral representation,

$$(I - P)A(I - P) = \sum_{\lambda \neq 0} \lambda P_\lambda,$$

where the P_λ are finite-dimensional, orthogonal projections. It follows that each P_λ commutes with u_x and at least on $P_\lambda \neq 0$; also $P_\lambda \leqq I - P$. This contradicts the maximality of $\{P_\alpha\}$.

Now the claim is that range $(I - P) = \kappa$. The theorem then will be established. Suppose ζ is an element of the range $(I - P)$ and $\Phi(A)\zeta \neq 0$ for some $A \in G_{h,s}$. Without loss of generality we may assume that A is self-adjoint. We have seen that $\Phi(A)$ commutes with all u_x and is a nonzero, self-adjoint, compact operator (compact because it is the uniform limit of operators of finite rank). Hence,

$$\Phi(A) = \sum \lambda P_\lambda ,$$

where for some $\lambda \neq 0$, $P_\lambda\zeta \neq 0$. As before P_λ commutes with all u_x so $P_\lambda u_x$ is a finite-dimensional subrepresentation of u_x.

Now range $P_\lambda \leqq$ range P (by maximality), $P\zeta \neq 0$ since $P_\lambda \leqq P$, $(I - P)\zeta \neq \zeta$, and so $\zeta \notin$ range $(I - P)$, which is a contradiction.

Conversely, let $\zeta \notin$ range $(I - P)$; then there exists α such that $P_\alpha\zeta \neq 0$. Since P_α is finite-dimensional, then $P_\alpha \in G_{h,s}$ and P_α commutes with all u_x. Thus

$$\begin{aligned}(\Phi(P_\alpha)\zeta, \eta) &= \lambda(P_\alpha u_x\zeta, u_x\eta)\\ &= (u_x{}^*P_\alpha u_x\zeta, \eta)\\ &= \lambda(P_\alpha\zeta, \eta)\\ &= (P_\alpha\zeta, \eta)\end{aligned}$$

and so

$$\Phi(P_\alpha) = P_\alpha .$$

Therefore,

$$\Phi(P_\alpha)\zeta = P_\alpha\zeta \neq 0$$

and $\zeta \notin m$.

In the process we have proved the following theorems.

THEOREM 9. *Let $\eta = \kappa^\perp$. Then $\eta = I - P$, where P is the supremum of all projections (finite-dimensional) which commute with all u_x as $x \in G$.*

THEOREM 10. *If $m_\lambda = \{\zeta \in h : \lambda \mid (u_x\zeta, t) \mid = 0$ for all $t \in h\}$, then $\eta = m_\lambda$.*

Say $\zeta \in \eta$; then $(\Phi(A)\zeta, t) = 0 = \lambda(Au_x\zeta, u_xt)$ for all $A \in G_{h,s}$. Take $A\zeta = (\zeta, \rho)\sigma$ and $0 = \lambda[(u_x\zeta, \rho)(\overline{u_xt, \sigma})]$. Put $t = \zeta$, $\rho = \sigma$. Then $\lambda(u_x\zeta, \rho)^2 = 0$ for all ρ, and

$$\lambda \mid (u_x\zeta, \rho)\cdot 1 \mid \leqq \sqrt{\lambda \mid (u_x\zeta, \rho) \mid^2}\sqrt{\lambda(1)^2} = 0.$$

Thus $\zeta \in m_\lambda$.

Now we use $\zeta \in m_\lambda$ to show $\zeta \in \eta$, and

$$\lambda \mid (u_x\zeta, t) \mid = 0 \quad \text{for all} \quad t,$$

$$\mid \lambda(u_x\zeta, t)(u_xt, \sigma) \mid \leqq \lambda \mid (u_x\zeta, t) \mid \mid (u_xt, \sigma) \mid$$

$$\leqq \| t \| \| \sigma \| \lambda \mid (u_x\zeta, t) \mid = 0.$$

If $A\zeta = (\zeta, \sigma)\sigma$, then $(\Phi(A)\zeta, t) = 0$ and so $\Phi(A)\zeta = 0$.

Hence $\Phi(A)\zeta = 0$ for any operator of finite rank; hence, because the latter set is dense in $G_{h,s}$, $\Phi(A)\zeta = 0$ for all $A \in G_{h,s}$.

In the case when G is a compact group, $\eta = (0)$ and we have the statement of the Peter-Weyl theorem for compact groups. A representation is called unitary if u_x is unitary for all $x \in G$ (i.e., $u_x^* u_x = I$).

Now let us consider the situation where the tape semigroup has different representations in different Hilbert spaces. The following theorem will tell us how this affects the representations.

THEOREM (Schur). *Let u_x and v_x be continuous, finite-dimensional, irreducible, unitary representations of an amenable semigroup on Hilbert spaces $L(h_u)$ and $L(h_r)$. If T is any Hilbert-Schmidt operator from h_u to h_r, then there exists Φ which is a linear operator on the space of Hilbert-Schmidt operators from h_u to h_r such that*

$$\Phi(T)u_x = v_x\Phi(T),$$

$$(Tu_x\zeta, \ v_x\eta) \to^{(a)} (\Phi(T)\zeta, \ \eta) \quad \textit{for all} \quad \zeta \in h_u \textit{ and all } \eta \in h_r.$$

Moreover, if u_x is a representation nonequivalent to v_x, then $\Phi(T) = 0$. If $u_x = v_x$, then $\Phi(T) = (1/n)\ (\mathrm{tr}\ T)\ I$, where $n = \dim h_u$.

2. Let us define now the term coordinate. A *coordinate* (respective to a given representation) is a function belonging to $C(G)$ of the form (u_xh'', h'), where h'' and h' are fixed vectors of h.

THEOREM 11. *Coordinates of nonequivalent representations are orthogonal respectively to any mean, i.e.,*

$$(u_xh, h'')\ \overline{(v_xk, k')} \to^{(a)} 0.$$

Now consider h_u, a finite-dimensional vector space. Then let $e_1, \cdots, e_n$ be an orthogonal basis for h_u. Then we have the following theorem.

THEOREM 12. *(u_xe_j, e_i), $1 \leqq i, j \leqq n$, are orthogonal respectively to any mean.*

4. Conclusion. We have now decomposition and orthogonality theorems for certain types of automata, essentially for those where the underlying semigroup of inputs is an amenable group. The theorems are algebraic as well as topological in their nature. A possible way to look further into the matter is to take specific examples of groups with a known mean and see

what are then the results of applying these theorems. The idea is to break down these types of machines into easier parts. Another direction open for investigation is to explore topological groups, not necessarily amenable, but having some other special conditions, and consider their decompositions.

The theorems on bounded isometric representations essentially say that an infinite machine acts like a finite one when restricted to certain subsets of the set of states. The theorems give a characterization of the zone in which the machines seem to act as finite ones. The orthogonality theorems essentially give averaging processes on automata where a mean can be constructed on the semigroups of inputs.

REFERENCES

[1] MICHAEL A. ARBIB, PETER FALB AND R. KALMAN, *Topics in Mathematical System Theory*, McGraw-Hill, New York, to appear.

[2] A. C. FLECK, *Preservation of structure by certain classes of functions on automata and related group theoretic properties*, Computer Laboratory, Michigan State University, East Lansing, 1961.

[3] J. HARTMANIS AND R. E. STEARNS, *Algebraic Structure Theory of Sequential Machines*, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.

[4] KARL HEINRICH HOFMANN AND PAUL S. MOSTERT, *Elements of Compact Semigroups*, Charles E. Merrill, Columbus, Ohio, 1966.

[5] KENNETH KROHN AND JOHN RHODES, *Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines*, Trans. Amer. Math. Soc., 116 (1965), pp. 450–464.

[6] KENNETH KROHN, RICHARD MATEOSIAN AND JOHN RHODES, *Methods of the algebraic theory of machines; decomposition theorem for generalized machines; properties preserved under series and parallel composition of machines*, J. Computer and System Science, 6 (1967), pp. 55–85.

[7] L. H. LOOMIS, *An Introduction to Abstract Harmonic Analysis*, Van Nostrand, Princeton, 1953.

[8] G. MEZEI, *Structure of monoids with applications to automata*, Proc. Symposium on Mathematical Theory of Automata, vol. XII, John Wiley, New York, 1962, pp. 267–299.

[9] L. PONTRYAGIN, *Topological Groups*, Princeton University Press, Princeton, 1958.

[10] G. P. WEEG, *The group and semigroup associated with automata*, Proc. Symposium on Mathematical Theory of Automata, vol. XII, John Wiley, New York, 1962, pp. 257–266.

[11] PAUL H. ZEIGER, *Cascade synthesis of finite-state machines*, Proc. 6th Annual Symposium on Switching Circuit Theory and Logical Design, University of Michigan, Ann Arbor, 1965, pp. 45–51.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

OPTIMAL CONTROL OF DISTRIBUTED SYSTEMS (A SURVEY OF SOVIET PUBLICATIONS)*

A. G. BUTKOVSKY†, A. I. EGOROV‡ AND K. A. LURIE§

1. Introduction. Many existing industrial processes and control systems operate under conditions in which their potential working capabilities are not exploited. These systems need to be improved in such a way that their potential working capabilities are fully utilized. We call a system optimal when it is the best possible one under certain given working conditions. Often, practical limitations lead to the consideration of various constraints which manifest themselves as restrictions on the values of phase variables and control functions. Historically, the problems of optimal control for systems governed by ordinary differential equations of finite order emerged from the aspiration to take various constraints into account [1]–[9]. The main technique for solving this class of problem has been based upon a special result of the calculus of variations known as Pontryagin's maximum principle [10] and upon the method of dynamic programming due to Bellman.

In physical situations, one often encounters systems whose parameters are distributed in both space and time. The dynamic behavior of these systems is governed by partial differential equations, integral equations, integrodifferential equations and sometimes by more general functional equations.

Sometimes there exist situations where a system is described by ordinary differential equations of which the order is very high. Under certain assumptions, the system description may be considerably simplified if we approximate the system by expressing it as one with distributed parameters. For example, the dynamic behavior of a very large number of subsequently combined aperiodic lumped elements may be approximately described by the heat equation.

Optimal control problems for systems with distributed parameters frequently arise in mechanics, mathematical physics and engineering. Some examples may be found in [12]–[19]. In these works, different methods are developed, which enable one to solve some problems of practical im-

* Received by the editors June 15, 1967, and in revised form March 20, 1968.

† Institute of Automation and Remote Control, Academy of Sciences, Moscow B-53, USSR.

‡ Institute of Automation, Kirghiz Academy of Science, Frunze, Kirghiz SSR, USSR.

§ A. F. Ioffe Physico-Technical Institute, Academy of Sciences, Politekhnicheskaya ul. 26, Leningrad K-21, USSR.

portance. In [20] the problem of optimal control of temperature distribution in solids is considered. An interesting result concerning necessary conditions for extremal problems in a function space with inequality constraints is discussed in [21].

The difficulty in the formulation of these problems is that in general considerations, the formulation should be broad enough so as to retain the necessary generality, and on the other hand, it should be narrow enough so as to permit devising effective means for solving the problem.

It is far more difficult to develop theories and techniques for distributed optimal control problems than for their lumped equivalents. The difficulty is that, besides the highly complicated character of the governing equations, there are certain features in optimal control problems involving partial differential equations which are not found in those involving ordinary differential equations. In this paper, some of these features will be indicated.

Up to the present time, variational problems for distributed systems were considered mainly in connection with direct methods of the calculus of variations. Requirements of optimal control theory made it necessary to investigate indirect approaches to constrained variational problems when the number of independent variables exceeds one.

We shall now proceed to describe a few typical controlled objects which require distributed control for their operation.

1.1. Heating of metal in a furnace. (See [12] and [13].) Consider the problem of optimal heating of a metal pig in a furnace. The temperature of the fixed heating media can be characterized by a function $u(y, t)$ depending on the space variable y, $0 \leqq y \leqq L$, and time t, $0 \leqq t \leqq T$, where L is the length of furnace and T is the total heating time. The state of the heated material is characterized by its temperature $Q = Q(y, t)$. The material moving through the furnace with a velocity $v(t)$ (positive in the direction of the y-axis) is heated according to the equation

$$b\frac{\partial Q}{\partial t} + bv\frac{\partial Q}{\partial y} + Q - u = 0, \tag{1.1}$$

where b is the time constant associated with heating an elemental layer of material. This parameter generally depends on the difference $y - \int_0^t v(t)\,dt$. The temperature of the material at the end of the furnace obviously depends on the temperature variation during the course of heating. Also, the value of the final temperature depends to a great extent on the velocity $v(t)$ of the material through the heating zone. The final temperature of the material also depends on the thickness S of the material layer as well

as on its thermal and physical constants such as thermal conductivity, specific heat, density, etc. The control problem is to regulate the final temperature distribution of the material. Such regulation is required in the rolling of steel slabs. The velocity variations of the material during its motion through the furnace together with the thickness variations of the slabs usually act as disturbances. The purpose of optimal control is to regulate the temperature distribution along the furnace in such a way so as to guarantee minimal (in some appropriate sense) deviation of the final temperature of the material from the one prescribed. For example, one may wish to minimize the integral

$$\int_0^T |\, Q^* - Q(L, t)\,|^\gamma \, dt, \qquad \gamma \geqq 1, \tag{1.2}$$

where T, γ and Q^* are given.

We shall consider only the following typical disturbances:

(a) disturbance caused by the time dependence of velocity v;

(b) disturbance caused by variations in the parameter b:

$$b = b\left(y - \int_0^t v(\tau)\, d\tau\right) = \begin{cases} b_1 & \text{if} \quad L \geqq y - \int_0^t v(\tau)\, d\tau \geqq \frac{1}{2} L, \\ b_2 & \text{if} \quad 0 \leqq y - \int_0^t v(\tau)\, d\tau < \frac{1}{2} L, \end{cases}$$

where b_1 and b_2 are known constants.

The latter situation corresponds to the process of heating two different groups of slabs following each other. Usually, the control function u does not depend on the space variable y. Also u satisfies an amplitude constraint of the form

$$A_1 \leqq u \leqq A_2 .$$

One can easily show with the aid of the maximum principle [12] that the optimal control function for case (a) is given by ($\gamma = 2$)

$$u = u(t) = Q^* + bv(t) \left.\frac{\partial Q}{\partial y}\right|_{y=L},$$

and for case (b), the optimal control is a discontinuous function given by

$$u = u(t) = \begin{cases} A_1 & \text{if} \quad b_1 > b_2 , \\ A_2 & \text{if} \quad b_1 < b_2 . \end{cases}$$

In the general situation, the optimal control consists of a mixture of the above controls. The transition from a control for case (a) to that for (b) is determined by the distance l between the point of discontinuity of the coefficient b and the end of the furnace, $y = L$.

The control problems associated with many other industrial processes such as multizonal furnaces for rapid heating, kilning and drying stoves and different kinds of heat exchangers can be posed in a similar manner.

1.2. Drying process. (See [27].) The process of drying moist material whose characteristic width is S, $0 \leqq x \leqq S$, in a drying stove of length L, $0 \leqq y \leqq L$, is described by the following system of equations:

$$\frac{\partial Q_1}{\partial t} = a_1 \frac{\partial^2 Q_1}{\partial x^2} + b_1 v \frac{\partial Q_1}{\partial y} + \beta \frac{\partial Q_2}{\partial t}, \tag{1.3}$$

$$\frac{\partial Q_2}{\partial t} = a_2 \frac{\partial^2 Q_2}{\partial x^2} + a_3 \frac{\partial^2 Q_1}{\partial x^2} + b_2 v \frac{\partial Q_2}{\partial y}. \tag{1.4}$$

The state of this process at time t can be characterized by the temperature distribution $Q_1(x, y, t)$ and the moisture concentration $Q_2(x, y, t)$. To complete the description of the process, (1.3) and (1.4) are supplemented by boundary conditions:

$$Q_1(x, 0, t) = Q_{\Gamma 1}(x, t), \tag{1.5}$$

$$Q_2(x, 0, t) = Q_{\Gamma 2}(x, t), \tag{1.6}$$

$$\lambda \frac{\partial Q_1}{\partial x}\bigg|_{x=S} = \alpha[u(y, t) - Q_1(S, y, t)], \tag{1.7}$$

$$\frac{\partial Q_1}{\partial x}\bigg|_{x=0} = 0, \tag{1.8}$$

$$\frac{\partial Q_2}{\partial x}\bigg|_{x=S} = \alpha_1 \frac{\partial Q_1}{\partial x}\bigg|_{x=S}, \tag{1.9}$$

$$\frac{\partial Q_2}{\partial x}\bigg|_{x=0} = 0, \tag{1.10}$$

and initial conditions

$$Q_1(x, y, 0) = Q_{01}(x, y), \tag{1.11}$$

$$Q_2(x, y, 0) = Q_{02}(x, y). \tag{1.12}$$

In these equations, the function $v(t) \geqq 0$ represents the velocity of the material in the direction of the positive y-axis, the functions $b_1 = b_1(y, t)$ and $b_2 = b_2(y, t)$ and constant coefficients a_1, a_2, a_3, α, α_1, β characterizing the thermal and diffusive properties of the material. The functions $Q_{\Gamma 1}(x, t)$, $Q_{\Gamma 2}(x, t)$, $Q_{01}(x, t)$ and $Q_{02}(x, y)$ are considered known.

The optimal control problem is to determine the control function $u(y, t)$ (temperature along the drying stove) to minimize (in some given sense) the influence of various kinds of disturbances, e.g., those corresponding to initial moisture concentration and porosity of the material, the velocity of material along the drying stove and so on.

Let $Q_2(y, t)$ denote the average moisture concentration over the cross section:

$$Q_2 = \frac{1}{S}\int_0^B Q_2(x, y, t)\, dx, \qquad 0 \leqq y \leqq L, \quad 0 \leqq t \leqq T. \tag{1.13}$$

One may wish to minimize the following performance index:

$$J = \int_0^T |\, Q^*(t) - Q_2(L, t)\,|^\gamma\, dt, \qquad \gamma \geqq 1, \tag{1.14}$$

where $Q^*(t)$ represents a prescribed function of the averaged concentration at the output end $y = L$ of the stove.

Usually, the following restrictions are added:

$$A_1 \leqq u(y, t) \leqq A_2, \quad \left|\frac{\partial u}{\partial y}\right| \leqq A_3, \quad Q_1(x, y, t) \leqq A_4, \tag{1.15}$$

where A_1, A_2, A_3, A_4 are known constants.

1.3. Chemical reactors. Consider now the optimal control of chemical reactors. It is required to obtain the maximum output of some component of the reaction product under certain restrictions on the reactant and certain parameters of the reaction process [27]. We shall examine a cyclic reactor with a fixed layer of catalyst. As the reactant is fed into the reactor, the reaction starts and accelerates under the action of the catalyst. The speed of reaction is temperature-dependent. The reaction, being endothermic in nature, tends to lower the temperature of the catalyst layer and thus slows down the reaction. The reaction stops completely when the temperature drops to a low critical value. Subsequently, the temperature of the catalyst layer increases and the process starts again. The highest value of the layer's temperature is restricted by the firmness of the catalyst layer and by some other undesirable reactions at high temperatures.

During the reaction process, one can only influence the layer's temperature by regulating the temperature of the reactant at the input of the reactor. In what follows, we shall consider a reaction of the first order. This means that its velocity depends only on the temperature Q_1 and not on the concentration Q_2. Thus, at any fixed point x, we have

$$\frac{\partial Q_2}{\partial t} = \exp\left(\alpha_0 - \frac{\beta_0}{Q_1}\right), \qquad 0 \leqq x \leqq L, \quad 0 \leqq t \leqq T, \tag{1.16}$$

where $Q_1 = Q_1(x, t)$ denotes the temperature at the point x and time t.

When the deviations of parameters from their average values are small, we can approximate (1.16) by the following relation:

$$\frac{\partial Q_2}{\partial t} = \exp(\alpha + \beta Q_1). \tag{1.17}$$

Suppose that the heat conduction between the layer of catalyst and the material is so intensive that both reagents are at the same temperature and the transmission of heat through the layer is mostly convective. Under these assumptions, one can use the first order equation

$$\frac{\partial Q_1}{\partial t} + v\frac{\partial Q_1}{\partial x} = -h \exp(\alpha + \beta Q_1) \tag{1.18}$$

to describe the process. Here, h characterizes the thermal effect of the reaction; v is a parameter depending on the velocity of material through the reactor and on the ratio of the reactant's specific heat to that of the catalyst's layer.

It is required to regulate the reactant temperature at the input in such a way as to maximize the extent of reactant conversion. This extent can be determined by the expression

$$\int_0^T \exp(x + \beta Q_1(t, x))\, dx. \tag{1.19}$$

In many cases, however, the optimality criterion must be of economic origin. Taking $R(Q_1(t, 0))$ to be the cost function associated with achieving temperature $Q_1(t, 0)$, one may wish to maximize the functional

$$\int_0^T \left[\int_0^L \exp(\alpha + \beta Q_1(t, x))\, dx - R(Q_1(t, 0))\right] dt, \tag{1.20}$$

where $R(Q_1)$ is an increasing function of its argument. Here, Q_1 is regarded as the control variable.

1.4. Metal processing. In metallurgy, great importance is attached to the rapid and high quality heating of metals in heating wells and furnaces.

The most frequent situations where optimal control over the heating process is necessary are the following (see [12]–[15], [27]).

1. The amount of work done by a rolling mill (or press, or sledge) depends to a large extent on the working temperature of the heating chamber.

In this case, it is necessary to optimize the heating process in such a way as to minimize the total time in heating the material up to the desired temperature distribution. This leads to the problem of time-optimal heating.

2. If the total heating time is fixed (e.g., when the furnace temperature

is determined by that in the rolling mill), then it is required to organize the heating process so as to obtain the best possible quality of heating over the given time interval.

An example of such a problem is as follows. Assume that the temperature distribution $Q(x, t)$, $-S \leqq x \leqq S$, $0 \leqq t \leqq T$, can be adequately described by the simple heat equation

$$\frac{\partial Q}{\partial t} = a \frac{\partial^2 Q}{\partial x^2}, \tag{1.21}$$

where a is the thermal conductivity. The boundary and initial conditions for this problem have the following form:

$$\lambda \frac{\partial Q}{\partial x}\bigg|_{x=S} = \alpha_1[u_1(t) - Q(S, t)], \tag{1.22}$$

$$\lambda \frac{\partial Q}{\partial x}\bigg|_{x=-S} = \alpha_2[u_2(t) - Q(-S, t)], \tag{1.23}$$

$$Q(x, 0) = Q_0(x), \tag{1.24}$$

where λ is the heat conduction coefficient; α_1 and α_2 are coefficients of heat exchange between the heating media and the material; Q_0 is the initial temperature distribution; and u_1, u_2 are the control variables corresponding to the temperatures of the heating media. It is assumed that u_1 and u_2 satisfy the following inequalities:

$$A_1 \leqq u_1(t) \leqq A_2, \tag{1.25}$$

$$A_3 \leqq u_2(t) \leqq A_4. \tag{1.26}$$

Also, an additional restriction may be imposed on the temperature gradient inside the material:

$$\left|\frac{\partial Q}{\partial x}\right| \leqq A_5. \tag{1.27}$$

Finally, a restriction is imposed on the total heating time T:

$$T \leqq A_6. \tag{1.28}$$

The parameters $A_1, \cdots, A_6$ are given constants. The control problem is to determine a control function (temperature of the heating medium) which minimizes the deviation (in some prescribed sense) of the material temperature distribution from a desired one at a fixed time T. In particular, find the control functions $u_1(t)$, $u_2(t)$, $0 \leqq t \leqq T$, so as to minimize the functional

$$J = \int_{-S}^{S} |Q^*(x) - Q(x, T)|^\gamma \, dx, \qquad \gamma \geqq 1, \tag{1.29}$$

under side conditions (1.25)–(1.28).

The solution of (1.21) corresponding to the zero initial condition is given by

$$(1.30) \qquad Q(x, t) = \int_0^t K(x, t - \tau)u(\tau)\, d\tau, \quad -S \leqq x \leqq S, \quad 0 \leqq t \leqq T,$$

where $K(x, t)$ is the Green's function. One can show [16] that the optimal control function $u(t)$, $A_1 \leqq u(t) \leqq A_2$, satisfies the following integral equation:

$$(1.31) \qquad u(t) = \frac{A_1 + A_2}{2} - \frac{A_1 - A_2}{2} \operatorname{sgn}\left[B(t) - \int_0^T N(t, \tau)u(\tau)\, d\tau\right],$$

where

$$B(t) = \int_{-S}^{S} Q^*(x)K(x, T - t)\, dx,$$

$$N(t, \tau) = \int_{-S}^{S} K(x, T - \tau)K(x, T - t)\, dx.$$

This equation may be solved using approximate methods.

To conclude our description of optimal control problems for distributed systems, let us describe a typical control problem in magnetohydrodynamics [18], [19].

1.5. Magnetohydrodynamic systems. Consider the rectilinear motion ($\mathbf{v} = (V(y), 0, 0)$) of a conducting fluid along a plane channel of width 2δ. The walls of the channel are insulated everywhere except in a region of length 2λ where two ideally conducting electrodes are placed opposite each other on different walls. The electrodes are connected to an external load R.

Upon imposing a transverse magnetic field $\mathbf{B} = -\mathbf{i}_3 B(x)$ onto the moving fluid, an electric current of density $\mathbf{j} = (\zeta^1, \zeta^2)$ is induced inside the channel. The total current flowing through the external load is given by

$$(1.32) \qquad I = \int_{-\lambda}^{\lambda} \zeta^2(x, \pm\delta)\, dx.$$

If the magnetic Reynolds number Re_m is small compared with unity, we may neglect the induced magnetic field. If, in addition, the mhd parameter of interaction is also small, it is possible to neglect the Lorentz force so that the fluid motion in the channel can be approximately described by purely hydrodynamic equations.

Let

$$(1.33) \qquad \mathbf{j} = -\operatorname{curl} \mathbf{i}_3 z^2, \quad j_x = \zeta^1, \quad j_y = \zeta^2.$$

The electric potential z^1 and current density $\mathbf{j}$ can be determined from the following equations:

$$\begin{gathered}\frac{\partial z^1}{\partial x} = -\rho\zeta^1, \quad \frac{\partial z^1}{\partial y} = -\rho\zeta^2 + \frac{VB}{c}, \quad \frac{\partial}{\partial x}\left(\frac{VB}{c} - \rho\zeta^2\right) + \frac{\partial \rho\zeta^1}{\partial y} = 0, \\ \frac{\partial z^2}{\partial x} = \zeta^2, \qquad \frac{\partial z^2}{\partial y} = -\zeta^1, \qquad \frac{\partial \zeta^1}{\partial x} + \frac{\partial \zeta^2}{\partial y} = 0, \\ I = z^2(\lambda, \pm\delta) - z^2(-\lambda, \pm\delta),\end{gathered} \tag{1.34}$$

where $\rho(x, y)$ is the specific resistivity of the fluid satisfying

$$\rho_{\min} \leqq \rho(x, y) \leqq \rho_{\max}. \tag{1.35}$$

The limits $\rho_{\max}$ and $\rho_{\min}$ are known constants corresponding, respectively, to the resistivity of the fluid in the absence of external ionization and that in the presence of total ionization.

Upon eliminating variables ζ^1 and ζ^2, (1.34) can be rewritten as a set of second order equations for z^1 and z^2:

$$\begin{gathered}\frac{\partial}{\partial x}\frac{1}{\rho}\frac{\partial z^1}{\partial x} + \frac{\partial}{\partial y}\frac{1}{\rho}\frac{\partial z^1}{\partial y} = \frac{1}{c}\frac{\partial}{\partial y}\frac{VB}{\rho}, \\ \frac{\partial}{\partial x}\rho\frac{\partial z^2}{\partial x} + \frac{\partial}{\partial y}\rho\frac{\partial z^2}{\partial y} = \frac{1}{c}\frac{\partial}{\partial x}VB.\end{gathered} \tag{1.36}$$

The boundary conditions can be established by considering the properties of the channel walls and by assuming that both components ζ^1, ζ^2 of the current density vanish at infinity.

The basic optimum control problem is to choose a control $\rho(x, y)$ from the class of piecewise continuous functions of two variables satisfying inequalities (1.35) such that the functional I (see (1.32)) takes on its maximum value.

The foregoing examples provide some motivation for the development of certain analytical approaches to control problems for distributed systems. Some of these approaches will be discussed in the sequel. The discussions are classified according to their extent of generality and also the particular forms of the basic equations in the problems. We shall first consider the case where these equations are of integral type.

2. Optimal control of systems governed by integral equations. Let us characterize the state of an object by a vector-valued function $q = Q(P) = (Q_1(P), \cdots, Q_n(P))$ defined on some domain D of an m-dimensional Euclidean space E_m, where $P \in D$. The admissible controls are vector-valued functions $u = U(P) = (U_1(P), \cdots, U_r(P))$ defined on D and taking values in a closed domain $\Omega_r \subset E_r$. The control $U(P)$ is related to

$Q(P)$ by the following equations:

$$\phi^i\left[Q(P), \int_D K(P, S, Q(S), U(S))\, dS\right] = 0, \qquad i = 1, \cdots, n, \tag{2.1}$$

where $\phi(q, y)$ and $K(P, S, q, u)$ are prescribed vector-valued functions of the arguments $P \in E_m$, $S \in E_m$, $q \in E_n$, $u \in \Omega_r$.

The system performance is measured by a specified functional of the state and control functions. The system is said to be optimal if this functional achieves its extremal value. Without loss of generality, we shall consider only the minimization problem. Let the functional in question be of the form:

$$\phi^0\left[\int_D F(S, Q(S), U(S))\, dS\right], \tag{2.2}$$

where $\phi^0(y)$ and $F(S, q, u)$ are specified functions of their arguments. Now, the optimal control problem can be formulated as follows: Determine an admissible control function $U(P)$ which minimizes the functional (2.2) under side conditions (2.1). This problem is discussed in [22]–[24]. The basic result is a necessary condition for minimum, which may be stated in the form of the following maximum principle.

THEOREM 1 (Maximum principle). *Let* $u_0 = u_0(P)$, $P \in D$, $u_0 \in \Omega_r$, *represent the control function and* $q_0 = Q_0(P)$ *the corresponding state function* (see (2.1)). *In order for the function* $u_0(P)$ *to be optimal, there must exist a function* $N(P)$, $P \in D$, *satisfying the following integral equation* (*linear relative to* $N(P)$):

$$\begin{aligned} -\phi_y^{0\prime}\left[\int_D F(S, Q_0(S), u_0(S))\, dS\right] \frac{\partial F(R, Q_0(R), u_0(R))}{\partial q} \\ + \phi_y'\left[Q_0(R), \int_D K(R, S, Q_0(S), u_0(S))\, dS\right] N(R) \\ + \int_D \phi_y'\left[Q_0(P), \int_D K(P, S, Q_0(S), u_0(S))\, dS\right] \\ \cdot \frac{\partial K(P, R, Q_0(R), u_0(R))}{\partial q} N(P)\, dP = 0 \end{aligned}$$

such that the function

$$\begin{aligned} \Pi(R, u) = -\phi_y^{0\prime}\left[\int_D F(S, Q_0(S), u_0(S))\, dS\right] F(R, Q_0(R), u) \\ + \int_D \phi_y'\left[Q_0(P), \int_D K(P, S, Q_0(S), u_0(S))\, dS\right] \\ \cdot K(P, R, Q_0(R), u) N(P)\, dP \end{aligned}$$

achieves its maximum with respect to the variable $u \in \Omega_r$ *at* $u = u_0(R)$ *for almost all* $R \in D$. *In other words, the following relationship*

$$\max_{u \in \Omega_r} \Pi(R, u) = \Pi(R, u_0(R))$$

holds almost everywhere in D.

If $Q(P)$ and $U(P)$ enter into (2.1) linearly, then the condition given by the maximum principle is not only necessary, but also sufficient for the optimality of $u_0(P)$ and $Q_0(P)$.

2.1. Example 1. Consider again the problem of optimal heating as described in §1 under metal processing. Here, it is assumed that the heat is supplied at one end of the bar by thermal radiation. For this case, the boundary conditions for (1.21) have the form

$$\begin{aligned} \lambda \frac{\partial Q}{\partial x}\Big|_{x=S} &= C\{[u(t)]^4 - [Q(S, t)]^4\}, \\ \frac{\partial Q}{\partial x}\Big|_{x=-S} &= 0. \end{aligned} \tag{2.3}$$

The temperature distribution inside the material is related to that on its surface through the following relationship:

$$Q(x, t) = \int_0^t K(x, t, \tau) Q(S, \tau)\, d\tau. \tag{2.4}$$

Equation (2.4) together with boundary condition (2.3) leads to a nonlinear integral equation for $Q(S, t)$:

$$\lambda \int_0^t K_x'(x, t, \tau)\,|_{x=S}\, Q(S, \tau)\, d\tau = C\{[u(t)]^4 - [Q(S, t)]^4\}, \quad 0 \leqq t \leqq T. \tag{2.5}$$

One may now seek the optimal control function $u(t)$, $0 \leqq t \leqq T$, which maximizes the functional (1.29) under side conditions (1.25)–(1.28).

Equation (1.30) may be applied to a broad class of distributed systems with a single control function. We are going to investigate the time-optimal control problem for this class of systems, namely, that of minimizing time T at which

$$Q(x, T) = Q^*(x) = \int_0^T K(x, T - t) u(t)\, dt, \tag{2.6}$$

where $|\, u(t)\,| \leqq L$ and Q^* is a specified function.

2.2. Application of L-problem of moments. The foregoing problem can be reduced to a Stieltjes L-problem of moments [24], [25]. Take any complete orthonormal set of functions $\varphi_i(x)$ $(i = 1, 2, \cdots, 0 \leqq x \leqq S)$

and express $Q^*(x)$ and $K(x, t)$ in terms of $\varphi_i(x)$ for every fixed t. Equation (2.6) will then be reduced to

$$\sum_{i=1}^{\infty} \alpha_i \varphi_i(x) = \sum_{i=1}^{\infty} \varphi_i(x) \int_0^T g_i(T - t)u(t)\, dt. \tag{2.7}$$

Let us now assume that the functions $g_k(t)$ $(k = 1, 2, \cdots, n; 0 \leqq t \leqq T)$ are chosen so that $\sum_{k=1}^{n} \Lambda_k g_k(t)$ (Λ_k, $k = 1, 2, \cdots$, not all zero) can vanish only at a finite number of points t in the interval $[0, T]$. Also, it is sufficient to assume that $g_k(t)$, $k = 1, 2, \cdots, n$, are linearly independent for arbitrary n.

The solution of (2.7) is equivalent to solving the following infinite system of integral equations:

$$\alpha_i = \int_0^T g_i(T - t)u(t)\, dt, \qquad i = 1, 2, \cdots. \tag{2.8}$$

Now, the optimal control problem can be restated as follows: Determine the control function $u(t)$, $|u(t)| \leqq L$, $0 \leqq t \leqq T$, satisfying (2.8) such that T takes on its minimum value.

It is known that a necessary and sufficient condition for the foregoing problem to be solvable is that, for any finite T,

$$\min_{\zeta_k^{(n)}} \int_0^T \left| \sum_{k=1}^{n} \zeta_k^{(n)} g_k(T_n - t) \right| dt = \lambda_n(T_n) = \frac{1}{L} \tag{2.9}$$

and

$$\sum_{k=1}^{n} \zeta_k^{(n)} \alpha_k = 1. \tag{2.10}$$

The optimal control function for this "reduced" problem has the following form:

$$u_n(t) = L \operatorname{sgn} \sum_{k=1}^{n} \xi_k^{(n)} g_k(T_n^0 - t), \tag{2.11}$$

where $\xi_k^{(n)}$ $(k = 1, 2, \cdots, n)$ is the solution of the problem, and the optimal time T_n^0 can be determined from the equation

$$\lambda_n(T_n^0) = \frac{1}{L}. \tag{2.12}$$

Suppose there exists a limiting function $\lambda(T) = \lim_{n\to\infty} \lambda_n(T)$ which is continuous for $0 \leqq T \leqq \infty$ (this condition does not hold in general), and $\lambda(T) \to \infty$ as $T \to \infty$. Then it is obvious that the equation

$$\lambda(T) = \frac{1}{L} \tag{2.13}$$

has a solution $T = T^0$ corresponding to the minimum transition time of the distributed system (2.6). Since the sequence of functions $u_n(T)$ given by (2.11) satisfy $|u_n(t)| \leqq L$, $0 \leqq t \leqq T_0$, $n = 1, 2, \cdots$, therefore the existence of the limiting control function $u^0(t)$ implies the existence of a solution to the original time-optimal control problem.

In many cases, one may treat the solution to the "reduced" problem corresponding to some finite n as an approximate solution to the original time-optimal control problem. To solve the moment problem for finite n, one may use the following method. Let us start from problem (2.9), (2.10); determine ζ_k^0, $k = 1, 2, \cdots, n$, and T^0 such that

$$\min_{\zeta_k} \int_0^T \left| \sum_{k=1}^n \zeta_k g_k(T - t) \right| dt = \int_0^T \left| \sum_{k=1}^n \zeta_k^0 g_k(T^0 - t) \right| dt = \frac{1}{L} \tag{2.14}$$

and

$$\sum_{k=1}^n \zeta_k^0 \alpha_k = 1. \tag{2.15}$$

We introduce the following notations:

$$\rho_\zeta(T) = \int_0^T \left| \sum_{k=1}^n \zeta_k g_k(T - t) \right| dt, \tag{2.16}$$

$$\zeta = (\zeta_1, \cdots, \zeta_n). \tag{2.17}$$

We take an arbitrary vector $\zeta = \zeta_{(0)}$ satisfying (2.15) and construct a graph of $\rho_{\zeta_{(0)}}(T)$ vs. T. Since $\rho_{\zeta_{(0)}}(T)$ increases monotonically from zero to infinity, there exists a T_0 such that $\rho_{\zeta_{(0)}}(T_0) = 1/L$. When the argument T is set equal to T_0, $\rho_\zeta(T_0)$ becomes a function of ζ only. We find the minimum of this function over ζ under side condition (2.15). The problem would have been solved at once if the minimum were achieved at the first step where $\zeta = \zeta_{(0)}$. However, such a vector is rarely guessed correctly. Instead, we generally come up with an intermediate vector $\zeta_{(1)}$ for which

$$\rho_{\zeta_{(1)}}(T_0) = \min_\zeta \rho_\zeta(T_0) < \frac{1}{L}.$$

For the next step of the iterative process, we take $\zeta = \zeta_{(1)}$ and find the value $T = T_1$ corresponding to the intersection between the graph $\rho_{\zeta_{(1)}}(T)$ and the horizontal line $1/L$. Again, we set the argument T of $\rho_{\zeta_{(1)}}(T)$ equal to T_1 and minimize $\rho_{\zeta_{(1)}}(T)$ over ζ_1, that is, find the value $\zeta_{(2)}$ such that

$$\rho_{\zeta_{(2)}}(T_1) = \min_\zeta \rho_\zeta(T_1).$$

Repeating the foregoing procedure, we obtain a sequence $T_0, T_1, \cdots$ converging to the optimal time T^0 and a sequence $\zeta_{(0)}, \zeta_{(1)}, \cdots$ converging to the vector ζ^0. The optimal control can be found by the formula

$$u_n(t) = L \operatorname{sgn} \sum_{k=1}^{n} \zeta_k^0 g_k(T^0 - t).$$

In the case where the function $\rho_\zeta(T)$ is convex relative to ζ for any fixed T and condition (2.15) is linear, one may use any standard minimization method (e.g., gradient or steepest descent methods) to find the absolute minimum of $\rho_\zeta(T)$. The foregoing iterative process can be performed automatically by using an optimizer for several variables and certain numerical devices.

Also, the iterative process can be simplified somewhat by restating problem (2.14)–(2.15) as follows: Determine the vector $\zeta = (\zeta_1, \cdots, \zeta_n)$ which gives the maximum value of T corresponding to the intersection between the graph $\rho_\zeta(T)$ and the horizontal line $1/L$, that is,

$$\max_{\zeta} T = T^0$$

under side condition $\rho_\zeta(T) = 1/T$.

It is easy to compute the gradient of $\rho_\zeta(T)$ from the formula

$$\frac{\partial \rho_\zeta(T)}{\partial \zeta_k} = \int_0^T g_k(T - t) \operatorname{sgn} \left[\sum_{k=1}^{n} \zeta_k g_k(T - t) \right] dt.$$

This relationship may be realized on a computer and it is then possible to perform a descent from any initial point under additional condition (2.15). In this case, it is possible to avoid the use of an optimizer for several variables.

Finally, consider the following situations where the condition specified by (2.6) cannot be fulfilled:

(i) $Q(x, T)$ cannot be made equal to $Q^*(x)$ for any $T > 0$;

(ii) the terminal time T is too small to permit $Q(x, T) = Q^*(x)$.

In situation (i), the optimal control always satisfies integral equation (1.31). In situation (ii), for sufficiently large T (i.e., for $T \geqq T^0$, where T^0 is the optimal time for problem (2.6)), (1.31) becomes meaningless (the argument of the signum function vanishes identically) and we must return to the solution of problem (2.6).

2.3. Example 2. Now, let us consider the problem of heating. Here, it is required to choose a control function $u(t)$ (temperature of the heating medium satisyfying $|u(t)| \leqq 1$, $0 \leqq t \leqq T$) to minimize the mean square deviation of the material's temperature from zero,

$$\phi^0 = \int_0^S Q^2(x, T)\, dx, \tag{2.18}$$

in the presence of the following side conditions:

$$\frac{\partial Q}{\partial t} = \frac{\partial^2 Q}{\partial x^2}, \tag{2.19}$$

$$-\lambda \frac{\partial Q}{\partial x}\bigg|_{x=0} = \alpha[u(t) - Q(0, t)], \quad \frac{\partial Q}{\partial x}\bigg|_{x=S} = 0, \tag{2.20}$$

$$Q(x, 0) = Q_0(x) = Q_0 = \text{const.} \tag{2.21}$$

The solution of (2.19) and (2.20) satisfying boundary condition (2.21) has the following form:

$$Q(x, t) = Q_0 \sum_{k=1}^{\infty} A_k e^{-\mu_k^2 t} \cos \mu_k x + \int_0^t \sum_{k=1}^{\infty} A_k \mu_k^2 e^{-\mu_k^2 (t-\tau)} \cos \mu_k x \cdot u(\tau)\, d\tau, \tag{2.22}$$

where μ_k, $k = 1, 2, \cdots$, is the sequence of positive roots of $\mu \tan \mu = \alpha S/\lambda = Bi$; α is the heat exchange coefficient; λ is the thermal conductivity, and the kernel $K(x, t)$ in (1.30) is given by

$$\sum_{k=1}^{\infty} A_k \mu_k^2 e^{-\mu_k^2 t} \cos \mu_k x = K(x, t). \tag{2.23}$$

Taking the first infinite sum in (2.22) (with $t = T$) as $Q^*(x)$ and using the orthogonality property of $\{\cos \mu_k x\}$, we can deduce from (1.31) that for sufficiently small T, the optimal control $u(t)$ satisfies the following integral equation:

$$u(t) = \operatorname{sgn}\left[Q_0 \sum_{k=1}^{\infty} B_k e^{-\mu_k^2 (2T-t)} - \sum_{k=1}^{\infty} \int_0^T \mu_k^2 B_k e^{-\mu_k^2 (2T-t-\tau)} u(\tau)\, d\tau \right], \tag{2.24}$$

where

$$B_k = \frac{\mu_k \sin^2 \mu_k}{\mu_k + \sin \mu_k \cos \mu_k}.$$

It has been shown in [20] that it is possible to transfer any constant initial temperature distribution, $Q_0(x) = Q_0 = \text{const.}$ $(0 \leqq x \leqq S)$, to the zero distribution if $T \geqq T^0$, where T^0 is the optimal time for transferring Q_0 to the zero distribution. In this case, the optimal control $u(t)$ satisfies $|u(t)| \equiv 1$ for $0 \leqq t \leqq T^0$, and has a denumerable infinite number of discontinuities which accumulate at the point T^0. For the case where $T < T^0$, the optimal control obtained from (2.24) has a finite number of discontinuities in the interval $[0, T]$. As $T \to T^0$, the number of discontinuities of

$u(t)$ tends to infinity. Also, the numerical solution of (2.24) indicates that the length of the time interval between any two successive discontinuities of $u(t)$ is always greater than the total remaining time to go after the second discontinuity.

Finally, it can be seen from (2.22) with $Q(x, T) \equiv 0$, $0 \leqq x \leqq S$, that due to linear independence of $\{\cos \mu_k x\}$, the problem of transferring any constant initial temperature distribution to the zero distribution can be reduced to the following problem of moments: Determine $u(t)$ satisfying $|u(t)| \leqq 1$ for $0 \leqq t \leqq T$, and

$$\frac{1}{\mu_k^2} = \int_0^T e^{\mu_k^2 t} u(t)\, dt, \qquad k = 1, 2, \cdots,$$

such that T takes on its minimum value T^0.

It follows from (2.9) and (2.10) (with $n \to \infty$) that we must solve

$$\min_{\zeta_k} \int_0^T \left| \sum_{k=1}^{n} \zeta_k\, e^{\mu_k^2 t} \right| dt = \frac{1}{L}$$

under side condition

$$\sum_{k=1}^{n} \frac{\zeta_k}{\mu_k^2} = 1.$$

The optimum control for the above problem with finite n has the form

$$u_n(t) = L \operatorname{sgn} \sum_{k=1}^{n} \zeta_k e^{\mu_k^2 t}.$$

Numerical solutions of this problem obtained by a digital computer have shown that for moderate values of the parameter Bi, and for most practical purposes, it is sufficient to consider only the moment problems up to the 4th order (due to rapid convergence of the series). The maximum deviation of $Q(x, T)$ from zero does not exceed one percent of the initial value Q_0.

The methods described above can easily be generalized to the case where several control variables are present in the system [27].

2.4. Application of L-problem of moments (restrictions on control functions and phase coordinates). Let us now discuss the control problem with restrictions on both the phase coordinates and the control functions.

Let the distributed system be described by (1.30) along with an additional equation

$$P(x, t) = \int_0^t F(x, t - \tau) u(\tau)\, d\tau,$$

where F is a prescribed function defined for $0 \leqq x \leqq S$, $0 \leqq t \leqq T$.

The time-optimal control problem with restrictions on control functions $|u(t)| \leqq L$ and phase coordinates $|P(x,t)| \leqq L$ $(0 \leqq x \leqq S, 0 \leqq t \leqq T)$ is formulated as follows: Given the function $Q^*(x)$, one has to find an admissible control function such that (2.6) is satisfied and

$$|P(x,t)| = \left| \int_0^t F(x, t-\tau)u(\tau)\, d\tau \right| \leqq L, \qquad 0 \leqq t \leqq T, \tag{2.25}$$

where x_1 is a point in the interval $[0, S]$ and the terminal time T is minimal.

This problem can also be reduced to an L-problem of moments in a normed linear space. Let us take n $(n \leqq \infty)$ linearly independent elements $g_1, \cdots, g_n$ of some normed linear space E. We wish to find the necessary and sufficient conditions required for the number $\alpha_1, \cdots, \alpha_n, L$ $(\sum_{k=1}^{n} \alpha_k^2 > 0, L > 0)$ to ensure the existence of a linear functional $l(g)$ satisfying the following relationships:

$$l(g_k) = \alpha_k, \quad \|l\| \leqq L, \quad k = 1, 2, \cdots, n. \tag{2.26}$$

In [25] it is shown that this problem is equivalent to the following problem:

Determine

$$\min_{\zeta_k} \left\| \sum_{k=1}^{n} \zeta_k g_k \right\| = \lambda \leqq \frac{1}{L} \tag{2.27}$$

under the additional condition

$$\sum_{k=1}^{n} \zeta_k \alpha_k = 1. \tag{2.28}$$

It can be shown that the problem (2.8), (2.10) discussed earlier is in fact equivalent to that of (2.26)–(2.28).

Here, the linear functional in question is of the form:

$$l(g) = \int_0^T g(T-t)u(t)\, dt, \quad \|l\| = \max_{[0,T]} |u(t)|. \tag{2.29}$$

The problem (2.25) can be reduced to the L-problem if we define

$$\|l\| = \max_{[0,T]} \left(u(t), \int_0^t F(x_1, t-\tau)u(\tau)\, d\tau\right) \tag{2.30}$$

to be the norm of the linear functional.

The problem is now reduced to minimization of the expression (2.27) under side condition (2.28).

The explicit form of the solutions will not be as simple as that for the problem without restrictions on the phase coordinates. It can be shown that

the corresponding norm for an element in E is

$$\| g \| = \min_{\psi} \left[\int_0^T \left| g(\tau) - \int_\tau^T F(x_1, t - \tau)\, d\psi(t) \right| d\tau + \int_0^T | d\psi(t) | \right]. \tag{2.31}$$

In this case, it is evident that the computational difficulties connected with the minimization of (2.27) are increased considerably because one has to minimize a functional depending not only on a finite number of parameters but also on a function.

Equation (2.31) can easily be generalized to the case where several phase coordinates of the system are restricted.

Note also that the method of moments may be generalized to include the case when restrictions (on both the control functions and phase coordinates) are of more general nature than those involving only absolute values. For example, the state or the control vector may be restricted to any convex domain with a nonempty interior.

2.5. Application of method of moments to oscillating systems. An important application of the method of moments is connected with optimal problems for oscillating systems [28], [29].

Consider an oscillating string governed by the wave equation

$$\frac{\partial^2 Q}{\partial t^2} = \frac{\partial^2 Q}{\partial x^2} \tag{2.32}$$

with boundary and initial conditions:

$$Q(0, t) = u(t), \qquad Q(\pi, t) = 0, \tag{2.33}$$

$$Q(x, 0) = Q_0(x), \qquad Q_t(x, 0) = Q_1(x). \tag{2.34}$$

Let the control function $u(t)$ be restricted by the inequality

$$\| u(t) \|_q = \left(\int_0^T | u(t) |^q \, dt \right)^{1/q} \leqq L. \tag{2.35}$$

It is required to set the string to rest in minimum possible time T, that is,

$$Q(x, T) = 0, \quad Q_t(x, T) = 0.$$

In the special case where restriction (2.35) is removed, the optimal regime is obvious from physical considerations, that is, the left end of the string should oscillate in such a way as if the string were semi-infinite to the left. The optimal transition time is exactly the time necessary for the initial disturbances to propagate away from the string "to infinity" through its left end.

One can see that in general, when restriction (2.35) is present, the optimal time will differ from that corresponding to the unrestricted control function by some integral number of cycles with period 2π. It has been shown in [28] that the optimal control $u_0(t)$ for the foregoing problem has the form:

(i) for $T = 2\pi n + \epsilon$, $n \neq 0$, $0 \leqq \epsilon \leqq 2\pi$,

$$u_0(t) = \begin{cases} \dfrac{1}{\pi(n+1)}\left(F(t) + \dfrac{c_2^{(P)}}{c_1^{(P)}}\right) & \text{for} \quad t \in [0, \epsilon], \\ \dfrac{1}{\pi n}\left(F(t) + \dfrac{c_2^{(P)}}{c_1^{(P)}}\right) & \text{for} \quad t \in [\epsilon, 2\pi], \end{cases} \tag{2.36}$$

where $c_1^{(P)}$, $c_2^{(P)}$ are constants, and $F(t)$ is a certain function which is extended to the whole axis with a period 2π (see [28]);

(ii) for $T \leqq 2\pi$,

$$u_0(t) = \frac{1}{T} F(t). \tag{2.37}$$

The same problem was solved by the "method of waves" [30]. The results do in fact coincide with those obtained earlier, namely,

(i) if $T = 2\pi n + \epsilon$, $n \neq 0$, $0 \leqq \epsilon \leqq 2\pi$, then

$$u_0(t) = \begin{cases} \dfrac{1}{n+1}\left\{\dfrac{1}{2} Q_0(t) + \dfrac{1}{2}\int_0^t Q_1(\zeta)\, d\zeta + K_0\right\} & \text{for} \quad t \in [0, \epsilon], \\ \dfrac{1}{n}\left\{\dfrac{1}{2} Q_0(t) + \dfrac{1}{2}\int_0^t Q_1(\zeta)\, d\zeta + K_0\right\} & \text{for} \quad t \in (\epsilon, 2\pi], \end{cases} \tag{2.38}$$

where K_0 is a constant and the function $u(t)$ is extended so that it has a period of 2π;

(ii) if $T \leqq 2\pi$, then

$$u_0(t) = \frac{1}{2} Q_0(t) + \frac{1}{2}\int_0^t Q_1(\zeta)\, d\zeta. \tag{2.39}$$

Similar problems for two-dimensional wave equations are discussed in [29] and [30].

3. Optimal control of systems governed by partial differential equations. Now we shall proceed to discuss optimization techniques for systems described by partial differential equations. First, we shall give an account of the optimality conditions for certain specific types of equations. The general approach will be discussed toward the end of the paper.

There is a considerable amount of work devoted to optimal processes governed by parabolic equations and systems. The first works pertaining to such processes are described in [12] and [13]. The problem of existence and

uniqueness of optimal control is discussed in [20], where the particular system considered is given by

$$\frac{\partial y}{\partial t} = \frac{\partial^2 y}{\partial x^2}, \qquad 0 \leqq x \leqq 1, \quad 0 \leqq t \leqq T, \tag{3.1}$$

$$y(0, x) = 0, \quad \frac{\partial y(t, 0)}{\partial x} = 0, \tag{3.2}$$

$$\frac{\partial y(t, 1)}{\partial x} = \alpha[p(t) - y(t, 1)], \qquad \alpha = \text{const.} > 0, \tag{3.3}$$

where the control $p(t)$ is a measurable function such that $| p(t) | \leqq 1$ almost everywhere on $[0, T]$. The optimality criterion is given by the functional

$$J_\gamma[p] = \int_0^T [y(T, x) - y_0(x)]^2 \, dx + \gamma \int_0^T p^2(t) \, dt, \quad \gamma = \text{const.} > 0,$$

where T is fixed, and the final state of the system is free.

It can be shown that given any admissible control, the solution to the corresponding initial boundary value problem (3.1)–(3.3) is unique. Here, by a solution we mean that the function $y(t, x)$ satisfies (3.1) inside the domain $Q(0 \leqq x \leqq 1, 0 \leqq t \leqq T)$ and initial condition (3.2) in the classical sense, and satisfies condition (3.3) in the weak sense, i.e.,

$$\lim_{\epsilon \to 0} \int_0^T \left[\frac{\partial y(t, 1 - \epsilon)}{\partial x} + \alpha y(t, 1 - \epsilon) - \alpha p(t) \right] \varphi(t) \, dt = 0$$

for any function $\varphi \in C_0^\infty(0, T)$, where

$$\sup | y(t, x) | \leqq 1, \quad \sup \left| \frac{\partial y(t, x)}{\partial x} \right| \leqq 2\alpha.$$

Having taken $J_\gamma[p], \gamma > 0$, as the optimality criterion, it is possible to establish the existence and uniqueness of the optimal control for any $y_0 \in L_2[0, 1]$. For the case where $\gamma = 0$, the optimal control exists but may not be unique. The uniqueness depends upon the nature of $y_0(x)$.

In [20], no method for the construction of J_γ-optimal control functions with the aid of certain optimality conditions is given. Instead, a numerical method based on approximating an arbitrary function by a piecewise constant function is presented. This approximation procedure leads to the investigation of the extremum of a certain function of n variables. The convergence of the procedure is demonstrated. Also, the problem where the initial condition serves as a control, and J_0 plays the role of an optimality criterion, as well as the time-optimal problem of transfer into a certain neighborhood of the function u_0 in $L_2[0, 1]$ are considered. Existence theorems are established but no effective method of solution is presented.

An alternative approach to the optimal control problem is based upon ideas due to L. I. Rosonoer [31] who had realized them for the case of one independent variable.

In [32] a process governed by a quasi-linear equation is discussed. The equation under consideration has the form:

$$\frac{\partial v}{\partial t} = f_0(t, x, v, u) + \sum_{k=1}^{n} f_k(t, x, v, u) \frac{\partial v}{\partial x_k} \tag{3.4}$$

with initial condition

$$v(0, x) = v_0(x) \tag{3.5}$$

and a boundary condition of the first kind:

$$v(t, x) \,|_{S_1} = v_1(t, x), \tag{3.6}$$

where $x = (x_1, \cdots, x_n)$ belongs to a domain D bounded by a surface S composed of two parts S_1 and S_2 ; and u is an m-dimensional control vector. The admissible controls $u = u(t, x)$ are assumed to be piecewise continuous and satisfy the constraints:

$$\varphi_k(u) \leqq 0, \qquad k = 1, \cdots, r \leqq m.$$

It is assumed that the solution corresponding to any admissible control is unique. The optimality criterion for the foregoing system may take on one of the following forms:

$$J_1 = \int_0^T \int_D \left[\mathcal{J}(t, x, v, u) + \sum_{k=1}^{n} \mathcal{J}_k(t, x, v, u) \frac{\partial v}{\partial x_k} \right] dx\, dt,$$

$$J_2 = \int_D \mathcal{J}_0(x, v(T, x), u(x))\, dx,$$

where the upper limit T is fixed.

A necessary condition for J_1-optimality of the control function $u(t, x)$ is that the inequality

$$\Delta H = H\left(t, x, v, \frac{\partial v}{\partial x_1}, \cdots, \frac{\partial v}{\partial x_n}, p, u + \Delta u\right) - H\left(t, x, v, \frac{\partial v}{\partial x_1}, \cdots, \frac{\partial v}{\partial x_n}, p, u\right) \leqq 0$$

holds for almost all $t \in [0, T]$ and $x \in D$, where $p(x, t)$ is determined from the equation

$$\frac{\partial p}{\partial t} = -\frac{\partial H}{\partial v} + \sum_{k=1}^{n} \frac{\partial}{\partial x_k} \left(\frac{\partial H}{\partial v_{x_k}} \right)$$

with boundary conditions:

$$p(T, x) = 0, \quad \left\{\sum_{k=1}^{n} \frac{\partial H}{\partial v_{x_k}} \cos(\mathbf{n}, x_k)\right\}_{S_2} = 0,$$

where $S_2 + S_1 = S$, $H = -\mathcal{J} + pf$, and $\mathbf{n}$ is the outer normal to S_2.

When (3.4) is linear and

$$\mathcal{J} = b_0(t, x)v + \sum_{k=1}^{n} b_k(t, x) \frac{\partial v}{\partial x_k} + \varphi_0(t, x, u),$$

the foregoing condition is also sufficient.

Similar results can be established for the J_2-optimal problem.

3.1. Parabolic systems. The optimal control problems for parabolic systems of the second order are discussed in [33]–[35]. The exact statement of the problem is as follows:

Let D be a region in an n-dimensional Euclidean space, bounded by a class $A^{(2)}$ surface Γ (see [36]). Let $L = (L_1, \cdots, L_m)$ be an elliptic operator of the form:

$$L_i y = \sum_{\nu=1}^{m} \sum_{j,k=1}^{n} a_{jk}^{i\nu} \frac{\partial^2 y_\nu}{\partial x_j \partial x_k}, \qquad i = 1, \cdots, m,$$

where the coefficients $a_{jk}^{i\nu}(t, x)$, for any $t \in [0, T]$, belong to class C^2 relative to the variable $x = (x_1, \cdots, x_n) \in D + \Gamma$. Let $M = (M_1, \cdots, M_m)$ denote the adjoint operator corresponding to L:

$$M_i z = \sum_{\nu=1}^{m} \left[\sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{jk}^{i\nu} \frac{\partial z^\nu}{\partial x_k}\right) + \sum_{j=1}^{n} \frac{\partial}{\partial x_j}(l_j^{\nu i} z_\nu)\right],$$

$$l_j^{\nu i} = -\sum_{k=1}^{n} \frac{\partial a_{jk}^{i\nu}}{\partial x_k}, \qquad i = 1, \cdots, m.$$

It can be easily verified that the following relationship is valid:

$$\sum_{i=1}^{m} \int_D (z_i L_i y - y_i M_i z)\, dx$$

$$= \sum_{i,\nu=1}^{m} \sum_{j=1}^{n} \int_\Gamma \left[\sum_{k=1}^{n} a_{jk}^{i\nu}\left(z_i \frac{\partial y_\nu}{\partial x_k} - y_\nu \frac{\partial z_i}{\partial x_k}\right) + l_j^{i\nu} y_\nu z_i\right] X_j(x)\, d\sigma,$$

where $X_j(x)$ denotes the direction cosines of the outer normal to the surface Γ. Following a procedure described in [36], it is possible to rewrite the above expression as

$$\int_D \sum_{i=1}^{n} (z_i L_i y - y_i M_i z)\, dx = \int_\Gamma \sum_{i=1}^{m} (z_i P_i y - y_i Q_i z)\, d\sigma,$$

where

$$P_i\, y = \sum_{\nu=1}^{m} \left[a_l^{i\nu} \frac{dy_\nu}{dl_{i\nu}} + b_{i\nu}\, y_\nu \right],$$

$$Q_i\, z = \sum_{\nu=1}^{m} \left[a_\lambda^{\nu i} \frac{dz_\nu}{d\lambda_{i\nu}} + d_{i\nu}\, z_\nu \right]. \tag{3.7}$$

In (3.7), the directions $l_{i\nu}$ are chosen arbitrarily except that $\cos(\mathbf{n}, l_{i\nu}) > 0$ ($\mathbf{n}$ is the outer normal to Γ) and the direction cosines belong to class C^1 on Γ. The directions $\lambda_{i\nu}$ are chosen in accordance with those of $l_{i\nu}$.

We shall consider a control process which is governed by the following system of parabolic equations:

$$L_{il}\, y = \frac{\partial y_i}{\partial t} - L_i\, y = f_i(t, x, y, y_x, u), \tag{3.8}$$

$$0 \leqq t \leqq T, \quad x \in D,$$

where the f_i are continuous in t and have continuous derivatives with respect to $y_1, \cdots, y_m, \partial y_1/\partial x_1, \cdots, \partial y_m/\partial x_n$. Moreover, the f_i satisfy a Lipschitz condition in u. The same conditions are imposed on $\partial f_i/\partial y_k$ and $\partial f_i/\partial y$. It is assumed that the control u takes on values in a bounded (open or closed) region U of a p-dimensional Euclidean space.

Let us assume further that the solution $y(t, x) = (y_1, \cdots, y_m)$ of (3.8) satisfies the initial condition

$$y(0, x) = a(x), \qquad x \in D, \tag{3.9}$$

where $a(x)$ is a continuous vector-valued function.

The boundary conditions are chosen in one of the following forms:

$$y_i(t, x) = \varphi_i(t, x), \qquad x \in \Gamma, \quad 0 \leqq t \leqq T, \quad i = 1, \cdots, m, \tag{3.10}$$

or

$$P_i(t, x)y = \varphi_i(t, x, y, v), \qquad x \in \Gamma, \quad 0 \leqq t \leqq T, \quad i = 1, \cdots, m, \tag{3.11}$$

where the operators P_i are defined by (3.8) in which the functions $a_l^{i\nu}(t, x)$, $b_{i\nu}(t, x)$ are continuously differentiable, and satisfy the same conditions as those imposed upon f_i in (3.9). The parameter v takes on values in some bounded (open or closed) domain V of a q-dimensional Euclidean space.

In what follows, we shall speak about the first or the second boundary value problem depending on whether the boundary conditions have been chosen in the form (3.10) or (3.11). In the first boundary value problem, the admissible control functions $u(t, x)$ are piecewise continuous and take on values in V. Also, the form of certain components of the vector $u(t, x)$

may be specified. In particular, they may depend either on t only or on x only. The surfaces of discontinuities are assumed to be smooth and each of them is either orthogonal to the t-axis or is such that in the neighborhood of any point on the surface we can introduce a nondegenerate coordinate transformation $\tau = t, \xi_i = \xi_i(t, x), u = 1, \cdots, n$, such that the surface becomes a portion of the plane $\xi_n = 0$.

In the second boundary value problem, the admissible controls will be represented by $\omega(t, x) = (u(t, x), v(t, x))$, where the $u(t, x)$ satisfy the above stated conditions, and the $v(t, x)$ are of the same type as $u(t, x)$, except that their values belong to some domain V. In each of these problems, it is assumed that the solution corresponding to each admissible control exists and is unique.

For the first and second boundary value problems, the respective criteria of optimality are as follows:

$$S_1 = \sum_{i=1}^{m} \left[\int_D \alpha_i(x) y_i(T, x)\, dx + \int_0^T \int_D \beta_i(t, x) y_i(t, x)\, dx\, dt + \int_0^T \int_\Gamma \gamma_i(t, x) P_i(t, x) y\, d\sigma\, dt \right],$$

$$S_2 = \sum_{i=1}^{m} \left[\int_D \alpha_i(x) y_i(T, x)\, dx + \int_0^T \int_D \beta_i(t, x) y_i(t, x)\, dx\, dt + \int_0^T \int_\Gamma \gamma_i(t, x) y_i\, d\sigma\, dt \right],$$

where α_i, β_i, γ_i are prescribed continuous functions. The final state of the system at time T satisfies the following conditions:

$$\begin{aligned} &\phi_\alpha(T, x, y(T, x)) = 0, \qquad \alpha = 1, \cdots, j, \\ &\int_D \psi_\beta(T, x, y(T, x), y_x(T, x))\, dx = C_\beta, \qquad \beta = 1, \cdots, k, \end{aligned} \tag{3.12}$$

where the C_β are given constants and $j + k = m$; the function ψ_β is assumed to be independent of y_x in the second boundary value problem. The final time T is not specified in general.

We shall say that an admissible control $u(t, x)[\omega(t, x)]$, $0 \leqq t \leqq T$, $x \in D$, in the first [second] boundary value problem transfers the system from state (3.9) into a set defined by (3.12) if the corresponding solution initiating from state (3.9) at the time $t = 0$ satisfies (3.12) at the time $t = T$.

In the sequel, two types of optimal control problems will be considered, namely,

(i) the final state satisfies (3.12) and the terminal T is not necessarily

fixed [33], [34];

(ii) the terminal time T is fixed and the final state of the system is free [35].

To formulate the optimality conditions for the above problems, we introduce the functions H and h defined by

$$H(t, x, w, u) = \sum_{i=1}^{m} z_i f_i(t, x, y, y_x, u),$$

$$h(t, x, p, v) = \sum_{i=1}^{m} z_i \varphi_i(t, x, y, v),$$

where

$$p = (z_1, \cdots, z_m, y_1, \cdots, y_m),$$

$$w = \left(z_1, \cdots, z_m, y_1, \cdots, y_m, \frac{\partial y_1}{\partial x_1}, \cdots, \frac{\partial y_m}{\partial x_n}\right).$$

Let $y(t, x)$ be the solution corresponding to the first [second] boundary value problem with an admissible control $u(t, x)$ [$\omega(t, x)$]. We introduce the functions $z_i(t, x)$ defined by the partial differential equations

$$M_{it} z = -\frac{\partial H(t, x, w, u)}{\partial y_i} + \sum_{k=1}^{n} \frac{d}{dx_k}\left(\frac{\partial H(t, x, w, u)}{\partial y_{ik}}\right) + \beta_i(t, x), \tag{3.13}$$

$$0 \leqq t \leqq T, \quad x \in D, \quad i = 1, \cdots, m,$$

with "initial" conditions:

$$z_i(T, x) = -\alpha_i(x) - \sum_{\alpha=1}^{j} a_\alpha(x) \frac{\partial \phi_\alpha}{\partial y_i} - \sum_{\beta=1}^{k} b_\beta \left(\frac{\partial \psi}{\partial y_i} - \sum_{\nu=1}^{n} \frac{d}{dx_\nu}\left[\frac{\partial \psi_\nu}{\partial y_{i\nu}}\right]\right), \quad x \in D, \tag{3.14}$$

where $M_{it}(z) = \partial z_i/\partial t + M_i z$, $y_{ik} = \partial y_i/\partial x_k$. The constants b_β and functions a_α are as yet undetermined. For the first boundary value problem, the boundary conditions for (3.13) are chosen as

$$z_i(t, x) = \gamma_i(t, x), \quad x \in \Gamma, \quad i = 1, \cdots, m, \tag{3.15}$$

where the γ_i are defined in S_1.

For the second boundary value problem, the boundary conditions for (3.13) are chosen as

$$Q_i(t, x) z = \frac{\partial h(t, x, p, v)}{\partial y_i} + \sum_{k=1}^{n} \frac{\partial H(t, x, w, u)}{\partial y_{ik}} X_k(x) - \gamma_i(t, x), \tag{3.16}$$

$$x \in \Gamma, \quad i = 1, \cdots, m,$$

where the Q_i are operators defined in (3.7); the $X_k(x)$ represent the direction cosines of the outer normal to Γ, and the γ_i are defined in $\mathfrak{S}_2$.

Both boundary value problems (3.13)–(3.16) associated with $z_i(t, x)$ are linear and satisfy the same general conditions as those in (3.8)–(3.11). One can therefore conclude that if the functions $a_\alpha(x)$ and constants b_β are specified, then the function $z(t, x)$ corresponding to each admissible control is unique.

Let $u(t, x)$ [$\omega(t, x)$] be an admissible control for the first [second] boundary problem (3.8)–(3.10) which transfers the system from state (3.9) to the set defined by (3.12). Let $y(t, x)$, $z(t, x)$ denote the corresponding solutions of the first or second boundary value problem (3.8)–(3.11) and (3.13)–(3.16). We introduce the functionals

$$J_1[u] = \int_0^T \int_D H(t, x, w(t, x), u)\, dx\, dt,$$

$$J_2[v] = \int_0^T \int_\Gamma h(t, x, p(t, x), v)\, d\sigma\, dt,$$

defined for admissible controls $u(t, x)$, $v(t, x)$.

We shall say that an admissible control $\omega(t, x)$ of the second boundary value problem (3.8)–(3.11), which transfers the system from state (3.9) to the set defined by (3.12), satisfies the condition of maximum if, for any other $\omega^1 = (u^1, v^1)$ (also capable of transferring the system from state (3.9) to the set defined by (3.12)), the inequalities

$$(3.17)\quad \Delta J_1 = \int_0^\tau \int_D [H(t, x, w(t, x), u^1) - H(t, x, w(t, x), u)]\, dx\, dt \leqq 0,$$

$$(3.18)\quad \Delta J_2 = \int_0^\tau \int_\Gamma [h(t, x, p(t, x), v^1) - h(t, x, p(t, x), v)]\, d\sigma\, dt \leqq 0,$$

are satisfied, where $\tau = \min\{T, T_1\}$ and $[0, T_1]$ is the time interval on which ω^1 is defined.

In a similar way, we can define the condition of maximum for the first boundary value problem. If we do not impose any restrictions on the form of dependency of the control functions on the arguments x and t, then inequalities (3.17), (3.18) are equivalent to the following conditions:

$$H(t, x, w(t, x), u(t, x)) \;((=))\; \sup_{u \in U} H(t, x, w(t, x), u), \quad x \in D,\ 0 \leqq t \leqq T,$$

$$h(t, x, p(t, x), v(t, x)) \;(=)\; \sup_{v \in V} h(t, x, p(t, x), v), \quad x \in \Gamma,\ 0 \leqq t \leqq T,$$

where the symbol $((=))$ denotes equality valid everywhere in the domain $(0 \leqq t \leqq T, x \in D)$ excepting possibly points lying on a finite number of n-dimensional surfaces, whose $(n + 1)$-dimensional volume is equal to

zero. The symbol $(=)$ is defined analogously except that we taken $n-1$ and Γ instead of n and D respectively.

The necessary conditions for optimal control are given by the following theorem (its proof requires the use of a rather stringent condition, namely, the completeness of the class of admissible controls [35]).

THEOREM 2. *In order that an admissible control* $u(t, x)$ $[\omega(t, x) = (u(t, x), v(t, x))]$ *of the first* [*second*] *boundary value problem, defined on the domain* $(0 \leqq t \leqq T, x \in D)$ *and transferring the system from state* (3.9) *to the set defined by* (3.12), *be optimal relative to the functional* $S_1[S_2]$, *it is necessary that there exist functions* $z_i(t, x)$, $a_\alpha(x)$ *and constants* b_β *such that*:

(i) *the functions* $y(t, x)$, $z(t, x)$, $u(t, x)$ $[\omega(t, x)]$, $Q(x)$ *and constants* b_β *form a solution of* (3.9) *and* (3.13) *with additional conditions* (3.9)–(3.12) *and* (3.14)–(3.16);

(ii) *the control* $u(t, x)$ $[\omega(t, x) = (u(t, x), v(t, x))]$ *satisfies the maximum condition relative to* $z(t, x)$:

(iii) *the condition*

$$\frac{dS_i}{dt} + \int_D \left[\sum_{\alpha=1}^{j} a_\alpha(x) \frac{d\phi(T, x, y(T, x))}{dT} + \sum_{\beta=1}^{k} b_\beta \frac{d\psi_\beta(T, x, y(T, x), y_k(T, x))}{dT} \right] dx = 0$$

is satisfied at the terminal time instant $t = T$, *where* $i = 1$ *for the first boundary value problem and* $i = 2$ *for the second boundary value problem.*

For problems of the second type, where the terminal time T is fixed and the terminal state is free, the optimality conditions can be deduced from Theorem 2. For this case, we set $b_\beta = 0$, $a_\alpha(x) = 0$ $(\alpha = 1, \cdots, j, \beta = 1, \cdots, k)$ in (3.14), and delete condition (iii) from Theorem 2. The requirement of completeness of the class of admissible controls can also be deleted. It turns out that when the control functions enter additively into the system equations and the boundary value problem is linear, the foregoing optimality conditions are not only necessary but also sufficient [33], [34]. Also, in this case, it is possible to derive an explicit expression for the increment of the functional. This fact may be used to solve certain linear problems in the theory of invariance.

In many physical situations, the automatic control systems contain both distributed and lumped distributed parameter elements. Optimality conditions for these kinds of system are obtained in [39], [51].

In [40] solutions are obtained for the problem of analytical design of regulators, when the process is describable by

$$\frac{\partial y_i}{\partial t} = f_i(x, y, y_x, y_{xx}, u), \qquad i = 1, \cdots, m, \tag{3.19}$$

where $x = (x_1, \cdots, x_n)$ is a point in a domain D and $f_i(x, 0, 0, 0, 0) \equiv 0$.

It is assumed that for a given set of initial and boundary conditions and for any admissible control, there exists a unique solution $y(t, x)$. In particular, the trivial (unperturbed) solution $y = 0$ corresponds to the solution with control function $u = 0$.

The optimality criterion is given by the functional

$$\mathcal{J} = \int_0^\infty w\, dt,$$

$$w = \int_D \int_D \sum_{i,j=1}^{n} \omega_{ij}(x, \xi) y_i(x, t) y_j(\xi, t)\, d\xi\, dx + \int_D \omega_{00}(x) u^2\, dx.$$

Also, we introduce a nonnegative function $\rho(y)$ ($\rho(0) = 0$) which serves as a measure of disturbance and, at the same time, characterizes the stability of the process. For the system under consideration, we may take ρ to be

$$\rho(y) = \sum_{k=1}^{m} \left[\int_D y_k^2(x) + \sum_{i=1}^{n} \left(\frac{\partial y_k}{\partial x_i} \right)^2 \right] dx.$$

The basic result of the paper can be summarized by the following theorem.

THEOREM 3. *If there exist a continuous positive definite functional $v(y)$ and a control function $u = u^0(y)$ such that the derivative dv/dt, calculated in accordance with differential equations* (3.19), *satisfies*

$$\left(\frac{dv}{dt} + w \right)_{u=u^0} = \min_u \left(\frac{dv}{dt} + w \right) = 0,$$

where the functional w is positive definite relative to the measure ρ for $u = u^0$, then the unperturbed motion $y = 0$ is asymptotically stable relative to the measure ρ and the functional $\mathcal{J}$ is minimal.

With the aid of this theorem, the author obtains the solution of the synthesis problem for the linear equation (3.19). Analogous results are obtained for distributed systems with stochastic parameters [41]. Other problems related to the optimal control of stochastic distributed systems are discussed in [42], [46].

In [47]–[50] the optimal control problem is formulated in the framework of functional analysis. In [48], for example, the process described by the Cauchy problem

$$\frac{dx}{dt} = F(t, x, u), \quad x(t_0) = x_0,$$

is considered, where x is an element of a Banach space B, u belongs to a certain subset U of a topological space and F represents a certain (generally unbounded) operator from $B \times V$ into B. The derivative is defined in the

sense of convergence in norm. The admissible controls are summable functions taking on values in U.

The optimality criterion is a functional of the form

$$J = \int_{t_0}^{t_1} f(t, x, u)\, dt.$$

In [49], optimality conditions are obtained in the form of Pontryagin's maximum principle. However, this result is valid only if one performs the variation of control in the narrow strips orthogonal to the t-axis, and not in regions of arbitrary shape. In other words, the "optimal" control in this theory is being compared not with all the admissible controls. In what follows, we shall present a generalization of Pontryagin's maximum principle for distributed systems [51]. Also, we shall point out certain features which are found only in optimal control problems associated with distributed systems.

3.2. General variational approach. In what follows, we shall call "distributed controls" the control functions entering into the basic differential equations, and "boundary controls" those entering into the initial and/or boundary conditions. One can of course conceive situations where both distributed and boundary controls are present.

In many cases, the control functions are subjected to certain restrictions; let these restrictions be of the form

$$G(u; x, y) \geqq 0, \tag{3.20}$$

where $u(x, y)$ represents a p-dimensional distributed control vector, x and y are independent variables (for simplicity, only the two-dimensional case will be discussed here). Similarly, the restrictions for boundary controls can be represented by

$$g(v; t) \geqq 0, \tag{3.21}$$

where $v(t)$ is the boundary control and t denotes a parameter along the boundary curve. The functions G and g are assumed to be differentiable.

A typical formulation of the control problem is as follows [51]. Let S denote a closed domain in the (x, y)-plane with piecewise continuous boundaries Σ_1, Σ_2. Let us consider the following system of partial differential equations defined on S:

$$\begin{aligned} \Xi_i &\equiv z_x^i - X_i(z, \zeta, u; x, y) = 0, \\ \mathrm{H}_i &\equiv z_y^i - Y_i(z, \zeta, u; x, y) = 0, \\ &\frac{\partial X_i}{\partial y} - \frac{\partial Y_i}{\partial x} = 0, \qquad i = 1, 2, \cdots, n. \end{aligned} \tag{3.22}$$

The latter equations contain total derivatives over all arguments included, not only over x and y entering explicitly. Let $z = (z^1, \cdots, z^n)$, $\zeta = (\zeta^1, \cdots, \zeta^\nu)$, $u = (u^1, \cdots, u^p)$, where z^i, ζ^j and u^k are functions of x and y, and u corresponds to a distributed control. The ordered pair (z, ζ) will be called the state of the system.

The equations in (3.22) represent a standard form for any system of partial differential equations (a special form of the Pfaffian system [52, pp. 323–324]). In other words, any system can be reduced to the form (3.22) (with an increase in the number of dependent variables if necessary). For example, the Helmholtz equation

$$z^1_{xx} + z^1_{yy} + uz^1 = 0$$

is equivalent to the system

$$z_x{}^1 = z^2, \quad z_y{}^1 = z^3, \quad z_x{}^2 = -\zeta^2 - uz^1, \quad z_y{}^2 = \zeta^1, \quad z_x{}^3 = \zeta^1, \quad z_y{}^3 = \zeta^2.$$

The wave equation

$$z^1_{yy} - (kz_x{}^1)_x = 0$$

is equivalent to the system

$$z_x{}^1 = -\zeta^1/k, \quad z_y{}^1 = \zeta^2, \quad z_x{}^2 = \zeta^2, \quad z_y{}^2 = -\zeta^1.$$

The form (3.22) is more general than that represented by a single partial differential equation of higher order. For example, the system

$$z_x{}^1 = \zeta^1, \quad z_y{}^1 = -\zeta^2 + u, \quad z_x{}^2 = \zeta^2, \quad z_y{}^2 = \zeta^1,$$

is equivalent to equations

$$\Delta z^1 = \frac{\partial u}{\partial y}, \quad \Delta z^2 = -\frac{\partial u}{\partial x}$$

only if u is differentiable. Note that the latter pair of equations contain derivatives of the control but not the control itself. The optimum control problems for higher order equations depending on controls (but not on their derivatives) are examined in [17] and [39].

Now, let us introduce the constraints associated with the control functions in (3.22). The first r_1 of these constraints can be expressed by equalities

$$G_k(u; x, y) = 0, \quad k = 1, \cdots, r_1, \tag{3.23}$$

and the remaining $r - r_1$ constraints are given by inequalities

$$G_k(u; x, y) \geqq 0, \quad k = r_1 + 1, \cdots, r \leqq p. \tag{3.24}$$

Suppose that the values of the first n_1 ($n_1 \leqq n$) functions z^i on Σ_1 are

known, where Σ_1 is assumed to be given, i.e.,

$$z^i|_{\Sigma_1} = z_1{}^i(t), \qquad i = 1, \cdots, n_1. \tag{3.25}$$

The number n_1 is determined by the conditions of the given problem.

The outer curve Σ_2 is not known a priori. It is assumed that there are n_2 ($n_2 \leqq n$) ordinary differential equations of the form

$$\Theta_{i_k} \equiv \frac{dz^{i_k}}{dt} - T_{i_k}(z, v, t) = 0, \qquad i_k = i_1, \cdots, i_{n_2}, \tag{3.26}$$

defined along Σ_2.

These equations include a set of functions (boundary controls)

$$v^l = v^l(t), \qquad l = 1, \cdots, \pi,$$

depending on the parameter t.

The values of z^{ik} at $t = 0$ are assumed to be known. Also, the boundary controls satisfy a set of constraints expressed by equalities

$$g_k(v; t) = 0, \quad k = 1, \cdots, \rho_1, \tag{3.27}$$

and inequalities

$$g_k(v; t) \geqq 0, \quad k = \rho_1 + 1, \cdots, \rho \leqq \pi. \tag{3.28}$$

The total number of these constraints is equal to $\rho \leqq \pi$. It is essential that the solutions to (3.22), (3.25), (3.26) under constraints (3.23), (3.24), (3.27), (3.28) exist. This requirement is satisfied if the physical problem is properly formulated. Note that there is a basic difference between the nature of the ζ and u variables in (3.22), namely, the control variable u can be directly manipulated externally, whereas the variable ζ can be manipulated only indirectly through the system. Generally, the solutions to (3.22), (3.25), (3.26) may not exist for arbitrary u and ζ defined on S. It is possible that the solutions corresponding to some u exist but they are not unique. In the subsequent discussions, we shall not consider such cases.

Now, the Mayer-Bolza problem can be formulated as follows [51]: In a suitable class of functions, determine the controls u, v and their corresponding solutions z and ζ such that the functional

$$J = \iint_S F(z, \zeta, u; x, y)\, dx\, dy + \oint_{\Sigma_1} f_1(z; t)\, dt + \oint_{\Sigma_2} f_2(z, v; t)\, dt \tag{3.29}$$

takes on its minimum value under side conditions (3.22)–(3.28). The functions X_i, Y_i, F, f_1, f_2 are assumed to be differentiable with respect to all their arguments.

In what follows, we shall clarify what we mean by a "suitable class" of functions, since this point is crucial for the existence of solutions to the optimal control problem.

To proceed further, it is necessary to define the class of admissible controls and to investigate the possible behavior of the state variables. The "distributed controls" are assumed to belong to a certain subset of the class of all piecewise continuous functions of two independent variables. Also, the discontinuities of distributed controls, if they exist, lie along smooth closed isolated curves. In the following discussions, we shall assume, for simplicity, that there is only one such discontinuity along a curve Σ_0 whose points lie inside S; moreover, this curve can be continuously deformed into any one of the boundary curves Σ_1 and Σ_2.

The state variables z^i are assumed to be continuous across the curve Σ_0. The variables ζ^j are in general discontinuous but their values on both sides of Σ_0 are related by the requirement that the tangential derivatives $\partial z^i/\partial t$ along Σ_0 should be continuous, namely,

$$[X_i x_t + Y_i y_t]_-^+ = 0, \qquad i = 1, \cdots, n. \tag{3.30}$$

Also, the boundary controls v are assumed to be members of a certain subset of the class of all piecewise continuous functions of t. For simplicity, it is assumed that there is only one point t_* of such discontinuity. This point corresponds to a corner point of the curve Σ_2 and z^i is continuous at that point. The latter assumption is necessary for discontinuity of dz^{i_k}/dt across t_*, because otherwise we would have permitted some line of discontinuity of $\partial z^i/\partial x$, $\partial z^i/\partial y$, starting from t on the boundary Σ_2 and penetrating inside S. Such a line of discontinuity may be connected only with the jump of distributed control u across it.[1] But we have already ruled out this possibility earlier by assuming that there is no such line intersecting the boundaries Σ_1 or Σ_2.

First, we shall transform inequality constraints (3.24) and (3.28) to equality constraints.

This can be accomplished by introducing (real) artificial control (slack) variables $u = (u_*^{r_1+1}, \cdots, u_*^{r})$, $v_* = (v_*^{\rho_1+1}, \cdots, v_*^{\rho})$ such that

$$G_k^* = G_k(u; x, y) - (u_*^k)^2 = 0, \qquad k = r_1 + 1, \cdots, r, \tag{3.31}$$

$$g_k^* = g_k(v; t) - (v_*^k)^2 = 0, \qquad k = \rho_1 + 1, \cdots, \rho. \tag{3.32}$$

Now, inequalities (3.24) and (3.28) can be replaced by the above equalities. Thus, we have transformed the problem involving variations of controls in a closed region to one with variations of controls in an open region, but with an increased number of control variables.

[1] Except, perhaps, some hyperbolic problems (see, for example, [61]).

The equations (3.22) and restrictions (3.23) and (3.31) are taken into consideration by introducing Lagrange multipliers:

$$\xi_i(x, y), \quad \eta_i(x, y), \quad i = 1, \cdots, n;$$

$$\Gamma_k(x, y), \quad k = 1, \cdots, r_1; \quad \Gamma_k^*(x, y), \quad k = r_1 + 1, \cdots, r.$$

In order to present the necessary conditions for stationarity, we introduce the "Hamiltonians":

$$H(z, \zeta, u, u_*) = \xi X + \eta Y - F - \Gamma G - \Gamma^* G^*, \tag{3.33}$$

$$h(z, v, v_*) = \theta T - f_2 - \gamma g - \gamma^* g^*, \tag{3.34}$$

where θ, γ, γ^* denote the Lagrange multipliers corresponding to (3.26), (3.27), (3.32), respectively.

Now, the Euler equations can be written as follows:

$$\frac{\partial \xi_i}{\partial x} + \frac{\partial \eta_i}{\partial y} = -\frac{\partial H}{\partial z^i}, \qquad i = 1, \cdots, n, \tag{3.35}$$

$$\frac{\partial H}{\partial \zeta^j} = 0, \qquad j = 1, \cdots, \nu, \tag{3.36}$$

$$\frac{\partial H}{\partial u^k} = 0, \qquad k = 1, \cdots, p, \tag{3.37}$$

$$\frac{\partial H}{\partial u_*{}^k} \equiv 2\Gamma_k^* u_k^* = 0, \qquad k = r_1 + 1, \cdots, r.$$

The natural boundary conditions along the boundary Σ_1 are

$$\frac{\partial f_1}{\partial z^i} + \xi_i y_t - \eta_i x_t = 0, \qquad i = n_1 + 1, \cdots, n, \tag{3.38}$$

and the corresponding conditions along the (unknown) boundary Σ_2 ($\delta_i{}^l = 0$, $(i \neq l)$; $\delta_i{}^i = 1$) are

$$\delta_i{}^{i_k} \frac{d\theta_{i_k}}{dt} + \frac{\partial h}{\partial z^i} - \xi_i y_t + \eta_i x_t = 0, \qquad i = 1, \cdots, n, \tag{3.39}$$

$$\frac{\partial h}{\partial v^l} = 0, \quad l = 1, \cdots, \pi, \qquad \frac{\partial h}{\partial v_*{}^l} \equiv 2\gamma_l^* v_*{}^l = 0, \quad l = \rho_1 + 1, \cdots, \rho, \tag{3.40}$$

$$F + \frac{f_2}{\rho_2} + \frac{\partial f_2}{\partial \mathbf{n}} = 0,$$

where ρ_2 denotes the radius of curvature, and $\mathbf{n}$ the outer normal to Σ_2.

The Weierstrass-Erdmann conditions along the curve Σ_0 of discontinuity

of the distributed controls are

$$(\xi_i y_t - \eta_i x_t)_-{}^+ = 0, \qquad i = 1, \cdots, n,$$
$$(H)_-{}^+ = z_x{}^{i-}(\xi_i)_-{}^+ + z_y{}^{i-}(\eta_i)_-{}^+ \tag{3.41}$$

(to sum up over i), and the corresponding conditions at the point t_* of discontinuity of boundary controls on the boundary Σ_2 are (in the second equation, to sum up over i_k)

$$\theta_{i_k}^-(t_*) = \theta_{i_k}^+(t_*),$$
$$\theta_{i_k}^-(t_*) \operatorname{grad} z^{i_k-}(t_*) = \theta_{i_k}^+(t_*) \operatorname{grad} z^{i_k+}(t_*), \qquad i_k = i_1, \cdots, i_{n_2}. \tag{3.42}$$

The necessary conditions for stationarity are supplemented by Weierstrass' necessary conditions for a strong relative minimum. The latter are given by the following inequalities involving the Hamiltonians:

$$H(z, \mathrm{Z}, U, U_*) \leqq H(z, \zeta, u, u_*), \tag{3.43}$$

$$h(z, V, V) \leqq h(z, V, V_*). \tag{3.44}$$

In the above expressions z, ζ, u, v denote the optimum values of corresponding variables, and Z, U, V denote any set of admissible functions. It is evident that the artificial variables are absent from (3.43) and (3.44). Also, the admissible values of Z corresponding to ζ variables enter into (3.43).

The totality of formulas (3.33)–(3.44) forms the analogue of Pontryagin's maximum principle for our problem.

Here, we notice a considerable difference between optimal control problems for partial differential equations and those for their ordinary differential equation analogues. As a matter of fact, it reveals the fact that the increment of the functional (about its optimal value) due to the variation of control inside some small domain generally depends not only on the value of the variation itself and the domain area, but also on the limiting form of the domain of variation (the limit is being taken when the area tends to zero).

To illustrate this point, consider the following example where the system equation is given by

$$\frac{\partial}{\partial x} u \frac{\partial z}{\partial x} + \frac{\partial}{\partial y} u \frac{\partial z}{\partial y} = 4\pi q.$$

This equation describes the potential distribution in a medium with dielectric permeability $u(x, y)$, induced by sources whose density is $q(x, y)$. Some functional of boundary values of z (or $\partial z/\partial n$) is minimized under definite boundary conditions.

Let $q(x, y)$ play the role of distributed control and let $u(x, y)$ be continuously differentiable. It is obvious that the increment of the functional due to the variation of $q(x, y)$ is completely determined by that of the total charge in the domain of variation provided that the area of this domain is sufficiently small. The charge is a scalar, and its increment depends only on that of the density $q(x, y)$ and on the area of the domain of variation and does not depend on its limiting form. One can see that under these conditions, the ζ variables disappear from the expression for the Hamiltonian (corresponding terms vanish in accordance with the Euler equations). The situation changes substantially if we choose the dielectric permeability $u(x, y)$ as the control function. The increment of the functional is now determined by the dipole moment of polarization charges disposed along the boundaries of the domain of variation (with sufficiently small area). The dipole moment is a vector, which depends essentially on the orientation or the form of boundaries of the domain of variation. Performing the comparison, one should take all possible directions of the dipole moment into account. This can be achieved provided the variation is accomplished inside a narrow strip (whose width tends to zero), and this strip is inclined under different possible angles with respect to certain fixed directions in the (x, y)-plane. The expression for the Weierstrass function (coinciding with the difference of Hamiltonians entering (3.43)) essentially depends on the admissible values Z of the ζ variable. There is, however, a natural way of elimination: we use the condition that z should be continuous across the boundary of the domain of variation. The same must be true for their tangential derivatives; this leads to the relationship

$$\begin{aligned} X_i(z, \mathrm{Z}, U; x, y)x_t + Y_i(z, \mathrm{Z}, U; x, y)y_t \\ = X_i(z, \zeta, u; x, y)x_t + Y_i(z, \zeta, u; x, y)y_t . \end{aligned} \tag{3.45}$$

We denote the direction cosines of the tangent to the boundary of the domain of variation (this boundary is assumed to be smooth) by x_t and y_t $(x_t^2 + y_t^2 = I)$. The values of these cosines may be arbitrary if no additional restrictions are imposed.

Equation (3.45) allows us to express the Z variable in terms of ζ, u, U, $x_t(y_t)$. The resulting expression is inserted into the Weierstrass condition, which can be carried out for arbitrary values of $x_t(y_t)$ or arbitrary inclinations of the strip of variation.

The variation of controls performed in a strip cannot, of course, be considered as the only way of variation. It turns out, however, that the strongest conditions of optimality are obtained from this choice of variation. When the controls are varied, say, in some elliptic regions tending to their center with some fixed eccentricity, the resulting optimality conditions will be

weaker. The weakest conditions correspond to the case with zero eccentricity (variation inside a circle, with no dependence on inclination). On the other hand, the strongest conditions correspond to the case with unit eccentricity (limiting case of a strip with maximum dependence on inclination).

In [51] a detailed account of the optimality conditions can be found.

In [18], [19] the foregoing theory is applied to the problem of optimal distribution of conductivity of the working fluid in the channel of a magnetohydrodynamical generator. In this problem, the effect of "anisotropy" just discussed is of great importance, because it leads to the strongest optimality criteria.

It may be added that this effect manifests itself only insofar as one deals with a strong extremum. The conditions for a weak extremum do not depend on the form of the domain of variation, and are of the same form as if the region of variation were circular (it seems reasonable that this is true for a wide class of control problems). This point together with the Weierstrass-Erdmann conditions shows that the basic difference between the solutions which give a strong or weak extremum manifests itself in the structure of the lines of discontinuity. For a weak extremum, these lines are well-behaved, and they acquire very oscillatory shape for a strong extremum.

In conclusion, mention should be made about approximate solutions to optimal control problems. One can indicate two approaches for obtaining such solutions. The first one is that in which optimal control is found approximately using certain conditions of optimality. A more common approach involves simplifying the basic equations and additional conditions from the very beginning. Then conditions of optimality are formulated for the simplified problem, and the corresponding optimal control is treated as an approximation to the exact one [24], [53]–[58]. In [53], for example, the minimum is found among admissible controls which are constant over the rectangles whose sides are parallel to the coordinate axes. A search of optimal division of the fundamental region into such rectangles is then performed. The search procedure is reduced to calculating the increment of the functional by means of variation of controls in strips parallel to the coordinate axes. These directions are by no means special (that is, they may be entirely unconnected with the internal peculiarities of a problem). It may turn out that the approximate control obtained by this approach will not be really approximate for a strong minimum. In [18] and [19], for example, the characteristic directions at any point are those of the electric field lines and the vector lines corresponding to the Lagrange multipliers. A certain condition imposed upon the angle between these directions provides the minimization criterion for arbitrary orientations of a strip of variation. As long as we are going to obtain conditions for a strong minimum, we should

not allow the possibility for an increment of the functional to change its sign. This possibility is however not excluded a priori, if the variation is always performed in strips parallel to the coordinate axes (a similar remark can be made about [20]). There is, of course, a variety of problems where the foregoing difficulty does not arise, namely, linear problems (with control functions entering additively) and other problems free from the influence of "anisotropy of variation".

Practical realization of distributed control systems is difficult, and this is why attempts are made to optimize not with the use of "ideal" controllers, but on the basis of most rational gathering of information about the process [59], [60]. This approach leads to some new mathematical problems which are beyond the scope of this paper.

REFERENCES

[1] A. A. FEL'DBAUM, *The simplest relay systems of automatic control*, Avtomat. i Telemeh., 10 (1949), pp. 249–266.

[2] A. YA. LERNER, *Improvement of the dynamical properties of automatic compensators with the aid of nonlinear couplings*, Ibid., 13 (1952), pp. 134–144, pp. 429–444.

[3] A. A. FEL'DBAUM, *Optimal processes in automatic control systems*, Ibid., 14 (1953), pp. 712–728.

[4] A. YA. LERNER, *On the limiting speed of response of automatic control systems*, Ibid., 15 (1954), pp. 461–477.

[5] A. A. FEL'DBAUM, *On the synthesis of optimal systems with the aid of phase space*, Ibid., 16 (1955), pp. 120–149.

[6] A. YA. LERNER, *Design of time-optimal control systems with constraints on the values of the phase coordinates*, Proc. 2nd All-Union Conference on the Theory of Automatic Control, vol. 2, Academy of Sciences Publishing House, Moscow, 1955, pp. 305–324.

[7] A. A. FEL'DBAUM, *The application of computers in automatic systems*, Avtomat. i Telemeh., 17 (1965), pp. 1046–1056.

[8] A. YA. LERNER, *Introduction to the Theory of Automatic Control*, Mashghiz, Moscow, USSR, 1958.

[9] A. A. FEL'DBAUM, *The Application of Computing Devices to Automation Theory*, Fizmatgiz, Moscow, USSR, 1959.

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

[11] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[12] A. G. BUTKOVSKY AND A. YA. LERNER, *Optimal control in distributed-parameter systems*, Dokl. Akad. Nauk SSSR, 134 (1960), pp. 778–781.

[13] ———, *Optimal control of distributed parameter systems*, Automat. Remote Control, 21 (1960), pp. 472–477.

[14] A. G. BUTKOVSKY, *Discussion of optimal control (in the Section on the Theory of Optimal Systems)*, Automatic and Remote Control, Proc. First Inter-

national Congress of I.F.A.C., Moscow, 1960, vol. 1, J. F. Coales, ed., Butterworths, London, 1961, pp. 545–546.

[15] A. Ya. Lerner, *The use of self-adjusting automatic control systems*, Automatic and Remote Control, Proc. First International Congress of I.F.A.C., Moscow, 1960, vol. 4, J. F. Coales, ed., Butterworths, London, 1961, pp. 226–230.

[16] A. I. Egorov, *Optimal control processes in distributed plants*, J. Appl. Math. Mech., 27 (1963), pp. 1045–1058.

[17] A. G. Butkovsky, *Optimal processes in distributed-parameter systems*, Automat. Remote Control, 22 (1961), pp. 17–26.

[18] K. A. Lurie, *Optimum control of conductivity of a fluid moving in a channel in a magnetic field*, J. Appl. Math. Mech., 28 (1964), pp. 316–327.

[19] ———, *On the problem of optimal conductivity distribution in a fluid moving in a transverse magnetic field*, PMTF Zh. Prikl. Meh. i Tehn. Fiz., 2 (1964), pp. 29–40.

[20] Yu. V. Egorov, *Certain problems in optimal control theory*, USSR Comput. Math. and Math. Phys., 3 (1963), pp. 1209–1232.

[21] A. Ya. Dubovitsky and A. A. Milyutin, *Extremum problems in the presence of constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453.

[22] A. G. Butkovsky, *The maximum principle for optimal systems with distributed parameters*, Automat. Remote Control, 22 (1961), pp. 1288–1301.

[23] ———, *The broadened principle of the maximum for optimal control problems*, Ibid., 24 (1963), pp. 292–304.

[24] ———, *The method of moments in the theory of optimal control of systems with distributed parameters*, Ibid., 24 (1964), pp. 1106–1113.

[25] N. I. Ahiezer and M. G. Krein, *Some Questions in the Theory of Moments*, Translated Mathematical Monograph, vol. 2, American Mathematical Society, Providence, 1962.

[26] A. A. Fel'dbaum, *Optimal Control Systems*, Academic Press, New York, 1965.

[27] A. G. Butkovsky, *Theory of Optimal Control for Distributed-Parameter Systems*, Nauka, Moscow, 1965.

[28] A. G. Butkovsky and L. N. Poltavsky, *Optimal control of a distributed oscillatory system*, Automat. Remote Control, 26 (1965), pp. 1835–1848.

[29] ———, *Optimal control of wave processes*, Ibid., 27 (1966), pp. 1542–1547.

[30] ———, *Optimal control of a two-dimensional distributed oscillatory system*, Ibid., 27 (1966), pp. 553–563.

[31] L. I. Rozoner, *The maximum principle of L. S. Pontryagin in the theory of optimal systems, I, II, III*, Ibid., 20 (1959), pp. 1288–1320, 1405–1441, 1516–1561.

[32] T. K. Sirazetdinov, *On the theory of optimal processes with distributed parameters*, Ibid., 25 (1964), pp. 431–463.

[33] ———, *The Pontryagin maximum principle in the theory of linear optimal processes with distributed parameters*, Trudy Kazan. Aviatsionnogo Instit., 80 (1963), pp. 51–63.

[34] A. I. Egorov, *Optimal processes in distributed-parameter systems and certain problems of invariance theory*, Izv. Akad. Nauk SSSR. Ser. Mat., 29 (1965), pp. 1205–1260.

[35] ———, *Optimal processes in distributed-parameter systems and certain problems of invariance theory*, Mat. Sb., 69 (1966), pp. 371–421.

[36] C. Miranda, *Equazioni alle Derivate Parziali di Tipo Ellitico*, Springer, Berlin, 1955.

[37] O. A. OLEINIK, *Boundary value problems for linear elliptic and parabolic equations with discontinuous coefficients*, Amer. Math. Soc. Trans. (2), 42 (1964), pp. 175–194.

[38] A. I. EGOROV, *Optimal control of processes in certain distributed-parameter systems*, Automat. Remote Control, 25 (1964), pp. 557–566.

[39] ———, *Optimal processes in systems containing distributed parameter plants, I, II*, Ibid., 26 (1965), pp. 972–988, 1178–1187.

[40] T. K. SIRAZETDINOV, *Analytic design of regulators in processes having distributed parameters*, Ibid., 26 (1965), pp. 1449–1457.

[41] ———, *On optimal control of stochastic processes with distributed parameters*, Izv. Vyssh. Uchebn. Zaved. Aviacion. Tehn., 3 (1965), pp. 38–45.

[42] V. L. VOLKOVICH AND YU. I. SAMOILENKO, *Complex Control Systems*, Collected Papers, Naukova Dumka Publishing House, Kiev, 1965.

[43] A. G. SENIN, *Statistical problem of synthesis for plants with distributed-parameters*, Automat. Remote Control, 25 (1964), pp. 593–600.

[44] M. I. FORTUS, *The extrapolation of a random field satisfying the wave equation*, Theor. Probability Appl., 27 (1963), pp. 204–207.

[45] YU. I. SAMOILENKO AND A. A. ZYUZIN-ZINCHENKO, *Distributed reception of information*, Telecommunications and Radio Engineering, 21 (1966), pp. 117–119.

[46] A. G. SENIN, *Some questions in the analysis and synthesis of measuring systems for extrapolating a random field*, Ivz. Sibirsk. Otd. Akad. Nauk SSSR Ser. Tehn. Nauk, 1964, no. 1, pp. 54–62.

[47] Z. I. KHALILOV, *A linear control problem in a Banach space*, Dokl. Akad. Nauk SSSR, 155 (1964), pp. 767–770.

[48] YU. V. EGOROV, *Optimal control in Banach space*, Soviet Math. Dokl., 150 (1963), pp. 630–633.

[49] A. YA. DUBOVITSKY AND A. A. MILYUTIN, *Extremum problems with constraints*, Ibid., 149 (1963), pp. 452–455.

[50] ———, *Second variations in extremal problems with constraints*, Ibid., 160 (1965), pp. 12–16.

[51] K. A. LURIE, *The Mayer-Bolza problem for multiple integrals and the optimization of the performance of systems with distributed-parameters*, J. Appl. Math. Mech., 27 (1963), pp. 1284–1299.

[52] P. K. RASHEVSKY, *Geometrical Theory of Partial Differential Equations*, Gostekhizdat, Moscow, 1947.

[53] YU. M. VOLIN AND G. M. OSTROVSKII, *A method of successive approximations for calculating optimal modes of some distributed-parameter systems*, Automat. Remote Control, 26 (1965), pp. 1188–1195.

[54] T. ABDIKERIMOV, *Optimal processes in certain discrete systems with distributed-parameters*, Ibid., 26 (1965), pp. 214–221.

[55] Z. I. VOSTROVA, *Optimal processes in sampled-data systems containing objects with distributed parameters*, Ibid., 27 (1966), pp. 767–779.

[56] M. P. LEONCHUK, *Numerical solution of problems of optimal processes with distributed parameters*, USSR Comput. Math. and Math. Physics, 4 (1964), pp. 189–198.

[57] B. M. RASPOPOV, *Problem of optimally fast response of uncoupled heat and mass transfer processes*, Automat. Remote Control, 26 (1965), pp. 1791–1796.

[58] N. N. GOLUB, *Control of heating "linearly" viscoelastic plates in the case of thermal stress limitations*, Ibid., 27 (1966), pp. 195–205.

[59] F. YA. GIMELSHEIN AND B. N. DEVYATOV, *The problem of optimum estimation of*

the state of controlled chemical engineering processes, Doklady Chemical Technology, 165 (1965), pp. 198–201.

[60] B. N. Devyatov, F. Ya. Gimelshein and G. S. Khorkova, *On the use of distributed control for the design of high-quality devices controlling processes of heat and mass transfer,* Izv. Sibirsk. Otd. Akad. Nauk SSSR, Ser. Tehn. Nauk, 1963, no. 2, pp. 60–77.

[61] N. M. Gyunter, *A Course in the Calculus of Variations,* ONTI, Moscow, USSR, 1941.

[62] E. T. Artikpaev and M. P. Leonchuk, *The application of optimal control theory to heat conduction problems,* Zh. Vychisl. Mat. i Mat. Fiz., 7 (1967), pp. 681–685.

[63] A. I. Egorov, *Necessary conditions for optimality for systems with distributed parameters,* Mat. Sb., 69 (1966), pp. 371–421.

[64] K. A. Lurie, *The Mayer Bolza problem for multiple integrals: some optimum problems for elliptic differential equations arising in magnetohydrodynamics,* Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967.

[65] F. A. Sholokhvich, *On controllability in Hilbert space,* Differencialńye Uravnenija, 3 (1967), pp. 479–484.

[66] V. R. Vinokurov, *Optimal control of processes described by integral equations, I, II, III,* Izv. Vyssh. Uchebn. Zaved. Matematika, 1967, no. 7, pp. 21–33, no. 8, pp. 16–23, no. 9, pp. 16–25.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

COMPLETELY CONTROLLABLE BILINEAR SYSTEMS*

R. E. RINK AND R. R. MOHLER†

Abstract. Sufficient conditions for complete controllability of systems that are bilinear in state and control are established by geometrical arguments. It is seen that bilinear systems generally are more controllable than systems that are jointly linear in control and state. Bilinear control frequently occurs quite naturally, but in other cases it can be implemented to improve controllability of a linear system by varying plant parameters.

1. Introduction. A physical process is considered to be bilinear if it has a mathematical model which is linear in the state variables and in the control variables, but not jointly linear in both; that is, products of state and control variables appear in the system equations.

Many physical processes have natural models which are bilinear. Perhaps the best known is the model for the neutron kinetics of a nuclear reactor [1, p. 16]. The control variable is the multiplication constant of the reactor, and this appears in the neutron-kinetic equations as a parameter which multiplies the neutron density. The same phenomenon occurs in problems of biological population. Also, numerous physiological processes can be described by bilinear models [2]. In other cases, it may be possible to implement a bilinear mode of control.

The general time-invariant bilinear system with n state variables and m control variables is written compactly as the vector equation:

$$\dot{x} = (A + \sum_{k=1}^{m} u_k B_k)x + Cu, \tag{1}$$

where A is the $n \times n$ matrix of real constants a_{ij}, C is the $n \times m$ matrix of the real constants c_{ik}, and B_k is the $n \times n$ matrix of the real constants b_{ijk} for fixed k.

It is assumed that the state vector x is unconstrained and that the class of allowable control policies $\{u(t)\}$ is the class of all piecewise continuous vector time functions with domain $[0, \infty)$ and range Ω, where Ω is a compact, connected set containing the origin in R^m.

Before proceeding with the controllability analysis of bilinear systems, several necessary terms are recalled. First a dynamical system is said to be completely controllable if it can be transferred from any initial state $x^0 \in R^n$ to any prescribed terminal state $x^1 \in R^n$ by some admissible control $u(t)$.

* Received by the editors April 28, 1967, and in revised form January 29, 1968.

† Department of Electrical Engineering, University of New Mexico, Albuquerque, New Mexico 87106. This work was supported by the National Science Foundation under Grant GK-1173.

The reachable zone from an initial state x^0, $R(x^0) \subset R^n$, is the set of all states to which the system can be transferred in finite time, starting at x^0. Similarly, the incident zone to a terminal state x^1, $I(x^1) \subset R^n$, is the set of all initial states from which x^1 is reachable in finite time.

The necessary and sufficient conditions for a linear system to be completely controllable are well known [3]. These conditions, however, only apply when the control variables are unconstrained. In fact, linear systems are almost never completely controllable when the control vector is constrained to a compact set. (An exception is the second order harmonic oscillator, for which any state can be achieved with a sufficiently large number of control changes.)

Turning now to bilinear systems, this fundamental difficulty disappears. For each fixed $u \in \Omega$, the bilinear system is a constant-parameter linear system with system matrix $A + \sum_{k=1}^{m} u_k B_k$. The terms $\sum_{k=1}^{m} u_k B_k$ in the system matrix permit manipulation of the eigenvalues of the fixed-control system. With an appropriate controller it is often possible to shift these eigenvalues from the left half complex plane to the right half plane.

2. Complete controllability. The controllability analysis presented here can be summarized by the following sufficient conditions.

MAIN RESULT. The bilinear system (1) is completely controllable if:

C1. There exist control values u^+ and u^-, such that the real parts of the eigenvalues of the system matrix are positive and negative, respectively, and such that equilibrium states $x^e(u^+)$, $x^e(u^-)$ are contained in a connected component of the equilibrium set;

C2. For each x in the equilibrium set with an equilibrium control, $u^e(x) \in \Omega$ such that $f(x, u^e(x)) = 0$, there exists a $v \in R^m$ such that g lies in no invariant subspace of dimension $\leqq n - 1$ of the matrix E, where

$$E = A + \sum_{k=1}^{m} u_k{}^e(x) B_k \tag{2}$$

and

$$g = Cv - \sum_{l=1}^{m} v_l [B_l (A + \sum_{k=1}^{m} u_k B_k)^{-1} Cu]. \tag{3}$$

Remark 1. For phase-variable systems, $x_1 = x$, $x_2 = \dot{x}$, $\cdots$, $x_n = x^{(n-1)}$, condition C2 is always satisfied if C is a nonzero matrix.

Remark 2. Condition C1 is satisfied if all the eigenvalues of the system matrix $A + \sum_{k=1}^{m} u_k B_k$ can be shifted across the imaginary axis of the complex plane without passing through zero, as u ranges continuously over a subset of Ω.

It is the objective of this section to substantiate these statements. First suppose there exists a fixed control value u^- in the interior of Ω such that the

eigenvalues of the system matrix $A + \sum_{k=1}^{m} u_k^- B_k$ all have negative real parts. Then the trajectories of the system with constant control $u = u^-$ cover all of R^n, and each trajectory approaches the unique equilibrium state

$$x^e(u^-) = -(A + \sum_{k=1}^{m} u_k^- B_k)^{-1} C u^-. \tag{4}$$

For any initial state $x^0 \in R^n$, the unique trajectory passing through x^0 with control u^- reaches any neighborhood of $x^e(u^-)$ in finite time.

Suppose, also, that there exists a fixed control value u^+ in the interior of Ω such that the eigenvalues of the matrix $A + \sum_{k=1}^{m} u_k^+ B_k$ all have positive real parts. Then the trajectories of the system with constant control $u = u^+$ cover all of R^n, and each trajectory corresponds to motion away from the unique equilibrium state

$$x^e(u^+) = -(A + \sum_{k=1}^{m} u_k^+ B_k)^{-1} C u^+. \tag{5}$$

For any terminal state x^1, the unique trajectory passing through x^1 with control u^+ reaches x^1 from any neighborhood of $x^e(u^+)$ in finite time.

If u^- and u^+ both exist, and if every point of some neighborhood of $x^e(u^+)$ can be reached from every point of some neighborhood of $x^e(u^-)$, then certainly any terminal state x^1 can be reached from any initial state x^0, and the system is completely controllable. Such equilibrium sets and their connectedness are described in the Appendix for bilinear systems. Then the controllability analysis can be completed by means of a local controllability theorem due to Lee and Markus [4], which provides sufficient conditions for local controllability of a system in a neighborhood of its equilibrium set.

THEOREM. *Consider*

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

where $f(x, u) \in C^1$ *in* $R^n \times \Omega$. *If, for* $x \in R^n$, *there exists an equilibrium control value* $u^e(x)$ *in the interior of* Ω *such that* (i) $f(x, u^e(x)) = 0$, *and* (ii) *there exists a* $v \in R^m$ *such that* Dv *lies in no invariant subspace of* E *of dimension* $\leqq (n - 1)$, *where*

$$D = \frac{\partial f}{\partial u}(x, u^e(x)) \quad \text{and} \quad E = \frac{\partial f}{\partial x}(x, u^e(x))$$

are real matrices, then $R(x)$ *and* $I(x)$ *are open connected subsets of* R^n.

If the equilibrium set is a connected proper subset of R^n and condition (ii) above is satisfied for every interior $u \in \Omega$, then the reachable zone from any equilibrium state x corresponding to an interior control value $u^e(x)$

includes every other equilibrium state corresponding to an interior control value in Ω. For, if $R(x)$ has an equilibrium state y on its boundary, then $I(y)$ must be disjoint from $R(x)$. But, by the Lee-Markus theorem, $I(y)$ is an open neighborhood of y if $u^e(y)$ is in the interior of Ω. Hence, $u^e(y)$ must be on the boundary of Ω.

For the general bilinear system, the matrix D is

$$D = \frac{\partial f}{\partial u}(x, u^e(x)) = [B_1 x \mid B_2 x \mid \cdots \mid B_m x] + C, \tag{6}$$

and E is defined by (2). Substitution of the expression for the equilibrium state (see (A1) in the Appendix) into (2) yields the Lee-Markus theorem for a bilinear system: *The system* (1) *is locally controllable at the equilibrium state corresponding to an interior* $u \in \Omega$ *if there exists a* $v \in R^m$ *such that*

$$-\sum_{l=1}^{m} v_l[B_l(A + \sum_{k=1}^{m} u_k B_k)^{-1} Cu] + Cv$$

lies in no invariant subspace of dimension $\leqq n - 1$ *of the matrix* E.

The practical implications of using this criterion for a general bilinear system appear quite formidable. However, it will now be shown that the application to systems of the phase-variable type is straightforward, and in fact such systems always satisfy the criterion when $C \neq 0$.

The matrix D for the phase-variable system is obtained by substituting the equilibrium-state expression (see (A6)) into (6). Hence,

$$D = \frac{-\sum_{k=1}^{m} c_k u_k}{a_1 + \sum_{k=1}^{m} u_k b_{1k}} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & & \vdots \\ \vdots & \vdots & & \vdots \\ b_{11} & b_{12} & \cdots & b_{1m} \end{bmatrix} + C. \tag{7}$$

The matrix E is simply of the canonical phase-variable form.

The criterion to be satisfied is that a $v \in R^m$ is to be found such that the vectors Dv, EDv, $\cdots$, $E^{(n-1)}Dv$ are linearly independent. By inspection of (7), if D is not identically zero it has nonzero entries only in the bottom row, and there exists a v such that Dv has the nth component nonzero and all others zero, E^2Dv has the $(n-2)$th component nonzero, etc. Such a set is certainly linearly independent.

There remains the question of whether D is identically zero. By inspection of (7), if $C = 0$ then $D = 0$ and the criterion is not satisfied. Suppose $C \neq 0$, and for the particular value of u under consideration, $\sum_{k=1}^{m} c_k u_k = 0$. Then $D = C$ and the criterion is satisfied. Finally, if $\sum_{k=1}^{m} c_k u_k \neq 0$, select

$v = u$, and then

$$Dv = Du = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \sum_{k=1}^{m} c_k u_k \left(1 - \dfrac{\sum_{k=1}^{m} b_{1k} u_k}{a_1 + \sum_{k=1}^{m} u_k b_{1k}} \right) \end{bmatrix}.$$

This is not identically zero, since A nonsingular implies $a_1 \neq 0$. Therefore the Lee-Markus criterion is always satisfied if C is not identically zero for phase-variable systems.

As observed in the Appendix for phase-variable systems, condition C1 of the main result may be satisfied even if one or more eigenvalues pass through zero. On the other hand, this condition cannot be satisfied with scalar control ($m = 1$) in a state-space of odd dimension n. For then an odd number of eigenvalues must be shifted across the imaginary axis, and at least one of these must pass through zero, since at most $(n - 1)/2$ can cross as complex-conjugate pairs. In the Appendix it is shown that for $m = 1$, however, the branches of the equilibrium curve are disjoint.

3. Examples.

3.1. Example 1. It is immediately obvious that the second order system

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

$$\dot{x}_2 = x_1 - x_2 + u$$

is completely controllable for unbounded control. For bounded control, however, this system is only locally controllable in some vicinity of the origin. If the process admits an appropriate bilinear mode of control, the system is completely controllable since the Lee-Markus criterion is already satisfied. For example, consider the bilinear system

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

$$\dot{x}_2 = -x_1 - (1 + 4u)x_2 + u,$$

where $|u| \leqq 1$ has the eigenvalues $\lambda_{1,2}(u) = -(1 + 4u)/2 \pm \frac{1}{2}\sqrt{(1 + 4u)^2 - 4}$ for constant u. For $u = -1$ both are real and positive, crossing the imaginary axis as a complex conjugate pair. Thus, u^+ and u^- exist and the equilibrium set is connected. The system is of the phase-variable type and $C \neq 0$, so that the Lee-Markus criterion is satisfied. Thus the new system is completely controllable.

As shown by the next example, however, not all bilinear systems with the capability of transferring system eigenvalues across the imaginary axis are completely controllable. Such is the case even though the system satisfies the Lee-Markus criterion.

3.2. Example 2. The bilinear system

$$\dot{x}_1 = 2ux_1 + x_2\,,$$

$$\dot{x}_2 = x_1 + 2ux_2 + u$$

has eigenvalues $\lambda_1(u) = 2u + 1$, $\lambda_2(u) = 2u - 1$. If the constraint set Ω is given by $|u| \leqq 1$, then the values u^+ and u^- exist. It is easily verified that the Lee-Markus criterion is satisfied for this system. The equilibrium set, however, is not connected, since $m = 1$ and both eigenvalues pass through zero as they cross the imaginary axis in the complex plane. Therefore, the criteria for complete controllability are not all satisfied.

The phase-plane portraits for $u = +1$ and $u = -1$ are shown superimposed in Fig. 1. The solid trajectories are for the system with $u = -1$, which has a stable node at $x_1 = -\frac{1}{3}$, $x_2 = -\frac{2}{3}$. The dashed trajectories are for the system with $u = +1$, which has an unstable node at $x_1 = \frac{1}{3}$, $x_2 = -\frac{2}{3}$. The equilibrium set is shown as the three heavy solid curves.

The reachable zones from various initial phases are easily determined on this figure by simply considering the directions of allowable motion at each point, which is the cone between the extremal directions corresponding to $u = +1$ and $u = -1$. It is evident that the system is not completely controllable. For example, the reachable zone from the point y is just the shaded region.

4. Conclusions. A bilinear system is completely controllable if the conditions given by the main result are satisfied. As one might expect, the conditions are not as simple as the popular conditions for complete controllability of linear systems with unconstrained control or for null controllability with constrained control. For phase-variable systems, however, the sufficient conditions are easy to apply.

In practice, a bilinear mode may be implemented by controlling significant plant parameters in a manner similar to the variable wing geometry of high-performance aircraft. For Example 1, a simple bilinear control made a locally controllable linear system completely controllable. With respect to many systems which are inherently bilinear, however, controllability is further complicated by numerous state constraints. Unfortunately, state constraints can appear in such diversity that meaningful controllability conditions would have to be specified for particular cases. For the classical neutron kinetics of a reactor [1], it is obvious by inspection

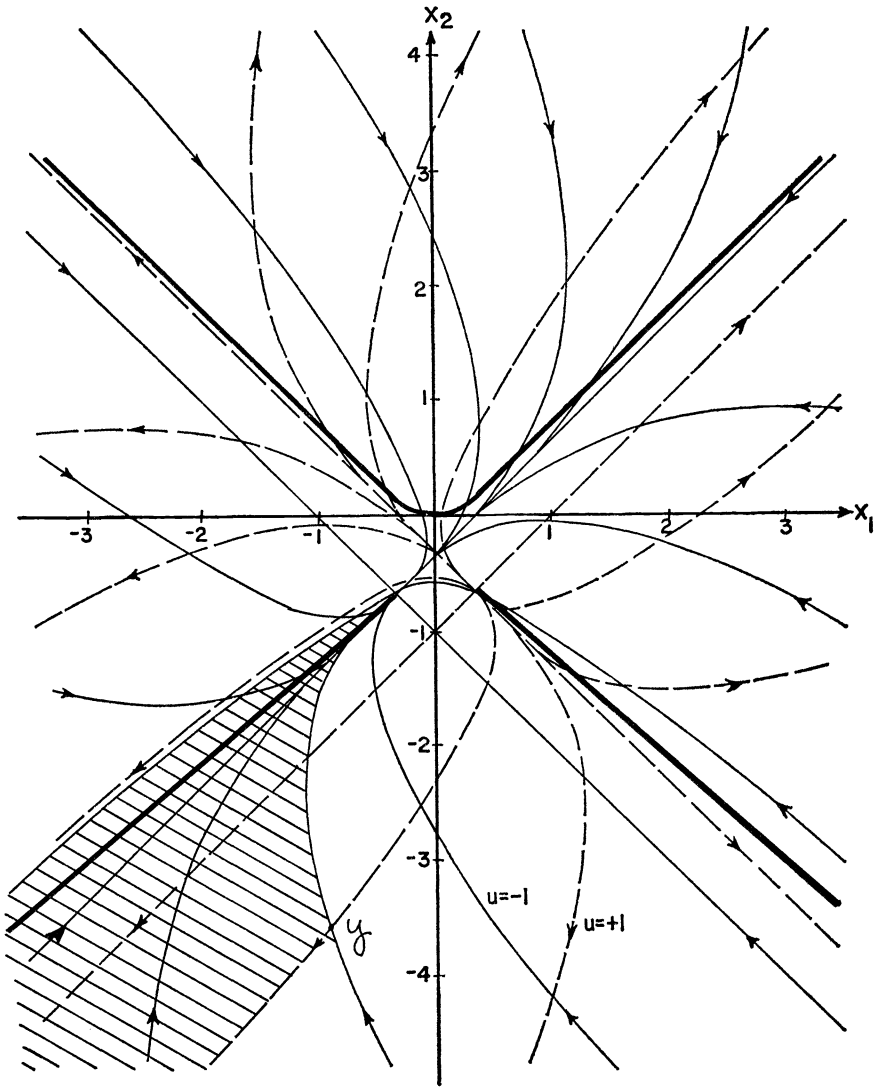


FIG. 1. *Phase-plane trajectories of Example 2 for* $u = +1$ *and* $u = -1$

of the model that every desired equilibrium state can be attained in finite time. Again such controllability is a consequence of the bilinear control mode.

Appendix. Equilibrium set of bilinear systems. In this Appendix the set of equilibrium points is described for bilinear systems. This description is

necessary to understand the connectedness property that is utilized to show complete controllability. For each fixed $u \in \Omega$, the state

$$x^e(u) = -(A + \sum_{k=1}^{m} u_k B_k)^{-1} Cu \tag{A1}$$

is the unique equilibrium state of the bilinear system (1) if the indicated inverse matrix exists. If the system matrix is singular for a control value $u = \hat{u}$, no equilibrium state exists corresponding to $\hat{u}$ unless $C\hat{u}$ happens to lie in the range of $A + \sum_{k=1}^{m} u_k B_k$, in which case an infinite number of stages are equilibrium states. It will be assumed here that Cu is not contained in the range of $A + \sum_{k=1}^{m} u_k B_k$ for any $u \in \Omega$ such that the latter matrix is singular. This implies, in particular, that A is nonsingular, since if A is singular, $C0 = 0$ is in the range of the singular matrix $A + \sum_{k=1}^{m} 0B_k$. It also implies that C is not identically zero if $A + \sum_{k=1}^{m} u_k B_k$ is singular for any $u \in \Omega$. With this assumption, (A1) defines a mapping from $\Omega - S$, where S is the subset of Ω on which the system matrix is singular, onto a subset of R^n, called the equilibrium set.

The equilibrium set is easily described *when u is a scalar*. Since $\det(A + uB)$ is a polynomial in u with degree at most n, the equation $\det(A + uB) = 0$ has at most n real roots $\{u^i\}$ in Ω. As u ranges over any interval in Ω not containing a root, the equilibrium points $x^e(u) = -u(A + uB)^{-1}C$ sweep out a smooth curve. As u approaches a value $u^1 \in \Omega$ for which $A + u^1 B$ is singular, the curve $x^e(u)$ tends to infinity asymptotic to the null space of $A + u^1 B$.

To prove this assertion, let $\{y_1, y_2, \cdots, y_r\}$ be a basis of the null space of $A + u^1 B$, and let $\{y_1, y_2, \cdots, y_r, y_{r+1}, \cdots, y_n\}$ be a basis of R^n. Then one has the unique representation

$$x^e(u) = \xi_1(u)y_1 + \xi_2(u)y_2 + \cdots + \xi_n(u)y_n \tag{A2}$$

for every $\{u^i\}$. The equation $(A + uB)x^e(u) = -uC$ must be satisfied, but as $u \to u^1$ the left-hand side has $\xi_i(u)(A + uB)y_i \to 0$ for $i = 1, 2, \cdots, r$ unless $|\xi_i(u)| \to \infty$. Since by assumption $u^1 C$ is not contained in the range of $A + u^1 B$, one or more of the functions $\xi_1(u), \xi_2(u), \cdots, \xi_r(u)$ must tend to infinity as $u \to u^1$. Thus, at each of the control values $\{u^i\}$, the otherwise smooth curve $x^e(u)$ blows up, and the equilibrium set $\{x^e(u)\ u \in \Omega - S\}$ consists of at most $n + 1$ smooth curves having no finite endpoints, with the possible exception of the equilibrium states corresponding to the minimum and maximum values of u in Ω.

Further, it will now be shown that these curves do not intersect one another. For, if there exists a common point $x^* = x^e(u^a) = x^e(u^b)$, with $u^a \neq u^b$, then

(A3) $$Ax^* + u^aBx^* + u^aC = Ax^* + u^bBx^* + u^bC = 0,$$

which implies that $(u^a - u^b)(Bx^* + C) = 0$. Thus, from (A3), $Ax^* = 0$, and since x^* is not the zero vector, A must be singular, contrary to the previous assumption. Therefore, the curves do not intersect one another, but constitute an unconnected set. Of course, if $\{u^i\}$ is empty, the equilibrium set is compact and connected.

Extending these considerations to the case where $m > 1$, suppose that $u^0 \in \Omega - S$ is a control value for which the mapping (A1) exists. There is an R^m-neighborhood of u^0 on which (A1) exists, since $\det (A + \sum_{k=1}^m u_k{}^0 B_k) \neq 0$ and the determinant of a matrix is a continuous function of the matrix entries. Hence (A1) defines a continuous mapping of an R^m-neighborhood u^0 into an R^n-neighborhood of $x^e(u^0)$, which geometrically corresponds to an equilibrium surface of dimension $d \leqq \min(m, n)$ passing through $x^e(u^0)$. As in the scalar case, this surface will tend to infinity as u approaches a vector-value u^1 for which $A + \sum_{k=1}^m u_k B_k$ is singular.

If $\det (A + \sum_{k=1}^m u_k B_k)$ is thought of as a function of a single component of the control vector, say u_1, with the remaining $m - 1$ components fixed, then

(A4) $$P(u_1 ; u_2{}^0, \cdots, u_m{}^0) = \det \left(A + u_1 B_1 + \sum_{k=2}^m u_k{}^0 B_k\right)$$

is a polynomial of degree at most n in u_1, and has at most n real roots $\{u_1{}^i(u_2{}^0, \cdots, u_m{}^0)\}$ in Ω. Since the roots of a polynomial vary continuously with the polynomial coefficients, as $u_2, u_3, \cdots, u_m$ range in a small neighborhood about the fixed values $u_2{}^0, u_3{}^0, \cdots, u_m{}^0$, each root $u_1{}^i$ of (A4) varies continuously and describes an $(m - 1)$-dimensional root-surface in Ω. (This statement needs qualification for points $(u_2{}^0, u_3{}^0, \cdots, u_m{}^0)$ at which the number of real roots of (A4) changes. Here two root surfaces intersect and the corresponding roots are complex at certain points in every neighborhood of $(u_2{}^0, \cdots, u_m{}^0)$.) In any event, whether the root surfaces intersect or not, they partition Ω into a number of connected cells, each of which is mapped by (A1) into a smooth, connected equilibrium surface in state-space. Unlike the scalar case, these equilibrium surfaces are not necessarily disjoint, as will now be demonstrated for the phase-variable type of bilinear system.

A phase-variable bilinear system is one governed by a single differential equation

$$\frac{d^n x}{dt^n} = \sum_{i=1}^n \left(a_i + \sum_{k=1}^m u_k b_{ik}\right) \frac{d^{(i-1)} x}{dt^{(i-1)}} + \sum_{k=1}^m c_k u_k .$$

With the usual phase variables $x_1 = x, x_2 = \dot{x}, \cdots, x_n = x^{(n-1)}$, the system

matrix is singular for control vectors which satisfy

$$\text{(A5)} \qquad \det\left(A + \sum_{k=1}^{m} u_k B_k\right) = (-1)^{n-1}\left(a_1 + \sum_{k=1}^{m} u_k b_{1k}\right) = 0.$$

For this type of system, it is easily verified that the mapping (A1) reduces to

$$\text{(A6)} \qquad x^e(u) = \frac{-\sum_{k=1}^{m} c_k u_k}{a_1 + \sum_{k=1}^{m} u_k b_{1k}} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

which just expresses the trivial fact that the phase-variable system is at equilibrium only when the dependent variable x is constant and all its derivatives are zero.

The mapping (A6) fails to exist when (A5) is satisfied, which can happen for at most a single value of u_1 with given values $u_2, u_3, \cdots, u_m$, due to the linear nature of (A5). Thus Ω is partitioned into at most two connected cells by the root-surface of (A5), each of which is mapped onto an interval of the x_1-axis by (A6). If (A5) has no roots $u \in \Omega$, then Ω is mapped onto a single compact interval of the x_1-axis by (A6).

Suppose Ω is partitioned by a root-surface S into Ω^+ and Ω^- with the denominator of (A6) positive in Ω^+ and negative in Ω^-. The $\{x^e(u) \mid u \in \Omega^+\}$ and $\{x^e(u) \mid u \in \Omega^-\}$ overlap (in fact, each is the entire x_1-axis) if the surface T defined by $\sum_{k=1}^{m} c_k u_k = 0$ intersects S at a nonzero angle in the interior of Ω. For, in this case, there exists a $\rho > 0$ such that, for any r with $|r| \leqq \rho$, the closure of Ω^+ and the closure of Ω^- contain sets U_r^+ and U_r^- which intersect S and on which $\sum_{k=1}^{m} c_k u_k = r$. Upon inspection of (A6), it is clear that

$$\{x^e(u) \mid u \in \bigcup_{|r| \leqq \rho} U_r^+\} \quad \text{and} \quad \{x^e(u) \mid u \in \bigcup_{|r| \leqq \rho} U_r^-\}$$

both cover the entire x-axis in R^n.

REFERENCES

[1] M. Ash, *Nuclear Reactor Kinetics*, McGraw-Hill, New York, 1965.

[2] F. S. Grodins, *Control Theory and Biological Systems*, Columbia University Press, New York, 1963.

[3] R. E. Kalman, Y. C. Ho and K. S. Narendra, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.

[4] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

SIAM J. CONTROL
Vol. 6, No. 3, 1968
Printed in U.S.A.

LOGARITHMIC VARIATION CRITERIA FOR THE STABILITY OF SYSTEMS WITH TIME-VARYING GAINS*

M. FREEDMAN† AND G. ZAMES‡

Summary. A new stability condition is derived for a feedback system consisting of a linear element $H(s)$ and a time-varying gain $k(t)$. It is shown that if $k(t)$ lies in an interval $[a + \epsilon, b - \epsilon]$, and if the shifted Nyquist diagram $H(j\omega - \sigma_{sh})$ does not intersect a critical interval of the complex plane and encircles it a number of times equal to the number of poles of $H(s)$ in the region Re $\{s\} > -\sigma_{sh}$, then the closed-loop system is stable, provided that the average logarithmic variation $\hat{k}$ satisfies the inequality

$$\hat{k} \triangleq \sup_{t \geq 0} \frac{1}{T} \int_t^{t+T} \left| \frac{d}{d\tau} \log \frac{k(\tau) - a}{b - k(\tau)} \right| d\tau < 4\sigma_{sh}$$

for some $T > 0$. Several alternative statements of this result, including a root locus interpretation, are given.

The proof, which involves a factorization of the open loop into positive operators, depends on a lemma on operators having prescribed phase characteristics, and on another lemma on the factorization of time-varying gains. As a by-product of the theory, it is shown that the property of being stable for all positive constant gains is equivalent to a factorizability property of the open loop.

1. Introduction. The behavior of a time-varying system over a succession of points in time can often be approximated by the behavior of a sequence of time-invariant systems, and methods of analysis depending on this fact are sometimes called "frozen-time" methods. In this paper some idea will be obtained of the range of validity of frozen-time approximations in a problem involving the stability of the feedback system of Fig. 1, where $H(s)$ represents a linear time-invariant operator and $k(t)$ is a time-varying gain.

For the system of Fig. 1, one might ask: If the system is stable for every constant gain in an interval (a, b) is it also stable for gains in (a, b) that are time-varying? This problem plays a role in the theory of time-varying systems rather similar to that of the Aizerman problem in the theory of nonlinear systems.

It seems intuitively reasonable that the system should remain stable for gains in (a, b) if the rate of time variation is slow enough. More generally, it would seem likely that validity of a frozen-time approximation ought to depend on suitable measures of the rates of time-variation and decay of

* Received by the editors November 8, 1967, and in revised form February 29, 1968.

† National Aeronautics and Space Administration, Electronics Research Center, Cambridge, Massachusetts 02139.

‡ Guggenheim Fellow, Office of Control Theory and Application, National Aeronautics and Space Administration, Electronics Research Center, Cambridge, Massachusetts 02139.

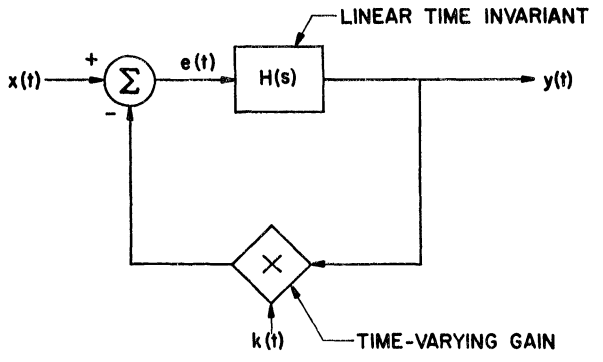


FIG. 1. *A feedback system*

memory. In the system of Fig. 1, exponential bounds on memory (i.e., on the impulse response) can be obtained from the geometry of the "shifted Nyquist diagram" $H(j\omega - \sigma_{sh})$ for various "shifts" $\sigma_{sh} > 0$. It is natural, therefore, to seek stability criteria involving $H(j\omega - \sigma_{sh})$ and some measure of the rate of change of $k(t)$. Here, some new stability criteria will be derived along these lines.

There now is available a large body of stability results for the system of Fig. 1. Many of the early results, such as the inequality of Bongiorno [1] and the circle criterion[1] of Zames [2a] employ no information about the gain $k(t)$ other than its range of values. Consequently, they offer no insight into the effects of "freezing time." However, they have the advantage of involving only the geometry of a Nyquist plot and hence are easy to check.

There are several results which go beyond [1] and [2a] in explicitly restricting the rate of change of the gain $k(t)$. For example, the condition of Brockett and Forys [3], generalized by Gruber and Willems [4], includes a restriction of the type $\dot{k}/k \leqq$ const., and an "RC" multiplier

$$Z(s) = \sum_i \frac{s + a_i}{s + b_i}$$

Zames [2b] obtained a condition involving exponential weighting factors, a shifted Nyquist diagram, and a factorization condition on $k(t)$ which, in effect, limits its rate of change. All of the previously mentioned conditions either make use of a multiplier for which no explicit construction is given, or do not reduce to the frozen-time Nyquist criterion[2] in the limit as the

[1] Similar criteria were derived independently by Narendra and Goldwyn [10], Sandberg, and J. Kuderewicz.

[2] By frozen-time Nyquist criterion the following hypothesis (which is not true in general) is meant: If the open-loop system is stable and if the Nyquist diagram of $H(j\omega)$ does not intersect or encircle an interval of the real axis lying between $-1/a$ and $-1/b$, then the closed-loop system is stable.

For a derivation of the ordinary Nyquist criterion, see Desoer [5].

rate of variation goes to zero. (For example, in the circle criterion the critical region does not approach a linear interval.)

In the present paper, although certain multipliers are employed much as in [2a] and [2b], a constructive procedure is developed so that all final stability conditions are free of multipliers, and are, in fact, explicit and geometric. The conditions on the gain $k(t)$ take the form of bounds on the average of $\log [k(t)]$. The idea of imposing an average variation constraint on $k(t)$ has been used before (for example, see Sandberg [6]). However, it is shown here that criteria of this sort are naturally expressed in terms of the logarithm of $k(t)$. In fact, the logarithmic variation is shown to be related to certain factorizability properties of $k(t)$.

Finally, a by-product of the approach taken here is the result that a system which is stable for all positive constant gains can be characterized by a factorizability property of its open loop operator.

The remainder of the paper is divided into the following numbered sections:

2. The main problem and its solution.
3. Basic method: A lemma on multipliers.
4. Construction of $M(s)$: A lemma on multipliers with prescribed phase characteristics.
5. Construction of $f(t)$: A factorization lemma for time-varying gains.
6. Proofs of main results.
7. Concluding remarks.

Appendix.

The main stability results are stated in §2. Their proofs are postponed until §6 to allow development of the supporting theory, which is contained in §3, §4 and §5.

2. The main problem and its solution.

DEFINITION 1. Let $L_p[0, \infty]$, where $p = 1, 2, \cdots, \infty$, be the linear space of real-valued functions $x(\cdot)$ on $[0, \infty)$ with the property that

$$\int_0^\infty |x(t)|^p \, dt < \infty \quad \text{if} \quad 1 \leqq p < \infty,$$

or

$$x(\cdot) \quad \text{is essentially bounded if} \quad p = \infty.$$

Let $L_2[0, \infty)$ be normed with the norm

$$\| x(\cdot) \| = \left\{ \int_0^\infty |x(t)|^2 \, dt \right\}^{1/2}.$$

The spaces $L_p(-\infty, \infty)$ on the interval $(-\infty, \infty)$ are similarly defined. The definition of the extended space L_{2e} is introduced via the notion of a

truncated function $x_T(\cdot)$ (see [2c, Part I] for a further discussion of extended spaces).

DEFINITION 2. For any real-valued function $x(\cdot)$ on $[0, \infty)$ and any $T \geqq 0$, let $x_T(\cdot)$ denote the truncated function defined by

$$x_T(t) = \begin{cases} x(t) & \text{for } t \leqq T, \\ 0 & \text{for } t > T. \end{cases}$$

Let L_{2e} be the space of those real-valued functions $x(\cdot)$ on $[0, \infty)$ whose truncations $x_T(\cdot)$ belong to $L_2[0, \infty)$ for all $T \geqq 0$, i.e.,

$$L_{2e} = \{x(\cdot) \mid x(\cdot):[0, \infty) \to \text{Reals}, \text{ and } x_T(\cdot) \in L_2[0, \infty) \text{ for all } T \geqq 0\}.$$

Remark 1. $x(\cdot) \in L_2[0, \infty)$ if and only if $\sup_{T>0} \| x_T(\cdot) \| < \infty$.

2.1. Feedback equations. The feedback system of Fig. 1 will be represented for all $t \geqq 0$ by the integral equations

$$e(t) = x(t) - k(t)y(t), \tag{1}$$

$$y(t) = h_0 e(t) + \int_0^t e(\tau)h_1(t - \tau)\, d\tau, \tag{2}$$

in which the following assumptions are made.

ASSUMPTION 1. $x(\cdot)$ is in $L_2[0, \infty)$. (The function $x(\cdot)$ represents the combined effects of an input and of possible nonzero initial conditions.)

ASSUMPTION 2. $e(\cdot)$ and $y(\cdot)$ are in L_{2e} (i.e., existence of solutions in L_{2e} for these "outputs" is being assumed[3]).

ASSUMPTION 3. h_0 is a real constant, and there is a real constant σ_0 (not necessarily positive) for which $h_1(t) \exp(\sigma_0 t)$ is in $L_1[0, \infty)$.

ASSUMPTION 4. $k(\cdot)$ is a real-valued function, absolutely continuous on $[0, \infty)$. (Since $k(\cdot)$ is absolutely continuous its derivative $\dot{k}(\cdot)$ exists almost everywhere, and $k(b) - k(a) = \int_a^b \dot{k}(t)\, dt$ for any real a and b (see Hobson [7, §406, pp. 592–593]).)

2.2. Shifted Nyquist diagrams. Some of the criteria to be established here will employ a Nyquist diagram plotted on a vertical line outside the region of convergence of the Laplace integral, but inside a region in which the Laplace transform has a suitable "continuation." The definitions of a Laplace transform, its continuation, and its Nyquist diagram will be developed next.

[3] It is convenient to separate questions of stability from those of existence, since the two can frequently be deduced from entirely different considerations.

DEFINITION 3. For any real constant σ_0, let LE_{σ_0} be the set of pairs $[h_0, h_1(t)]$, where h_0 is a real constant, and $h_1(t) \exp(\sigma_0 t)$ is in $L_1[0, \infty)$.

DEFINITION 4. Let s denote a point of the complex plane. The real and imaginary parts of s will sometimes be denoted by σ and ω. For any pair $[h_0, h_1(t)]$ in some LE_{σ_0} the *Laplace transform* $H(s)$ *with domain* $\operatorname{Re}\{s\} \geqq -\sigma_0$ is defined by the equation.

$$H(s) = h_0 + \int_0^\infty h_1(t) \exp(-st)\, dt \tag{3}$$

(for all complex s with $\operatorname{Re}\{s\} \geqq -\sigma_0$) and is analytic for $\operatorname{Re}\{s\} > -\sigma_0$.

Certain additional assumptions will be made concerning $H(s)$ in order to allow the definition of a Nyquist diagram.

DEFINITION 5. A function $H(s)$ is said to be *meromorphic* in a region of the complex plane if it is analytic in the region except, at most, at a countable number of poles. For any real constant σ_{cn}, the *meromorphic continuation of* $H(s)$ *to* $-\sigma_{cn}$, if it exists, is the function $H_{cn}(s)$ with domain $\operatorname{Re}\{s\} > -\sigma_{cn}$ which is meromorphic throughout its domain, and is equal to $H(s)$ whenever s is in the domains of $H(s)$ and $H_{cn}(s)$. (From elementary continuation theory, if $H_{cn1}(s)$ and $H_{cn2}(s)$ are two continuations, then $H_{cn1}(s) = H_{cn2}(s)$ on the intersection of their domains.) If $H_{cn}(s)$ is, in fact, analytic for $\operatorname{Re}\{s\} > -\sigma_{cn}$, then $H_{cn}(s)$ will be called an *analytic continuation of* $H(s)$ *to* $-\sigma_{cn}$.

Two classes of transforms are now introduced for which Nyquist diagrams will later be defined.

DEFINITION 6. For any real constant σ_{cn}, let $Mer_{\sigma_{cn}}$ be the class of functions $H(s)$ having the following properties: (i) $H(s)$ is the Laplace transform of a pair $[h_0, h_1(t)]$ in some LE_{σ_0}; (ii) $H(s)$ has a meromorphic continuation to $-\sigma_{cn}$; and (iii) for any $-\sigma_{sh} > -\sigma_{cn}$, $H(\sigma + j\omega)$ has a finite number of poles in $\operatorname{Re}\{s\} \geqq -\sigma_{sh}$, and approaches h_0 uniformly with σ in $\operatorname{Re}\{s\} \geqq -\sigma_{sh}$ as $|\omega| \to \infty$.

Let $Anl_{\sigma_{cn}}$ be the subset of $Mer_{\sigma_{cn}}$ consisting of those functions that are analytic for $\operatorname{Re}\{s\} > -\sigma_{cn}$.

If $H(s)$ is in $Mer_{\sigma_{cn}}$, then $H(s)$ is analytic on almost every vertical line in $\operatorname{Re}\{s\} > -\sigma_{cn}$.

DEFINITION 7. If $H(s)$ is in $Mer_{\sigma_{cn}}$ and has no poles on some line $\operatorname{Re}\{s\} = -\sigma_{sh}$, then the σ_{sh}-shifted Nyquist diagram of $H(s)$ consists of (i) the directed curve $H_{cn}(j\omega - \sigma_{sh})$ with ω in $(-\infty, \infty)$ directed from $-\infty$ to ∞, and (ii) the point h_0.

2.3. A stability theorem. The main problem here is: To find conditions involving $k(t)$ and a σ_{sh}-shifted Nyquist diagram of $H(s)$ under which the system is stable in the sense that (i) $e(\cdot)$ is in $L_2[0, \infty)$ with $\| e \| \leqq$ const. $\| x \|$, and (ii) $\lim_{t\to\infty} y(t) = 0$.

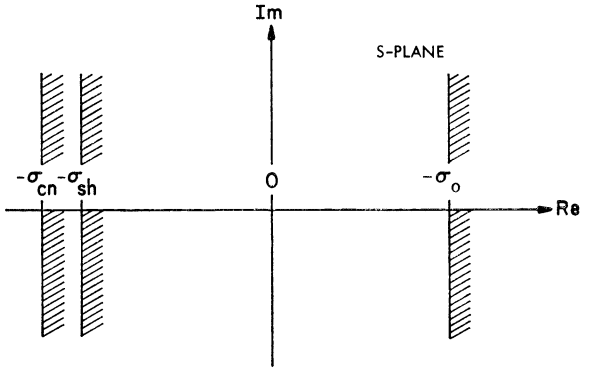


FIG. 2. *Regions in the complex plane*

The notion of stability adopted here is natural for integral equations. It implies asymptotic stability in the context of differential equations. With some additional minor assumptions, it also implies bounded-input bounded-output stability (see [2b]).

The main stability results will be stated in this section, but their proofs will be postponed (to §6) until after the supporting theory is established.

THEOREM 1. *Suppose that* (1) *and* (2), *and the related Assumptions* 1–4, *are valid. If*:

(i) *the Laplace transform* $H(s)$ *of the pair* $[h_0, h_1(t)]$ *is in* $Anl_{\sigma_{cn}}$ *for some constant* $\sigma_{cn} > 0$ (see Fig. 2);

(ii) *for some constant* σ_{sh}, $0 < \sigma_{sh} < \sigma_{cn}$, *the* σ_{sh}*-shifted Nyquist diagram of* $H(s)$ *does not intersect the negative real axis of the complex plane including the origin*;

(iii) *there are constants* $\underline{k} > 0$ *and* $\bar{k}$ *with the property that, for all* $t \geqq 0$, $\underline{k} \leqq k(t) \leqq \bar{k}$;

(iv) *there is a constant* $T > 0$ *for which*

$$\hat{K} = \sup_{t \geqq 0} \frac{1}{T} \int_t^{t+T} \left| \frac{d}{d\tau} \log k(\tau) \right| d\tau < 4\sigma_{sh}; \tag{4}$$

then $e(\cdot)$ *is in* $L_2[0, \infty)$, *and in fact* $\| e \| \leqq \text{const.} \| x \|$.

2.4. A stability theorem for gains in a prescribed interval.

DEFINITION 8. If $H(s)$ is the Laplace transform of $[h_0, h_1(t)]$ of (1) and (2), and if $1 + ah_0 \neq 0$, let $H^*(s)$ be the function with domain $\text{Re}\{s\} > -\sigma_0$ defined by the equation

$$H^*(s) = [1 + bH(s)][1 + aH(s)]^{-1}. \tag{5}$$

As the reciprocal of a nonzero analytic function in an open region is meromorphic, $H^*(s)$ is meromorphic for $\text{Re}\{s\} > -\sigma_0$. In fact, $H^*(s)$ is in Mer_{σ_0} by the following remark.

Remark 2. Let $H(s)$ and $H^*(s)$ be as in the preceding definition, with $1 + ah_0 \neq 0$, and suppose that $b \neq a$. Then $H(s)$ is in some $Mer_{\mu_{cn}}$ if and only if $H^*(s)$ is in $Mer_{\mu_{cn}}$ and, in fact, their continuations to $-\mu_{cn}$ satisfy the equation

$$H^*_{cn}(s) = [1 + bH_{cn}(s)][1 + aH_{cn}(s)]^{-1}. \tag{5a}$$

(For the proof of this remark, see the Appendix.)

THEOREM 2. *Suppose that* (1) *and* (2) *and the related Assumptions* 1–4 *are valid. If*

(i) *for all* $t \geqq 0$, $k(t)$ *lies in an interval* $[a + \epsilon, b - \epsilon]$, *where* $0 < \epsilon < (b - a)/2$;

(ii) *there are constants* σ_{cn} *and* σ_{sh} (see Fig. 2) *with* $0 < \sigma_{sh} < \sigma_{cn}$, *and the following properties hold*:

(a) $1 + ah_0 \neq 0$, *and the function*

$$H^*(s) \triangleq [1 + bH(s)][1 + aH(s)]^{-1}$$

with domain $\operatorname{Re}\{s\} > -\sigma_0$ *is in* $Anl_{\sigma_{cn}}$;

(b) *the* σ_{sh}*-shifted Nyquist diagram of* $H^*(s)$ *does not intersect the negative real axis of the complex plane including the origin*:

(iii) *for some constant* $T > 0$,

$$\hat{k} \triangleq \sup_{t \geqq 0} \frac{1}{T} \int_t^{t+T} \left| \frac{d}{d\tau} \log \frac{k(\tau) - a}{b - k(\tau)} \right| d\tau < 4\sigma_{sh}; \tag{6}$$

then $e(\cdot)$ *is in* $L_2[0, \infty)$ *and, in fact,* $\| e \| \leqq \text{const.} \| x \|$.

COROLLARY 1. *If, in addition to the hypotheses of Theorem* 2, $h_0 = 0$ *and* $h_1(t)$ *is in* $L_2[0, \infty)$, *then* $\lim_{t\to\infty} y(t) = 0$.

Remark 3. Let $\dot{k}$ denote the derivative of k. Then

$$\frac{1}{T} \int_t^{t+T} \left| \frac{d}{d\tau} \log \frac{k(\tau) - a}{b - k(\tau)} \right| d\tau = \frac{1}{T} \int_t^{t+T} \left| \frac{\dot{k}(b - a)}{(k - b)(k - a)} \right| d\tau. \tag{6a}$$

Remark 4. As $\hat{k}$ is defined by an average of $| d \log (\cdot)/dt |$, there is nothing to prevent $| d \log (\cdot)/dt |$ from being large over small subintervals of $[0, \infty)$.

2.5. Encirclement conditions. Instead of plotting the Nyquist diagram of $H^*(s) = [1 + bH(s)]\,[1 + aH(s)]^{-1}$, it is frequently more convenient to plot the diagram of $H(s)$. It will appear that it is always possible to replace conditions on $H^*(s)$ by conditions on $H(s)$.

DEFINITION 9. For any finite open interval (a, b) of the real numbers, the *critical region for* (a, b) consists of the set of real points x of the finite complex plane with the property that $-1/x$ is in the closed interval $[a, b]$.

For example, if $0 < a < b$, then the critical region is the interval $[-1/a, -1/b]$; if $a < 0 < b$, then the critical region is the union of intervals $(-\infty, -1/b] \cup [-1/a, \infty)$.

THEOREM 2′. *Theorem* 2 *remains valid if the assumptions* (a) *and* (b) *in* (ii) *are replaced by the following conditions*:

(ii)′. *There are constants* μ_{cn} *and* σ_{sh} *with* $0 < \sigma_{sh} < \mu_{cn}$ *and the following properties hold*:

(a) $H(s)$ *is in* $Mer_{\mu_{cn}}$ *(so that* $H(s)$ *has a finite number of poles*[4] *in* $\operatorname{Re}\{s\} \geqq -\sigma_{sh}$ *with no poles on the line* $\operatorname{Re}\{s\} = -\sigma_{sh}$);

(b) *the* σ_{sh}*-shifted Nyquist diagram of* $H(s)$ *does not intersect*[5] *the critical region for* (a, b), *but encircles it a number of times equal to the number of poles of* $H(s)$ *in* $\operatorname{Re}\{s\} > -\sigma_{sh}$.

Furthermore, for all σ_{sh} satisfying $0 < \sigma_{sh} < \sigma_{cn}$, except for at most a countable number of σ_{sh} for which there is a pole of $H(s)$ on the line $\operatorname{Re}\{s\} = -\sigma_{sh}$, conditions (ii) imply conditions (ii)′.

Remark 5. Under the assumptions of Theorem 2, $H(s)$ has a finite number of poles in the half-plane $\operatorname{Re}\{s\} \geqq -\sigma_{sh}$ whether or not $H(s)$ is rational, as seen from Theorem 2′. Furthermore, no generality is lost in considering the Nyquist diagram of $H(s)$ instead of that of $H^*(s)$.

2.6. Root locus conditions.

DEFINITION 10. $H(s)$ is *rational* if it can be expressed as a ratio of finite order polynomials. For a rational $H(s)$, the *root locus of* $H(s)$ consists of those points s for which there is a real constant c with $H(s) = -1/c$.

THEOREM 2″. *Let* $H(s)$ *be the Laplace transform of* $[h_0, h_1(t)]$ *in* (1) *and* (2), *and suppose that* $1 + ah_0 \neq 0$. *If* $H(s)$ *is rational, then the assumptions in* (ii) *of Theorem* 2 *are equivalent to the following condition*:

(ii)″ *The portion of the root locus for* c *in* $[a, b]$ *lies in the region* $\operatorname{Re}\{s\} < -\sigma_{sh}$.

Example. Suppose $H(s) = (s + p)^{-2}$, $p > 0$. For what gains does Theorem 2″ predict stability?

This example is easily evaluated in terms of the root locus. The root locus is confined to the two lines $\operatorname{Re}\{s\} = -p$ and $\operatorname{Im}\{s\} = 0$, and, in fact, lies in the region $\operatorname{Re}\{s\} = -p$ if $c \geqq 0$ and $\operatorname{Re}\{s\} \leqq -p + \sqrt{-c}$ if $c < 0$. Therefore, the system is stable for any constant gain in the interval $(-p^2, \infty)$. If (a, b) is a subinterval of $(-p^2, \infty)$, then the system is stable for any time-varying gain in a subinterval $[a + \epsilon, b - \epsilon]$ of (a, b), provided (i) $\hat{k} < 4p$ if $a \geqq 0$, or (ii) $\hat{k} < 4(p - \sqrt{-a})$ if $a < 0$. Thus for $a < 0$, a tradeoff is obtained between the interval (a, b) and the rate $\hat{k}$.

Remark 6. Suppose that the system of equations (1) and (2) is stable for every constant gain $k(t) = \text{const.}$ in an interval (a, b). The following question is raised: Does the system remain stable for every time-varying

[4] The poles of $H(s)$ are defined to be the poles of its continuation.

[5] This nonintersection condition implies that $1 + ah_0 \neq 0$.

gain $k(t)$ in (a, b), at least if the rate of variation is small enough? By Theorem 2′ the answer is affirmative, provided the Nyquist conditions are satisfied with a nonzero shift.

3. Basic method: A lemma on multipliers. In this section, a lemma will be stated which forms the skeleton of our method of providing stability. The lemma is based on the now well-known idea of splitting the open-loop system into two positive operators. Application of this idea will depend on finding suitable "multiplier" functions for $H(s)$ and $k(t)$.

DEFINITION 11. Let $\mathcal{L}_{\sigma_0}$ be the class of operators $\mathbf{H}: L_{2e} \to L_{2e}$ satisfying

$$(\mathbf{H}x)(t) = h_0 x(t) + \int_0^t x(\tau) h_1(t - \tau)\, d\tau \tag{7}$$

for all $x \in L_{2e}$ and for all $t \geqq 0$, where $[h_0, h_1(t)]$ is in LE_{σ_0}. (Note that $\mathbf{H}$ is nonanticipative, i.e., $(\mathbf{H}x)_T = (\mathbf{H}x_T)_T$ for all x in L_{2e} and all $T \geqq 0$.) The pair $[h_0, h_1(t)]$ will be referred to as the kernel of $\mathbf{H}$. The Laplace transform of $[h_0, h_1(t)]$ is defined as in (3). (Note that $H(s)$ is in Anl_{σ_0}.)

Let $\mathfrak{K}$ be the class of absolutely continuous real-valued functions $k(\cdot)$ on $[0, \infty)$, with each $k(\cdot)$ having constants $\underline{k} > 0$ and $\bar{k} \geqq \underline{k}$ for which $\underline{k} \leqq k(t) \leqq \bar{k}$ for all $t \geqq 0$.

LEMMA 1. *Let* (1) *and* (2) *and the related Assumptions* 1–4 *be valid. If there are constants* σ_{cn}, σ_{sh} *and* r *with* $0 < r < \sigma_{sh} < \sigma_{cn}$, $H(s)$ *is in* $Anl_{\sigma_{cn}}$, *and there is an operator* $\mathbf{M}$ *in* $\mathcal{L}_{\sigma_{sh}}$ *and a function* $f(\cdot)$ *in* $\mathfrak{K}$ *(the multipliers) with the following properties*:

(i)

$$\operatorname{Re}\{M(j\omega - \sigma_{sh})\} \geqq 0, \tag{8}$$

$$\operatorname{Re}\{M(j\omega - \sigma_{sh})H(j\omega - \sigma_{sh})\} \geqq \delta > 0 \tag{9}$$

for all $\omega \in (-\infty, \infty)$, *where* δ *is a constant*;

(ii) *the function* $f(\cdot)k(\cdot)$ *is in* $\mathfrak{K}$, *and* $f(t)\exp(-2rt)$ *and* $f(t)k(t)\cdot\exp(-2rt)$ *are monotonically nonincreasing functions of* t, *for* t *in* $[0, \infty)$, *then* e *is in* $L_2[0, \infty)$, *and* $\| e \| \leqq \text{const.} \| x \|$.

The proof appears in the Appendix.

4. Construction of $M(s)$: A lemma on multipliers with prescribed phase characteristics. It will be shown that an operator $\mathbf{M}$ in $\mathcal{L}_{\sigma_{sh}}$ can be found with *any* prescribed phase function $\arg\{M(j\omega - \sigma_{sh})\}$, provided the phase and its derivative are suitably restricted. This fact will make it possible to construct a multiplier $M(j\omega - \sigma_{sh})$ to rectify the phase of $H(j\omega - \sigma_{sh})$ so as to lie in the interval $(-90°, 90°)$. It will also be possible to show that a feedback system which is stable for all positive constant gains can be characterized by the fact that $\mathbf{H} + \epsilon\mathbf{I}$ can be factored into strongly positive oper-

ators, where H is the open-loop operator, $\mathbf{I}$ is the identity operator, and $\epsilon > 0$ is any constant.

LEMMA 2 (Multipliers with prescribed phase). *If*

(i) $\Phi_0(\omega)$ *is a real-valued, continuous, a.e. differentiable, odd function of* ω *for* ω *in* $(-\infty, \infty)$,

(ii) $\Phi_0(\omega)$ *and* $\Phi_0'(\omega)$ *are in* $L_2(-\infty, \infty)$,

then:

(a) *there is a* $\lambda(t)$ *in* $L_1(-\infty, \infty)$ *with* $\lambda(t) = 0$ *for* $t < 0$, *and with a Fourier transform* $\Lambda(\omega)$ *having the property that* $\operatorname{Im}\{\Lambda(\omega)\} = \Phi_0(\omega)$;

(b) *there is a* $z(t)$ *in* $L_1(-\infty, \infty)$ *with* $z(t) = 0$ *for* $t < 0$, *and with a Fourier transform* $Z(\omega)$ *having the property* $1 + Z(\omega) = \exp[\Lambda(\omega)]$;

(c) *if* $-\pi < \Phi_0(\omega) \leqq \pi$, *then there is a* $z(t)$ *in* $L_1(-\infty, \infty)$ *with* $z(t) = 0$ *for* $t < 0$, $1 + Z(\omega) \neq 0$, *and* $\arg\{1 + Z(\omega)\} = \Phi_0(\omega)$.

(*In fact, if* $Z(s)$ *is the Laplace transform of* $z(t)$, *then* $1 + Z(s)$ *is minimum phase, i.e., has no right half-plane zeros.*)

Proof of Lemma 2. (a) The proof is based on two main ideas: (i) if the inverse Fourier transform of a complex-valued function vanishes on $(-\infty, 0)$, then the real and imaginary parts of the function are in 1–1 correspondence; (ii) if a function and its derivative are in $L_2(-\infty, \infty)$, then the inverse Fourier transform is in $L_1(-\infty, \infty)$.

The *inverse limit-in-the-mean Fourier transform* of $j\Phi_0(\omega)$ is defined by the equation

$$\phi_0(t) = \frac{1}{2\pi} \lim_{W \to \infty} \int_{-W}^{W} j\Phi_0(\omega) \exp(j\omega t)\, d\omega.$$

Since $\Phi_0(\omega)$ is in $L_2(-\infty, \infty)$, the limit-in-the mean exists, and $\phi_0(t)$ is also in $L_2(-\infty, \infty)$. Since $j\Phi_0(\omega)$ is purely imaginary and odd, $\phi_0(t)$ is real and odd. Let $\phi_e(t)$ be the function on $(-\infty, \infty)$ defined by

$$\phi_e(t) = \begin{cases} \phi_0(t) & \text{for } t \geqq 0, \\ -\phi_0(t) & \text{for } t < 0. \end{cases}$$

Now it will be shown below that $\phi_0(t)$ is in $L_1(-\infty, \infty)$. Consequently, $\phi_e(t)$ is also in $L_1(-\infty, \infty)$ and has a Fourier transform $\Phi_e(\omega)$. Since $\phi_e(t)$ is real-valued and even, $\Phi_e(\omega)$ is purely real. Let $\lambda(t) = \phi_e(t) + \phi_0(t)$. Therefore $\lambda(t) = 0$ for $t < 0$, $\lambda(t)$ has a Fourier transform $\Lambda(\omega) = \Phi_e(\omega) + j\Phi_0(\omega)$, and $\Phi_0(\omega) = \operatorname{Im}\{\Lambda(\omega)\}$.

It remains to be shown that $\phi_0(t)$ is in $L_1(-\infty, \infty)$. Note first that since $\Phi_0(\omega)$ and its derivative $\Phi_0'(\omega)$ are in $L_2(-\infty, \infty)$, $t\phi_0(t)$ is in $L_2(-\infty, \infty)$ (see Titchmarsh [8a, Theorem 68, p. 92]). Consequently,

$$\int_{-\infty}^{\infty} |\phi_0(t)|\, dt \leqq \left[\int_{-\infty}^{\infty} \frac{1}{1+t^2}\, dt\right]^{1/2} \left[\int_{-\infty}^{\infty} (1+t^2)\,|\phi_0(t)|^2\, dt\right]^{1/2} < \infty$$

by the Schwarz inequality. Therefore, $\phi_0(t)$ is in $L_1(-\infty, \infty)$.

(b) If $x(t)$ and $y(t)$ are any two functions in $L_1(-\infty, \infty)$, let $(x * y)(t)$ denote their convolution $\int_{-\infty}^{\infty} x(t-\tau)y(\tau)\,d\tau$. Let $z_n(t)$ be the function defined on $[0, \infty)$ by the sum

$$z_n(t) = \lambda(t) + \frac{(\lambda * \lambda)(t)}{2!} + \cdots + \frac{\overbrace{(\lambda * \cdots * \lambda)(t)}^{n \text{ times}}}{n!}, \qquad n = 1, 2, \cdots. \tag{10}$$

Now if x and y have the property of belonging to $L_1(-\infty, \infty)$ and vanishing on $(-\infty, 0)$, then $x * y$ also has this property, and $\| x * y \|_1 \leqq \| x \|_1 \cdot \| y \|_1$. When this result is applied to (10) it can be concluded that the sequence $\{z_n(t)\}$ converges faster than the sequence of real numbers $\{s_n\}$ defined by

$$s_n \triangleq \| \lambda \|_1 + \frac{\| \lambda \|_1^2}{2!} + \cdots + \frac{\| \lambda \|_1^n}{n!} \tag{11}$$

in the sense that $\| z_n - z_m \|_1 \leqq | s_n - s_m |$ for any positive integers m and n. Now $\{s_n\}$ is a Cauchy sequence in the real numbers (whose limit is $\exp \{\| \lambda \|_1\} - 1$). It follows that $\{z_n\}$ is a Cauchy sequence in $L_1(-\infty, \infty)$ and, since $L_1(-\infty, \infty)$ is complete, there is a limit $z(t)$ in $L_1(-\infty, \infty)$, with $z(t) = 0$ for $t < 0$, to which $\{z_n\}$ converges. Furthermore, if the Fourier transforms of $z(t)$ and $\lambda(t)$ are denoted $Z(\omega)$ and $\Lambda(\omega)$, then $Z(\omega)$ is in $L_\infty^c(-\infty, \infty)$ (which is defined to be the space of complex-valued, essentially bounded functions on $(-\infty, \infty)$), and

$$\begin{aligned} Z(\omega) &= \Lambda(\omega) + \frac{\Lambda^2(\omega)}{2!} + \cdots \\ &= \exp [\Lambda(\omega)] - 1; \end{aligned} \tag{12}$$

part (b) of the lemma is proved. (Term-by-term transformation of the series for $z(t)$ is permitted, since the Fourier transformation is a continuous map of $L_1(-\infty, \infty)$ into $L_\infty^c(-\infty, \infty)$.)

(c) This follows immediately from (a) and (b).

LEMMA 3 (Construction of the multiplier). *If there are constants σ_{cn} and σ_{sh}, $0 < \sigma_{sh} < \sigma_{cn}$, for which $H(s)$ is in $Anl_{\sigma_{cn}}$, there are no poles of $H(s)$ on the line* $\operatorname{Re}\{s\} = -\sigma_{sh}$, *and the σ_{sh}-shifted Nyquist diagram of $H(s)$ does not intersect the negative real axis, including the origin, then there is an operator* $\mathbf{M}$ *in $\mathcal{L}_{\sigma_{sh}}$ with the properties that*

$$\operatorname{Re}\{M(j\omega - \sigma_{sh})\} \geqq \delta_1 > 0, \tag{13}$$

$$\operatorname{Re}\{M(j\omega - \sigma_{sh})H(j\omega - \sigma_{sh})\} \geqq \delta_2 > 0, \tag{14}$$

where δ_1 and δ_2 are constants, for all ω in $(-\infty, \infty)$.

The Nyquist diagram assumption implies that the argument of $H(j\omega - \sigma_{sh})$ lies in the interval $(-\pi, \pi)$. If the conclusions of the lemma are to be fulfilled, $M(j\omega - \sigma_{sh})$ and $M(j\omega - \sigma_{sh})H(j\omega - \sigma_{sh})$ must clearly have arguments in $(-\pi/2, \pi/2)$. Upon recalling that

$$\arg\{M(j\omega - \sigma_{sh})H(j\omega - \sigma_{sh})\} = \arg\{M(j\omega - \sigma_{sh})\} + \arg\{H(j\omega - \sigma_{sh})\},$$

our initial attempt at constructing $\mathbf{M}$ might involve choosing

$$\arg\{M(j\omega - \sigma_{sh})\} = -\tfrac{1}{2}\arg\{H(j\omega - \sigma_{sh})\}.$$

The construction which will actually be adopted here will depart from this choice for large $|\omega|$, in order to meet the added constraint that $\mathbf{M}$ must be in $\mathfrak{L}_{\sigma_{sh}}$. In fact, the construction which will be employed here will ensure that $\arg\{M(j\omega - \sigma_{sh})\}$ and its derivative will be in $L_2(-\infty, \infty)$, whereupon Lemma 2 will be used to show that $\mathbf{M}$ is in $\mathfrak{L}_{\sigma_{sh}}$.

The following remarks will be needed.

Remark 7. If $F(\omega)$ is a continuous complex-valued function on $(-\infty, \infty)$, $F(\omega) \neq 0$, $|\arg\{F(\omega)\}| < \pi/2$ and $\lim_{|\omega|\to\infty} F(\omega)$ exists and is a positive nonzero number, then there is a constant $\delta > 0$ with the property that $\operatorname{Re}\{F(\omega)\} \geqq \delta$.

Remark 8. If $H(s)$ is an analytic function of the complex variable s on $\operatorname{Re}\{s\} = -\sigma_{sh}$, where σ_{sh} is a constant, and if $H(j\omega - \sigma_{sh}) \neq 0$ for any ω, then $\arg\{H(j\omega - \sigma_{sh})\}$ is continuously differentiable with respect to ω.

Remark 8 follows immediately from the fact that

(i) $\arg\{H(s)\} = \operatorname{Im}\{\log H(s)\}$, so that

$$\frac{d}{d\omega}\operatorname{Arg}\{H(j\omega - \sigma_{sh})\} = \operatorname{Im}\frac{d\{H(j\omega - \sigma_{sh})\}/d\omega}{H(j\omega - \sigma_{sh})};$$

(ii) $H(j\omega - \sigma_{sh})$ is infinitely differentiable with respect to ω in a region of analyticity;

(iii) $H(j\omega - \sigma_{sh})$ has no zeros on $\operatorname{Re}\{s\} = -\sigma_{sh}$.

Proof of Lemma 3. The proof is in the three following steps. Step 1: Construction of a phase function $\Phi_0(\omega)$; Step 2: Determination of an operator $\mathbf{M}$ in $L_{\sigma_{sh}}$ with $\arg\{M(j\omega - \sigma_{sh})\} = \Phi_0(\omega)$; Step 3: Verification of inequalities (13) and (14).

Step 1: *Construction of* $\Phi_0(\omega)$. Under the assumptions on $H(s)$, $H(j\omega - \sigma_{sh})$ approaches a real constant h_0 as ω goes to ∞ (by the Riemann-Lebesgue theorem). The Nyquist condition ensures that $h_0 \neq 0$, and so $\arg\{H(j\omega - \sigma_{sh})\}$ approaches 0. Let $W > 0$ be a constant with the property that $|\arg\{H(j\omega - \sigma_{sh})\}| \leqq \pi/6$ for $\omega > W$. Let $\Phi_0(\omega)$ be a function on $(-\infty, \infty)$ defined by the equations

$$\Phi_0(\omega) = \begin{cases} -\tfrac{1}{2} \arg \{H(j\omega - \sigma_{sh})\} & \text{for } |\omega| \leqq W, \\ c\,\dfrac{W}{\omega} & \text{for } |\omega| > W, \end{cases} \tag{15}$$

where $c = -\frac{1}{2} \arg \{H(jW - \sigma_{sh})\}$. By this construction it is ensured that $|\Phi_0(\omega)| < \pi/2$ and $|\Phi_0(\omega) + \arg \{H(j\omega - \sigma_{sh})\}| < \pi/2$.

Step 2: *Determination of* **M** *in* $\mathcal{L}_{\sigma_{sh}}$. **M** will be determined using Lemma 2. From the definition of $\Phi_0(\omega)$ (see (15)) and Remark 8, it follows that $\Phi_0(\omega)$ is continuous for all ω, and continuously differentiable, except possibly at $\omega = W$. Further, $\Phi_0(\omega)$ and $\Phi_0'(\omega)$ asymptotically behave like $\omega^{-1} \cdot$const. and $\omega^{-2} \cdot$const., respectively. Therefore, it can be concluded that $\Phi_0(\omega)$ and $\Phi_0'(\omega)$ are in $L_2(-\infty, \infty)$. Thus $\Phi_0(\omega)$ satisfies the hypotheses of Lemma 2c, and there is a unique function $z_1(t)$ in $L_1[0, \infty)$ with Laplace transform $Z_1(s)$, $1 + Z_1(j\omega)\mathcal{L}_{\sigma_{sh}} \neq 0$ for any ω, and $\arg \{1 + Z_1(j\omega)\} = \Phi_0(\omega)$. Let **M** be the operator in $\mathcal{L}_{\sigma_{sh}}$ with kernel $[1, z(t) \exp(-\sigma_{sh}t)]$.

Step 3: *Verification of inequalities* (13) *and* (14). It can be checked that the functions $M(j\omega - \sigma_{sh})$ and $M(j\omega - \sigma_{sh})H(j\omega - \sigma_{sh})$ satisfy the hypotheses of Remark 7. Therefore, inequalities (13) and (14) are valid, and the lemma is proved.

4.1. A characterization of stable systems. An operator **H** in $\mathcal{L}_{\sigma_{cn}}$ is called *strongly positive* if there is a constant $\delta > 0$ for which $\mathrm{Re}\{H(j\omega)\} \geqq \delta$. The following useful characterization of a feedback system stable for positive gains is obtained from Lemma 3 and the Nyquist criterion[6]: *The feedback equations* (1) *and* (2) *are stable* (*in the sense of Theorem* 1), *for all positive constant gains* $k(t) = c \geqq 0$, *if and only if the operator* $\mathbf{H} + \epsilon\mathbf{I}$ *can be factored into two strongly positive operators for every* $\epsilon > 0$, *where* **I** *is the identity operator.*

5. Construction of $f(t)$: A factorization lemma for time-varying gains. The problem to be solved here is: Given $k(t)$ in $\mathcal{K}$ and a constant $r > 0$, is it possible to factor $k(t)$ into two parts $k(t) = k_+(t)k_-(t)$, so as to make $k_+(t) \exp(2rt)$ nondecreasing and $k_-(t) \exp(-2rt)$ nonincreasing? If such a factorization is possible, then a suitable multiplier function $f(t)$ for Lemma 1 can be obtained by setting $f(t) = 1/k_+(t)$.

LEMMA 4. *If* $k(t)$ *is in* $\mathcal{K}$, *and there are constants* $r > 0$ *and* $T > 0$ *for which the inequality*

$$\frac{1}{T}\int_t^{t+T} \left| \frac{d}{d\tau} \log k(\tau) \right| d\tau \leqq 4r \quad \textit{for all} \quad t \geqq 0 \tag{16}$$

[6] In the context of differential equations a similar characterization is considered in R. W. Brockett and J. W. Willems [11].

is valid, then there are functions $k_+(t)$ *and* $k_-(t)$ *in* $\mathfrak{K}$ *with the following properties*:

(i) $k(t) = k_+(t)k_-(t)$;

(ii) $k_+(t) \exp(2rt)$ *is nondecreasing*;

(iii) $k_-(t) \exp(-2rt)$ *is nonincreasing.*

Instead of Lemma 4, the following equivalent lemma will be proved.

LEMMA 4′. *Let* $\ell(t)$ *be a real-valued, absolutely continuous, bounded (i.e., from above and below) function of* t *for* t *in* $[0, \infty)$. *Suppose there are constants* $T > 0$ *and* $r > 0$ *for which the inequality*

$$\frac{1}{T}\int_t^{t+T} |\dot{\ell}(\tau)|\, d\tau \leqq 4\tau \quad \text{for all} \quad t \geqq 0 \tag{17}$$

is valid, where $\dot{\ell}(\tau)$ *denotes the derivative of* $\ell(\tau)$. *Then, there are two real-valued, absolutely continuous, bounded functions* $\ell_+(t)$ *and* $\ell_-(t)$ *on* $[0, \infty)$, *with the following properties for almost all* $t \geqq 0$:

$$\ell(t) = \ell_+(t) + \ell_-(t), \tag{18}$$

$$\frac{d\ell_+(t)}{dt} \geqq -2r, \tag{19}$$

$$\frac{d\ell_-(t)}{dt} \leqq 2r. \tag{20}$$

Clearly Lemma 4′ is equivalent to Lemma 4, as the assumptions on $\ell(t)$ are equivalent to those on $\log k(t)$, etc.

Proof of Lemma 4′. The functions $\ell_+(t)$ and $\ell_-(t)$ will be defined by the equations

$$\ell_+(t) = \frac{\ell(t)}{2} + \tilde{\ell}(t), \tag{21}$$

$$\ell_-(t) = \frac{\ell(t)}{2} - \tilde{\ell}(t), \tag{22}$$

where the function $\tilde{\ell}(t)$ will be constructed presently, with suitable restrictions on $|\tilde{\ell}(t)|$ and $d\tilde{\ell}(t)/dt$. It is clear that (18) will be satisfied.

Construction of $\tilde{\ell}(t)$. Condition (17) is equivalent to the following inequality, in terms of a convolution:

$$\int_0^t |\dot{\ell}(\tau)|\, |w(t-\tau)|\, d\tau \leqq 4r \quad \text{for all} \quad t \geqq 0, \tag{23}$$

where $w(t)$ is a "smoothing kernel," defined on $[0, \infty)$ by the equation

$$w(t) = \begin{cases} 1/T & \text{on } [0, T], \\ 0 & \text{on } (T, \infty). \end{cases} \tag{24}$$

Let

$$\tilde{\ell}(t) \triangleq \frac{1}{2}\int_0^t |\dot{\ell}(\tau)| \left\{1 - \int_0^{t-\tau} w(\sigma)\, d\sigma\right\} d\tau. \tag{25}$$

With the definition of $\ell_+(t)$ and $\ell_-(t)$ by (21) and (22) now complete, it remains to be shown that these functions and their derivatives are suitably bounded.

Boundedness of $\ell_+(t)$ and $\ell_-(t)$. The smoothing kernel has the property that

$$0 \leqq 1 - \int_0^{t-\tau} w(\sigma_1)\, d\sigma_1 \leqq Tw(t-\tau) \quad \text{for all} \quad t \geqq \tau. \tag{26}$$

From (26), (25) and (23), it follows that $0 \leqq \tilde{\ell}(t) \leqq 2Tr$, and so $\tilde{\ell}(t)$ is bounded. Since $\ell(t)$ is bounded by hypothesis, it follows from (21) and (22) that $\ell_+(t)$ and $\ell_-(t)$ are bounded.

Bounds on $d\ell_+/dt$ and $d\ell_-/dt$. From (25),

$$\frac{d\tilde{\ell}(t)}{dt} = \frac{1}{2}|\dot{\ell}(t)| - \frac{1}{2}\int_0^t |\dot{\ell}(\tau)|\, w(t-\tau)\, d\tau. \tag{27}$$

Therefore,

$$\frac{d\ell_+(t)}{dt} = \frac{1}{2}\{\dot{\ell}(t) + |\dot{\ell}(t)|\} - \frac{1}{2}\int_0^t |\dot{\ell}(\tau)|\, w(t-\tau)\, d\tau, \tag{28}$$

$$\frac{d\ell_-(t)}{dt} = \frac{1}{2}\{\dot{\ell}(t) - |\dot{\ell}(t)|\} + \frac{1}{2}\int_0^t |\dot{\ell}(\tau)|\, w(t-\tau)\, d\tau. \tag{29}$$

Now (19) of the lemma follows from (28), (23), and the inequality $\dot{\ell}(t) + |\dot{\ell}(t)| \geqq 0$. Similarly, (20) follows from (29), (23), and the inequality $\dot{\ell}(t) - |\dot{\ell}(t)| \leqq 0$.

6. Proofs of the main results.

Proof of Theorem 1. Theorem 1 will follow from Lemmas 1, 3 and 4.

Under the hypotheses of Theorem 1, the kernel $[h_0, h_1(t)]$ has a Laplace transform which satisfies the hypotheses of Lemma 3. Therefore, there is an operator $\mathbf{M}$ in $\mathfrak{L}_{\sigma_{sh}}$ and a constant $\delta > 0$ satisfying inequalities (8) and (9) of Lemma 1.

Furthermore, under the hypotheses of Theorem 1, $k(t)$ is in $\mathfrak{K}$ and the hypotheses of Lemma 4 are satisfied. Therefore, there are functions $k_+(t)$ and $k_-(t)$ in $\mathfrak{K}$ with the properties that: (i) $k(t) = k_+(t)k_-(t)$; (ii) $k_+(t)$ exp $(2rt)$ is nondecreasing; and (iii) $k_-(t)$ exp $(-2rt)$ is nonincreasing. Letting $f(t) = 1/k_+(t)$, we observe that $f(t)$ satisfies the hypotheses of Lemma 1. Thus, all the hypotheses of Lemma 1 are satisfied, and Theorem 1 follows.

Proof of Theorem 2. Theorem 2 is established by transforming the feedback equations into a form in which Theorem 1 applies. The following new terms are defined:

$$k^* = (k - a)(b - k)^{-1}, \tag{30}$$

$$H^*(s) = [1 + bH(s)][1 + aH(s)]^{-1}, \tag{31}$$

$$x^* = (1 + k^*)x, \tag{32}$$

$$e^* = e + ay, \tag{33}$$

$$y^* = e + by. \tag{34}$$

By Remark 2, $H^*(s)$ is the Laplace transform of a pair $[h_0^*, h_1^*(t)]$. It can be verified that

$$\begin{aligned} e^*(t) &= x^*(t) - k^*(t)y^*(t), \\ y^*(t) &= h_0^* e^*(t) - \int_0^t e^*(\tau)h_1^*(t - \tau)\, d\tau, \end{aligned} \tag{35}$$

where the equations (35) represent a system satisfying the hypotheses of Theorem 1.

Consequently, e^* is in $L_2[0, \infty)$ and $\| e^* \| \leqq$ const. $\| x^* \|$. Since $e_T = (1 + aH)^{-1} e_T^*$ and $(1 + aH)^{-1}$ is bounded from the assumptions on H^*, it follows that $\| e_T \| \leqq$ const. $\| e_T^* \|$. Clearly also $\| x_T^* \| \leqq [1 + \bar{k}^*] \cdot \| x_T \|$, and so it may be concluded that e is in $L_2[0, \infty)$, and, in fact, $\| e \| \leqq$ const. $\| x \|$.

Proof of Corollary 1. Corollary 1 follows from the fact that the convolution of two L_2 functions is the Fourier transform of an L_1 function, which must approach zero at infinity by the Riemann-Lebesgue theorem.

Proof of Theorem 2′. Only the nontrivial case $a \neq 0$ will be considered. To begin with, it can be checked that $1 + ah_0 \neq 0$ under either assumption (ii) or (ii)′, and that the nonintersection conditions are equivalent. Next, it will be recalled that, for any constant σ_1, $H(s)$ is in Mer_{σ_1} if and only if $H^*(s)$ is in Mer_{σ_1}, by Remark 2. Theorem 2′ can now be deduced from the following.

Remark 9. If $H(s)$ and $H^*(s)$ are in Mer_{σ_1}, then, for all $\sigma_{sh} < \sigma_1$ for which there are no poles of $H(s)$ on the line Re $\{s\} = -\sigma_{sh}$, the following statements are equivalent:

(i) The σ_{sh}-shifted Nyquist diagram of $H(s)$ does not cut the point $-1/a$, but encircles it a number of times equal to the number of poles of $H(s)$ in Re $\{s\} \geqq -\sigma_{sh}$.

(ii) $H^*(s)$ is in $Anl_{[\sigma_{sh}+\epsilon]}$ for some $\epsilon > 0$.

Proof of Remark 9. By the principle of the argument, (i) is equivalent to the condition that $1 + aH(s)$ has no zeros in Re $\{s\} \geqq -\sigma_{sh}$. As $aH^* = b$

$+ (b - a)[1 + aH(s)]^{-1}$ for $\text{Re}\{s\} \geqq -\sigma_1$, (i) is equivalent to the condition that $H^*(s)$ is analytic for $\text{Re}\{s\} \geqq -\sigma_{sh}$. Finally, for any function $H^*(s)$ in Mer_{σ_1}, $H^*(s)$ is analytic for $\text{Re}\{s\} \geqq -\sigma_{sh}$ if and only if $H^*(s)$ is in $Anl_{[\sigma_{sh}+\epsilon]}$ for some $\epsilon > 0$, by a simple argument involving the distribution of the zeros of a function in Mer_{σ_1}.

7. Concluding remarks. A stability condition has been derived in terms of the geometry of the shifted Nyquist diagram (Theorems 2 and 2′) or of the root locus (Theorem 2″). For slowly time-varying systems the condition approaches a "frozen-time" Nyquist condition (Remark 6). The proof involves the construction of a multiplier having a prescribed argument (Lemmas 2 and 3) and factorization of the gain $k(t)$. It has been shown that the fact that the system of Fig. 1 is stable for all positive constant gains can be characterized by the property that $\mathbf{H} + \epsilon\mathbf{I}$ can be factored into two strongly positive operators (§4.1).

Although it has been assumed for simplicity that the gain $k(t)$ is differentiable, the results of this paper can be extended to a gain having jump discontinuities provided its logarithm has a finite variation var $[\log k(\cdot)]$ on every finite interval $[t, t + T]$. The variation condition on $k(t)$ then becomes

$$\sup_{t \geqq 0} \frac{1}{T} \text{var } [\log k(\cdot)]_t^{t+T} < 4\sigma_{sh}.$$

It is emphasized that although the conditions derived here restrict the transform $H(s)$ to be meromorphic in a half-plane, there is nothing to prevent $H(s)$ from having essential singularities in the other half-plane, and so the results are applicable to distributed systems.

Appendix. Proof of Remark 2. Only the nontrivial case $a \neq 0$ will be considered. The proof will be developed in a series of propositions. Suppose first that $H(s)$ is in $Mer_{\mu_{cn}}$.

P1. $H(s)$ is the transform of a pair $[h_0, h_1(t)]$ in some LE_{μ_0}, by definition of $Mer_{\mu_{cn}}$.

P2. $\lim_{|\omega| \to \infty} H(j\omega + \mu_0) = h_0$, by P1 and the Riemann-Lebesgue theorem. Also, $H(s)$ is bounded in $\text{Re}\{s\} \geqq -\mu_0$.

P3. $\lim_{|s| \to \infty} H(s) = h_0$ in $\text{Re}\{s\} \geqq -\mu_0$, by P2 and a theorem of Phragmèn-Lindelöf (Titchmarsh [8b, §5.63]).

P4. Let μ_{sh} be any constant satisfying $-\mu_{sh} > -\mu_{cn}$. Then $H(\sigma + j\omega)$ approaches h_0 as $|\omega| \to \infty$ uniformly with σ in $\text{Re}\{s\} \geqq -\mu_{sh}$, by definition of $Mer_{\mu_{cn}}$.

P5. $\lim_{|s| \to \infty} H(s) = h_0$ in $\text{Re}\{s\} \geqq -\mu_{sh}$, by P3 and P4.

P6. The zeros of $1 + aH(s)$ in $\text{Re}\{s\} \geqq \mu_{sh}$ lie in a closed bounded subregion of $\text{Re}\{s\} \geqq -\mu_{sh}$, by P5 and the assumption that $1 + ah_0 \neq 0$.

P7. $1 + aH(s)$ has at most a finite number of zeros in $\text{Re}\{s\} \geqq -\mu_{sh}$,

since a function meromorphic in a closed bounded region can have at most a finite number of zeros in the region.

P8. By P7, $H^*(s)$ is analytic in some $\text{Re}\{s\} \geqq -\mu_0^*$, where $-\mu_0^*$ is a large enough constant.

P9. $H^*(s)$ is the transform of a pair in LE_{μ_1}, where $-\mu_1 = \max(-\mu_0, -\mu_0^*)$ by an argument based on a theorem of Paley and Wiener [9, Theorem XVIII, p. 60 (with a shift $-\mu_1$)].

P10. Clearly $H^*_{cn}(s)$ as defined by (5a) is the meromorphic continuation of $H^*(s)$ to $-\mu_{cn}$.

P11. In $\text{Re}\{s\} \geqq -\mu_{sh}$, $H^*_{cn}(s)$ has a finite number of poles by P7, and $H^*_{cn}(\sigma + j\omega)$ approaches h_0 uniformly with σ as $|\omega| \to \infty$, by P4 and the assumption that $1 + ah_0 \neq 0$.

From P9, P10 and P11, it can be concluded that $H^*(s)$ is in $Mer_{\mu_{cn}}$. The converse part of Remark 2 has a similar proof, which will be omitted.

Proof of Lemma 1. The following proof falls into the category of "energy balance" arguments, having as its starting point an equation which can be interpreted as a statement of conservation of energy of a network (for a discussion, see [2c, Part I]).

Since $f(\cdot)$ is in $\mathcal{K}$, there are constants $\underline{f} > 0$ and $\bar{f} > \underline{f}$ for which $\underline{f} \leqq f(t) \leqq \bar{f}$ for all $t \geqq 0$. Let $\rho \triangleq [1 - r/\sigma_{sh}]\underline{f}$ and observe that this choice of ρ is small enough to ensure that $[f(t) - \rho] \exp(-2\sigma_{sh}t)$ is a nonincreasing function of t and that $f(t) - \rho \geqq 0$.

The hypotheses of this lemma assure that $M(s)H(s)$ is an $Anl_{\sigma_{sh}}$ for $\sigma_{sh} > 0$. This condition is sufficient to imply that **MH** is a bounded map of $L_2[0, \infty)$ to $L_2[0, \infty)$; in fact, the following is true, although the proof will be omitted.

Remark A1. Let **Z** be in $\mathfrak{L}_{\sigma_0}$ for some σ_0, and further assume that $Z(s)$ is in $Anl_{\sigma_{sh}}$ for $\sigma_{sh} > 0$. Then, for any $\sigma_1 < \sigma_{sh}$ and $x(t)$ with $x(t) \exp(\sigma_1 t) \in L_2[0, \infty)$, $(\mathbf{Z}x)(t) \exp(\sigma_1 t)$ is in $L_2[0, \infty)$ and its Fourier transform is $Z(j\omega - \sigma_1)X(j\omega - \sigma_1)$, where $X(s)$ denotes the Laplace transform of $x(t)$.

In particular, if $x(t)$ is in $L_2[0, \infty)$ and $\sigma_{sh} > 0$, then $\mathbf{Z}x(t)$ is in $L_2[0, \infty)$ with Fourier transform $Z(j\omega)X(j\omega)$. Furthermore, by the Parseval theorem, for any x in $L_2[0, \infty)$ the inequality $\| \mathbf{Z}x \| \leqq \gamma(\mathbf{Z}) \cdot \| x \|$ is obtained, where $\gamma(\mathbf{Z}) \triangleq \sup_{-\infty<\omega<\infty} |Z(j\omega)|$.

The following "conservation of energy" equation is obtained from (1) and (2), for any $T \geqq 0$:

$$\int_0^T x(t) \cdot f(t) \cdot [\mathbf{MH}e](t) \cdot dt$$

$$= \rho \int_0^T e(t) \cdot [\mathbf{MH}e](t) \cdot dt + \int_0^T e(t) \cdot (f(t) - \rho) \cdot [\mathbf{MH}e](t)\, dt \tag{A1}$$

$$+ \int_0^T k(t) \cdot y(t) \cdot f(t) \cdot [\mathbf{My}](t)\, dt.$$

The first term in (A1) is bounded using the Schwarz inequality and the fact that **M** and **H** are nonanticipative:

$$\text{(A2)} \qquad \int_0^T x(t)\cdot f(t)\cdot[\mathbf{MH}e](t)\,dt \leqq \|x_T\|\cdot\|(\mathbf{MH}e)_T f\| \leqq \bar{f}\cdot\gamma(\mathbf{MH})\cdot\|x_T\|\cdot\|e_T\|,$$

where a dot indicates either the product of two real numbers or of two real-valued functions. Here $\gamma(\mathbf{MH}) = \sup_{-\infty<\omega<\infty} |M(j\omega)H(j\omega)|$ as in Remark A1. It will be shown below that the remaining terms in (A1) satisfy the following inequalities:

$$\text{(A3)} \qquad \int_0^T e(t)\cdot[\mathbf{MH}e](t)\,dt \geqq \delta\|e_T\|^2 \quad \text{for some} \quad \delta > 0,$$

$$\text{(A4)} \qquad \int_0^T e(t)\cdot(f(t)-\rho)\cdot[\mathbf{MH}e](t)\,dt \geqq 0,$$

$$\text{(A5)} \qquad \int_0^T y(t)\cdot k(t)\cdot f(t)\cdot[\mathbf{M}y](t)\,dt \geqq 0.$$

From (A1) through (A5) it can be concluded that

$$\|e_T\| \leqq A\|x_T\|,$$

where

$$\text{(A6)} \qquad A = \frac{\bar{f}\cdot\gamma(\mathbf{MH})}{f\cdot\delta\cdot[1 - r/\sigma_{sh}]},$$

and since (A6) is valid for all $T \geqq 0$, the lemma is established.

Inequalities in (A3) through (A5) will be proved using Parseval's theorem. In (A4) and (A5) the exponentially weighted form of Parseval's theorem will be needed, and the second mean value theorem (Hobson [7, Chap. VII, §422] will be used.

Inequality (A3). Let $E_T(s)$ be the Laplace transform of $e_T(t)$. Recalling that **MH** is nonanticipative and in $Anl_{\sigma_{sh}}$, one obtains

$$\text{(A7)} \qquad \int_0^T e(t)\cdot|\mathbf{MH}e|(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} M(j\omega)\cdot H(j\omega)\cdot|E_T(j\omega)|^2\,d\omega$$

(Parseval's theorem and Remark A1)

$$\text{(A8)} \qquad = \frac{1}{\pi}\int_0^{\infty} \operatorname{Re}\{M(j\omega)\cdot H(j\omega)\}\,|E_T(j\omega)|^2\,d\omega.$$

Now $M(s)H(s)$ is in $Anl_{\sigma_{sh}}$. Hence, $M(s)H(s)$ is analytic in $\operatorname{Re}\{s\} > -\sigma_{sh}$ and approaches a limit as $|s| \to \infty$. Consequently, the real part $\operatorname{Re}\{M(s)H(s)\}$ cannot take its minimum inside the region. Since $M(s)H(s)$

is continuous on the boundary, i.e., on $\mathrm{Re}\{s\} = -\sigma_{sh}$, it follows from (9) in Lemma 1(i) that $\mathrm{Re}\{M(j\omega)H(j\omega)\} \geqq \delta$. From (A8) therefore,

$$\text{(A9)} \qquad \int_0^T e(t)\cdot[\mathbf{MH}e](t)\,dt \geqq \frac{\delta}{\pi}\int_0^\infty |E_T(j\omega)|^2\,d\omega = \delta\,\|e_T\|^2.$$

Inequality (A4). By construction, $(f(t) - \rho)\exp(-2\sigma_{sh}t)$ is monotone nonincreasing. By the second mean value theorem, there is a point T' in $[0, T]$ for which

$$\begin{aligned}&\int_0^T e(t)\cdot(f(t) - \rho)\cdot[\mathbf{MH}e](t)\,dt\\ \text{(A10)}\qquad &= \int_0^T \{(f(t) - \rho)\cdot\exp(-2\sigma_{sh}t)\}\cdot\{e(t)\cdot[\mathbf{MH}e](t)\cdot\exp(2\sigma_{sh}t)\}\,dt\\ &= \{f(0) - \rho\}\int_0^{T'} e(t)\cdot[\mathbf{MH}e](t)\cdot\exp(2\sigma_{sh}t)\cdot dt.\end{aligned}$$

Now $f(0) - \rho \geqq 0$ by construction. Also, the integral in (A10) is nonnegative by (9), the argument here being similar to the one used in proving (A3), but with Parseval's theorem in its exponential version. Therefore (A10) is nonnegative.

Inequality (A5). The proof here is similar to the proof of (A4), drawing on (8) and assumption (ii) of Lemma 1, (i.e., on the fact that $f(t)k(t)\exp(-2rt)$ is nonincreasing).

REFERENCES

[1] J. J. Bongiorno, Jr., *An extension of the Nyquist-Barkhausen stability criterion to linear lumped-parameter systems with time-varying elements*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 166–170.

[2a] G. Zames, *On the stability of nonlinear, time-varying feedback systems*, Proc. NEC, 20 (1964), pp. 725–730.

[2b] ———, *Nonlinear time-varying feedback systems: Conditions for L_∞-boundedness derived using conic operators on exponentially weighted spaces*, Proc. 3rd Allerton Conference on Circuit and System Theory, University of Illinois, Urbana (1965), pp. 460–471.

[2c] ———, *On the input-output stability of time-varying nonlinear feedback systems, Parts I and II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, 465–476.

[2d] G. Zames and P. L. Falb, *Stability conditions for systems with monotone and odd monotone nonlinearities*, Ibid., AC-12 (1967), pp. 221–223.

[3] R. W. Brockett and L. J. Forys, *On the stability of systems containing a time-varying gain*, Proc. 2nd Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1964, pp. 413–430.

[4] M. Gruber and J. L. Willems, *On a generalization of the circle criterion*, Proc. 4th Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1966, pp. 827–848.

[5] C. A. DESOER, *A general formulation of the Nyquist criterion*, IEEE Trans. Circuit Theory, CT-12 (1965), pp. 230–234.

[6] I. W. SANDBERG, *On the stability of linear systems containing a time-varying element with restricted rate of time variation*, IEEE International Convention Record, 14 (1966), Part 7, pp. 173–182.

[7] E. W. HOBSON, *The Theory of Functions of a Real Variable*, vol. 1, Dover, New York, 1957.

[8a] E. C. TITCHMARSH, *Theory of Fourier Integrals*, 2nd ed., Clarendon Press, Oxford, 1962.

[8b] ———, *The Theory of Functions*, 2nd ed., Oxford University Press, Oxford, 1964.

[9] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Providence, 1934.

[10] K. S. NARENDRA AND R. M. GOLDWYN, *A geometrical criterion for the stability of certain nonlinear nonautonomous systems*, IEEE Trans. Circuit Theory, CT-11 (1964), pp. 406–408.

[11]. R. W. BROCKETT AND J. W. WILLEMS, *Frequency domain stability criteria, I and II*, IEEE Proc. Joint Automatic Control Conference, 2(1965), pp. 735–747.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

A TIGHT UPPER BOUND ON THE RATE OF CONVERGENCE OF THE FRANK-WOLFE ALGORITHM*

M. D. CANON† AND C. D. CULLUM‡

Introduction. The Frank-Wolfe algorithm [1] is a well-known iterative procedure for computing the minimum of a convex function $f(x)$ over a convex polyhedron. It was shown by Frank and Wolfe that if $\hat{x}$ is a solution to the problem, and x_k, $k = 1, 2, \cdots$, are the iterates obtained via the algorithm, then $f(x_k) - f(\hat{x}) \leqq \alpha/k$ for some positive constant α, and for all k sufficiently large. Empirical data, moreover, indicate that the rate of convergence of $f(x_k) - f(\hat{x})$ to zero is not significantly better than $1/k$. The purpose of this paper is to make precise the phrase "not significantly better than $1/k$". In particular, we shall prove under usually satisfied assumptions that no matter what $\epsilon > 0$ and $\alpha > 0$ are, $f(x_k) - f(\hat{x}) \geqq \alpha/k^{1+\epsilon}$ for an infinite number of k.

For simplicity in presentation, we have taken f to be a positive definite quadratic form. To the specialist, however, it will be clear that the same result can be obtained as long as f is convex, twice continuously differentiable, and the Hessian matrix, $\partial^2 f(\hat{x})/\partial x^2$, evaluated at the optimal solution $\hat{x}$, is positive definite.

Finally, to avoid possible confusion, we point out that Frank and Wolfe also gave in [1] a separate algorithm for solving quadratic programming problems which converges in a finite number of steps. The algorithm referred to in this paper is the Frank-Wolfe convex programming algorithm.

1. Problem statement. We shall consider the following quadratic programming problem P:

Minimize

$$f(x) = \tfrac{1}{2}\langle x, Qx\rangle + \langle x, d\rangle$$

subject to the constraints

$$Bx \leqq c.$$

Here Q is a *positive definite* symmetric $n \times n$ matrix, B, d and c are $s \times n$, $n \times 1$, and $s \times 1$ matrices, respectively. We shall assume that problem P has a solution, and so without loss in generality we shall assume that the set of feasible solutions is *bounded.*

* Received by the editors October 16, 1967, and in revised form May 9, 1968.

† International Business Machines Corporation, San Jose Research Laboratory, San Jose, California 95114.

‡ International Business Machines Corporation, Thomas J. Watson Research Laboratory, Yorktown Heights, New York 10598.

For the purpose of analysis, it is convenient to restate problem P in the following equivalent form. The set $\Omega = \{x \mid Bx \leqq c\}$ is a bounded convex polyhedron. Let a^i, $i = 1, 2, \cdots, l$, be the extreme points of this set; Ω is the convex hull of its extreme points, i.e.,

$$\Omega = \{x \mid x = A\xi, \sum \xi^i = 1, \xi \geqq 0\},$$

where A is an $n \times l$ matrix whose columns are the a^i, $i = 1, 2, \cdots, l$. Problem P can be stated in the following form:

Minimize

$$f(x) = \tfrac{1}{2}\langle x, Qx\rangle + \langle x, d\rangle$$

subject to the constraints:

$$\begin{aligned} x &= A\xi, \\ \sum \xi^i &= 1, \\ \xi &\geqq 0. \end{aligned}$$

Since $\Omega \neq \varnothing$[1] is compact and convex, and $f(x)$ is continuous and strictly convex, problem P has a unique solution in x, which henceforth will be denoted by $\hat{x}$.

Let $\nabla f(x) = Qx + d$. When specialized to problem P, the Frank-Wolfe algorithm reads as follows:

FW-ALGORITHM: Given $x_k \in \Omega$, select a^{i_k} such that

$$\langle a^{i_k}, \nabla f(x_k)\rangle = \min_{i=1,\cdots,l} \langle a^i, \nabla f(x_k)\rangle,$$

and let $s_k = (a^{i_k} - x_k)$. If $\langle s_k, \nabla f(x_k)\rangle \geqq 0$, then x_k solves problem P. Otherwise, let λ_k be such that

$$\min_{\lambda>0} f(x_k + \lambda s_k) = f(x_k + \lambda_k s_k),$$

i.e.,

$$\lambda_k = -\frac{\langle s_k, \nabla f(x_k)\rangle}{\langle s_k, Qs_k\rangle}, \tag{1.1}$$

and set

$$\begin{aligned} \mu_k &= \min\{\lambda_k, 1\}, \\ x_{k+1} &= x_k + \mu_k s_k. \end{aligned} \tag{1.2}$$

Repeat the above procedure with x_{k+1} replacing x_k.

For later reference, let us note that if $x_k = A\xi_k$, with $\sum \xi_k{}^i = 1$ and

[1] In this paper $\varnothing$ is the empty set, $\partial\Omega$ is the relative boundary of Ω, and Ω its relative interior.

$\xi_k \geqq 0$, and l_{i_k} is a vector whose i_kth component is one and whose remaining components are zero, then it follows from (1.2) that $x_{k+1} = A[(1 - \mu_k)\xi_k + \mu_k l_{i_k}] \triangleq A\xi_{k+1}$, with $\sum \xi^i_{k+1} = 1$ and $\xi_{k+1} \geqq 0$. We may then write:

$$x_{k+1} = A\xi_{k+1}, \quad \sum \xi^i_{k+1} = 1, \quad \xi_{k+1} \geqq 0, \qquad k = 1, 2, \cdots,$$

where

$$\xi_{k+1} = (1 - \mu_k)\xi_k + \mu_k l_{i_k}, \qquad k = 1, 2, \cdots. \tag{1.3}$$

Concerning the above algorithm, Frank and Wolfe established the following result.

THEOREM 1. *Let $\hat{x}$ be the solution to problem* P. *There is an integer K and a constant $\alpha > 0$ such that for all $k \geqq K$, $f(x_k) - f(\hat{x}) \leqq \alpha/k$.*

Remark. In particular, it follows that $f(x_k) \to f(\hat{x})$, so that $x_k \to \hat{x}$, since $\hat{x}$ is the unique solution to P.

The principal result of this paper is contained in the next theorem.

THEOREM 2. *Let $\hat{x}$ be the solution to problem* P, *and suppose that*

(i) $\hat{x} \in \partial\Omega$,

(ii) $x_k \in \dot{\Omega}$ *for infinitely many* $k \in \{1, 2, \cdots\}$.

Then, for every constant $\alpha > 0$ and for every $\epsilon > 0$, $f(x_k) - f(\hat{x}) \geqq \alpha/k^{1+\epsilon}$ for infinitely many $k \in \{1, 2, \cdots\}$.

Some remarks are in order concerning the hypotheses made in Theorem 2. Properties (i) and (ii) guarantee that the algorithm does not terminate in a finite number of steps. The real reason for imposing them, however, is more subtle, and will be discussed in detail in §3. We point out that in the majority of practical problems property (i) holds and, although it may not appear obvious, the same remark also applies to (ii). In any case, if property (i) holds, and $x_1 \in \dot{\Omega}$ is chosen so that

$$f(x_1) \leqq f(a^i) \quad \text{for} \quad i = 1, 2, \cdots, l,$$

then it is easily shown that property (ii) holds.

2. Proof of Theorem 2. We shall assume throughout this section that the hypotheses of Theorem 2 are satisfied. We begin by proving two lemmas which will be useful.

LEMMA 1. *Let b_n, $n = 1, 2, \cdots$, be a sequence of real numbers such that the series $\sum |b_n|$ diverges. Then, for every $\epsilon > 0$, $\sum_{n=k}^{\infty} b_n^2 \geqq 1/k^{1+\epsilon}$ for infinitely many $k \in \{1, 2, \cdots\}$.*

Proof. The proof is by contraposition. Suppose there is an $\epsilon > 0$ and an integer K such that for $k \geqq K$,

$$\sum_{n=k}^{\infty} b_n^2 \leqq 1/k^{1+2\epsilon}. \tag{2.1}$$

We shall prove that the series $\sum |b_n|$ converges. Equivalently, since

$$\sum |b_n| = \sum |b_n| n^{(1+\epsilon)/2} n^{-(1+\epsilon)/2} \leqq (\sum b_n{}^2 n^{1+\epsilon})^{1/2} (\sum 1/n^{1+\epsilon})^{1/2}$$

by the Schwarz inequality, and the series $\sum 1/n^{1+\epsilon}$ converges for $\epsilon > 0$, we shall prove that the series $\sum b_n{}^2 n^{1+\epsilon}$ converges. Using (2.1) we have

$$\sum_{k=K}^{m} k^{\epsilon} \sum_{n=k}^{m} b_n{}^2 \leqq \sum_{k=K}^{m} k^{\epsilon} \sum_{n=k}^{\infty} b_n{}^2 \leqq \sum_{k=K}^{m} 1/k^{1+\epsilon}.$$

Now

$$\sum_{k=K}^{m} k^{\epsilon} \sum_{n=k}^{m} b_n{}^2 = \sum_{n=K}^{m} b_n{}^2 \sum_{k=K}^{n} k^{\epsilon},$$

and since

$$\sum_{k=K}^{n} k^{\epsilon} \geqq \int_{K}^{n} x^{\epsilon}\, dx = \frac{1}{1+\epsilon} (n^{1+\epsilon} - K^{1+\epsilon}),$$

it follows that

$$\frac{1}{1+\epsilon} \sum_{n=K}^{\infty} b_n{}^2 (n^{1+\epsilon} - K^{1+\epsilon}) \leqq \sum_{k=K}^{\infty} 1/k^{1+\epsilon}.$$

Since $\epsilon > 0$, the series on the right converges, and so the series on the left converges. By (2.1), the series $\sum b_n{}^2$ converges, and, therefore, the series $\sum b_n{}^2 n^{1+\epsilon}$ converges.

LEMMA 2. *Let* $0 \leqq b_n < 1$ *for* $n = 1, 2, \cdots$. *Then,* $\prod_{n=1}^{m} (1 - b_n) \to 0$ *if and only if the series* $\sum b_n$ *diverges.*

(A proof of this lemma can be found in any good text on infinite series.)

Returning now to the iterative procedure described in §1, we prove the following elementary propositions.

PROPOSITION 1. *There is an integer* K *such that for all* $k \geqq K$, $\mu_k = \min \{\lambda_k, 1\} = \lambda_k < 1$.

Proof. Since $x_k \neq \hat{x}$ for any $k = 1, 2, \cdots$, it follows that $\mu_k > 0$ for $k = 1, 2, \cdots$, so that $f(x_{k+1}) < f(x_k)$ for $k = 1, 2, \cdots$. If $\mu_k = 1$, then $x_{k+1} = x_k + (a^{i_k} - x_k) = a^{i_k}$ which, in view of our above remark, can occur for only a finite number of k.

By Proposition 1, (1.2) and (1.3) can be written in the form

$$\begin{aligned} x_{k+1} &= x_k + \lambda_k s_k, \\ \xi_{k+1} &= (1 - \lambda_k)\xi_k + \lambda_k l_{i_k}, \end{aligned} \qquad k \geqq K, \tag{2.2}$$

where

$$\lambda_k = -\frac{\langle s_k, \nabla f(x_k) \rangle}{\langle s_k, Q s_k \rangle} \tag{2.3}$$

and

$$s_k = (a^{i_k} - x_k).$$

Substituting (2.3) into (2.2) we have, for $k \geqq K$,

$$f(x_{k+1}) = f(x_k) - \frac{1}{2}\frac{\langle s_k, \nabla f(x_k)\rangle^2}{\langle s_k, Qs_k\rangle} = f(x_k) - \frac{1}{2}\langle s_k, Qs_k\rangle\lambda_k^2. \tag{2.4}$$

PROPOSITION 2. *There is a constant β such that $\langle s_k, Qs_k\rangle \geqq \beta > 0$ for $k = 1, 2, \cdots$.*

Proof. Since Q is a positive definite symmetric matrix, it is well known that

$$\langle s_k, Qs_k\rangle \geqq \gamma \| s_k \|^2 \quad \text{for} \quad k = 1, 2, \cdots,$$

where $\gamma > 0$ is the smallest eigenvalue of Q. Thus, to prove the proposition, it suffices to show that there is a constant β such that $\| s_k \| \geqq \beta > 0$ for $k = 1, 2, \cdots$. We claim that $\hat{x} \neq a^i$ for any $i = 1, 2, \cdots, l$, for otherwise $\hat{x}$ would be an extreme point of Ω, and the iterative procedure would terminate in a finite number of steps. Hence, there is a β such that

$$\| a^i - \hat{x} \| \geqq 2\beta > 0 \quad \text{for} \quad i = 1, 2, \cdots, l.$$

Since $x_k \to \hat{x}$, the result now follows.

PROPOSITION 3. *The series $\sum \lambda_k$ diverges.*

Proof. The proof is by contradiction. Suppose that the series $\sum \lambda_k$ converges. By (2.2),

$$\xi_{k+1} \geqq \prod_{n=K}^{k}(1 - \lambda_n)\xi_K$$

for all $k \geqq K$. Moreover, by Proposition 1, we have that $0 \leqq \lambda_k < 1$ for all $k \geqq K$. Thus, by Lemma 2, $\prod_{n=K}^{k}(1 - \lambda_n) \to \delta > 0$, and so

$$\xi_{k+1} \geqq \delta\xi_K \tag{2.5}$$

for all $k \geqq K$. Recall that $\hat{x} \in \partial\Omega$, so that $\hat{x}$ belongs to some face (or edge) of the polyhedron Ω. Let a^i, $i \in I \subset \{1, 2, \cdots, l\}$, be the extreme points of Ω which also lie in this face. Now, by assumption, $x_k \in \dot{\Omega}$ for infinitely many $k \in \{1, 2, \cdots\}$. Consequently, there is an integer $L \geqq K$ and a ξ_L with

$$x_L = A\xi_L,$$
$$\sum \xi_L{}^i = 1,$$
$$\xi_L \geqq 0,$$

and such that for some $s \in \bar{I}$ (the complement of I) $\xi_L{}^s > 0$. To simplify notation we assume that $L = K$, so that

$$\xi_{k+1}^s \geqq \delta\xi_K{}^s > 0 \tag{2.6}$$

for all $k \geqq K$. The sequence $\{\xi_k\}_{k=1}^{\infty}$ is bounded and, therefore, has a convergent subsequence, say $\xi_{k_m} \to \hat{\xi}$. Since $x_{k_m} = A\xi_{k_m}$ with $\sum \xi^i_{k_m} = 1$ and $\xi_{k_m} \geqq 0$, and $x_{k_m} \to \hat{x}$, we have that $\hat{x} = A\hat{\xi}$ with $\sum \hat{\xi}^i = 1$ and $\hat{\xi} \geqq 0$. Using (2.6) we conclude that $\hat{\xi}^s > 0$ and, hence, that $s \in I$; but $s \in \bar{I}$, and so we have arrived at a contradiction.

Proof of Theorem 2. By (2.4), for $k \geqq K$,

$$f(x_k) - f(\hat{x}) = [f(x_K) - f(\hat{x})] - \frac{1}{2}\sum_{n=K}^{k-1} \langle s_n, Qs_n\rangle \lambda_n^2,$$

and, since $f(x_k) \to f(\hat{x})$, it follows that the right-hand side converges to zero. We may then write

$$f(x_k) - f(\hat{x}) = \frac{1}{2}\sum_{n=k}^{\infty} \langle s_n, Qs_n\rangle \lambda_n^2.$$

Let β be the positive constant envisaged in Proposition 2, and let $\alpha = \beta/2$. Then

$$\frac{f(x_k) - f(\hat{x})}{\alpha} \geqq \sum_{n=k}^{\infty} \lambda_n^2,$$

and, since the series $\sum \lambda_k$ diverges (Proposition 3), it follows from Lemma 1 that for any $\epsilon > 0$, $f(x_k) - f(\hat{x}) \geqq \alpha/k^{1+\epsilon}$ for infinitely many $k \in \{1, 2, \cdots\}$. It is now easy to see that this inequality must hold infinitely often for any real $\alpha > 0$ and $\epsilon > 0$.

3. Some further observations. In the preceding section it was shown that under usually satisfied assumptions the series $\sum \lambda_k$ diverges, and so, in the sense of Lemma 1, we obtained an upper bound on the rate at which $\sum_k^{\infty} \lambda_n^2$ tends to zero. It is clear, therefore, that the result of Theorem 2 remains valid under the simple hypothesis that the series $\sum \lambda_k$ diverges. The following example illustrates a case when this series converges.

Example. Take $n = 2$, $l = 3$, and

$$a^1 = \text{col } (0, 1), \quad a^2 = \text{col } (2, 0), \quad a^3 = \text{col } (-2, 0).$$

Let

$$f_\epsilon(x) = (x^1)^2 + (x^2 - \epsilon)^2,$$

where $0 \leqq \epsilon \leqq 1$ and will be specified later. We denote this problem by P_ϵ. The solution $\hat{x}_\epsilon$ to P_ϵ is clearly given by $\hat{x}_\epsilon = (0, \epsilon)$, and the level sets will be circles centered at $(0, \epsilon)$.

Case 1. Let $0 < \epsilon < 1$. First note that $\hat{x}_\epsilon$ lies in the interior of the convex hull of the vectors a^1, a^2, a^3, and that it has a unique representation as a convex combination of the a^i, i.e.,

$$\hat{x}_\epsilon = \epsilon a^1 + \tfrac{1}{2}(1 - \epsilon)a^2 + \tfrac{1}{2}(1 - \epsilon)a^3.$$

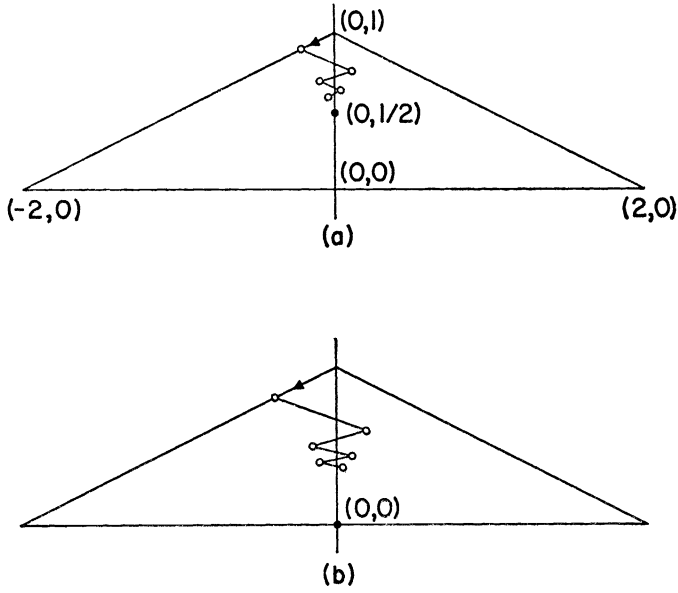


FIG. 1

As a starting point for the iterative procedure, we take $x_1 = a^1 = \text{col } (0, 1)$, i.e.,

$$\xi_1 = \text{col } (1, 0, 0).$$

In the case $\epsilon = \frac{1}{2}$, the first few iterates x_k are shown in Fig. 1(a). The reader should have no difficulty in seeing that for every $k = 1, 2, \cdots$, and every $0 < \epsilon < 1$, either $s_k = (a^2 - x_k)$ or $s_k = (a^3 - x_k)$, or equivalently, $s_k \neq (a^1 - x_k)$ for any $k = 1, 2, \cdots$. It is also clear that $0 < \lambda_k(\epsilon) < 1$, $k = 1, 2, \cdots$, and that the iterative procedure does not terminate in a finite number of steps. By (1.3),

$$\xi_{k+1}^1 = (1 - \lambda_k)\xi_k^1, \qquad k = 1, 2, \cdots,$$

and so, recalling that $\xi_1^1 = 1$, we have that

$$\xi_k^1 = \prod_{n=1}^{k-1} (1 - \lambda_n).$$

Since $\hat{x}$ has a unique representation, $\hat{\xi}$, $\xi_k \to \hat{\xi}$. Since $\hat{\xi}^1 = \epsilon$, it follows that

$$\prod_{n=1}^{k-1} (1 - \lambda_n(\epsilon)) \to \epsilon,$$

so that by Lemma 1, the series $\sum \lambda_k(\epsilon)$ converges for every $0 < \epsilon < 1$. Indeed, it can be shown that the rate at which $f_\epsilon(x_k) - f_\epsilon(\hat{x}_\epsilon)$ converges to zero is geometric.

Case 2. Let $\epsilon = 0$. In this case (see Fig. 1(b)), $\hat{x}_\epsilon = (0, 0)$ belongs to the boundary of Ω. If the iterative procedure is again initiated at $x_1 = a^1$, then it is clear that $x_k \in \dot{\Omega}$ for all $k = 2, 3, \cdots$. By Proposition 3, for $\epsilon = 0$, the series $\sum \lambda_k(\epsilon)$ diverges.

Concerning the above example, one may conjecture that, loosely speaking, the rate at which $f_\epsilon(x_k) - f_\epsilon(\hat{x})$ tends to zero becomes more and more like 1/k as ϵ tends to zero.

Acknowledgment. The authors wish to thank their colleague D. Chazan for supplying the proof for Lemma 1.

REFERENCE

[1] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

EXISTENCE THEOREMS FOR OPTIMAL CONTROLS OF THE MAYER TYPE*

LAMBERTO CESARI†

In the present paper we give existence theorems for weak and usual solutions of (one-dimensional problems) of calculus of variations and optimal control written in the Mayer form.

In previous papers [1], [2] we have given a few of such theorems for problems written in the Lagrange form. These theorems were based on certain properties of variable sets in E_n, particularly on a modification of Kuratowski's concept of upper semicontinuity, which we denoted as property (Q). Subsequently this property (Q) has been investigated by others [3], [6], [9]. We required the property (Q) as a hypothesis in our "Closure Theorem II", as well as in the ensuing existence theorems in [1], [2]. Closure Theorem II is connected with the use of Helly's selection process on certain components of a minimizing sequence in our paper [1] (the same Helly's selection process has been subsequently used for the same purpose also by McShane [7] and Nishiura [8]).

In the present paper we show first (§2) that certain growth conditions as those used in our earlier paper [1], as well as others used in the papers by McShane [7] and Nishiura [8], necessarily imply property (Q), or variants of it, for the sets under investigation. We also show that a much weaker form of the same growth condition (§2, §4), though not strong enough to assure property (Q), is still capable of assuring the equiabsolute continuity of the components of the minimizing sequence for which not Helly's but Ascoli's selection process is being used. In §3 we prove a "Closure Theorem III" which extends slightly our previous Closure Theorem II of [1] under a variant of property (Q).

In §4 and §5 we prove new existence theorems for optimal usual and weak solutions of Mayer-type problems of the calculus of variations and optimal control, and again in these theorems we require property (Q), or variants of it, as a hypothesis. These properties (Q) are certainly satisfied under the growth conditions mentioned above. The existence theorems of §4 and §5 extend in a number of ways previous existence theorems. As in [1] we first prove existence theorems for usual optimal solutions (§4), and then we obtain existence theorems for weak solutions (Gamkrelidze's sliding regimes) as simple corollaries (§5).

* Received by the editors January 29, 1968, and in revised form April 16, 1968.

† Department of Mathematics, The University of Michigan, Ann Arbor, Michigan 48104. This research was supported in part by US-AFOSR Grant 942-65 at The University of Michigan.

1. The concept of upper semicontinuity of variable sets and some of its variants. As usual let t, x, u be variables, $t \in E_1$, $x = (x^1, \cdots, x^n) \in E_n$, $u = (u^1, \cdots, u^m) \in E_m$. Given any set F in a Euclidean space, we shall denote by $\operatorname{cl} F$ and $\operatorname{co} F$ the closure and the convex hull of F, respectively, so that $\operatorname{cl} \operatorname{co} F$ denotes the closure of the convex hull of F.

Let A be a fixed closed subset of the tx-space $E_1 \times E_n$, and, for every $(t, x) \in A$, let $U(t, x)$ be a given subset of E_m. For every $(\bar{t}, \bar{x}) \in A$ and $\delta > 0$, let $N_\delta(\bar{t}, \bar{x})$ denote the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from $(\bar{t}, \bar{x})$, and let $U(\bar{t}, \bar{x}, \delta) = \bigcup U(t, x)$, where the union $\bigcup$ is taken over all $(t, x) \in N_\delta(\bar{t}, \bar{x})$. Thus $U(\bar{t}, \bar{x}) \subset U(t, x, \delta)$ for every $\delta > 0$.

The set $U(t, x)$ is said to be upper semicontinuous (in the sense of Kuratowski) at the point $(\bar{t}, \bar{x}) \in A$, or, briefly, $U(t, x)$ *has property* (U) *at* $(\bar{t}, \bar{x})$, provided

$$U(\bar{t}, \bar{x}) = \bigcap_\delta \operatorname{cl} U(t, x, \delta). \tag{1}$$

We say that $U(t, x)$ has property (U) in A if $U(t, x)$ has property (U) at every point $(\bar{t}, \bar{x}) \in A$. A set $U(t, x)$ possessing property (U) is necessarily closed as the intersection of closed sets. Note that in verifying property (U) at $(\bar{t}, \bar{x})$ all we have to prove is that

$$U(\bar{t}, \bar{x}) \supset \bigcap_\delta \operatorname{cl} U(\bar{t}, \bar{x}, \delta),$$

since the opposite inclusion is trivial. In other words, we have to prove that $\bar{z} \in \bigcap_\delta \operatorname{cl} U(\bar{t}, \bar{x}, \delta)$ implies $\bar{z} \in U(\bar{t}, \bar{x})$.

In [1] and other papers we made extensive use of property (U), together with an analogous property, or property (Q), for convex closed sets. For every $(t, x) \in A$ let $Q(t, x)$ be a given subset of the x-space E_n. We say that $Q(t, x)$ *has property* (Q) *at the point* $(\bar{t}, \bar{x}) \in A$ provided

$$Q(\bar{t}, \bar{x}) = \bigcap_\delta \operatorname{cl} \operatorname{co} Q(\bar{t}, \bar{x}, \delta). \tag{2}$$

We say that $Q(t, x)$ has property (Q) in A if it has the property above at every point $(\bar{t}, \bar{x}) \in A$. A set $Q(t, x)$ possessing property (Q) is necessarily closed and convex as the intersection of closed convex sets. Note that, in verifying property (Q) at $(\bar{t}, \bar{x})$, all we have to prove is that

$$Q(\bar{t}, \bar{x}) \supset \bigcap_\delta \operatorname{cl} \operatorname{co} Q(\bar{t}, \bar{x}, \delta),$$

since the opposite inclusion is trivial. In other words, we have to prove that $\bar{z} \in \bigcap_\delta \operatorname{cl} \operatorname{co} Q(\bar{t}, \bar{x}, \delta)$ implies $\bar{z} \in Q(\bar{t}, \bar{x})$. In [1] we used property (Q) as one of the assumptions in Closure Theorem II [1, p. 386], as well as in the existence theorems, and analogously we use property (Q) in Closure Theorem III of the present paper, and in the existence theorems.

If we interpret Closure Theorem II (or III) as a property of lower semicontinuity (see Remark 8 of §3.1 below), then Brunovsky [3] has proved that the convexity requirement in property (Q) is not only sufficient, but

also necessary, under suitable restrictions, for lower semicontinuity. For further use of property (Q) see also La Palm [6] and Olech and Lasota [9].

We should note in passing that since $Q(t, x) \subset E_n$, $Q(\bar{t}, \bar{x}, \delta) \subset E_n$, the set $\operatorname{co} Q(\bar{t}, \bar{x}, \delta)$ above is the union of all points $z \in E_n$ of the form $z = \sum \lambda_\gamma z_\gamma$, where $\sum$ ranges over all $\gamma = 1, \cdots, \nu$, where $\lambda_\gamma \geqq 0$, $\gamma = 1, \cdots, \nu$, $\lambda_1 + \cdots + \lambda_\nu = 1$, and $z_\gamma \in Q(t_\gamma, x_\gamma)$, $(t_\gamma, x_\gamma) \in N_\delta(\bar{t}, \bar{x})$, $\gamma = 1, \cdots, \nu$, for some ν, say, $\nu = n + 1$.

In the present paper we shall need a slight variant of property (Q).

For every $(t, x) \in A$ let $\tilde{Q}(t, x)$ be a given subset of the $\tilde{z}$-space E_{n+1}, $\tilde{z} = (z^0, z) = (z^0, z^1, \cdots, z^n)$, and let $Q(t, x)$ denote the projection of $\tilde{Q}(t, x)$ on the z-space E_n. We say that $Q(t, z)$ *has property* (Q) *with respect to* $\tilde{Q}(t, x)$ *at the point* $(\bar{t}, \bar{x}) \in A$ provided

$$\tilde{z} = (\bar{z}^0, \bar{z}) \in \bigcap_\delta \operatorname{cl} \operatorname{co} \tilde{Q}(\bar{t}, \bar{x}, \delta) \tag{3}$$

implies $\bar{z} \in Q(\bar{t}, \bar{x})$. We say that $Q(t, x)$ has property (Q) with respect to $\tilde{Q}(t, x)$ in A if it has the above property at every point $(\bar{t}, \bar{x}) \in A$. Note that (3) implies trivially that

$$\bar{z} \in \bigcap_\delta \operatorname{cl} \operatorname{co} Q(\bar{t}, \bar{x}, \delta),$$

since the latter relation is the projection of (3) in E_n. For the sake of brevity we may refer to the property (Q) relative to $\tilde{Q}(t, x)$ as property (Q_r).

(i) If $Q(t, x)$ satisfies property (Q), then $Q(t, x)$ satisfies property (Q) relative to $\tilde{Q}(t, x)$.

Indeed for property (Q_r) we must only verify that $\bar{z} \in Q(\bar{t}, \bar{x})$ when both $\bar{z} \in \bigcap_\delta \operatorname{cl} \operatorname{co} Q(\bar{t}, \bar{x}, \delta)$ and (3) holds. Thus property (Q_r) is a weakening of property (Q) for the sets $Q(t, x)$. In other words, property (Q) for $Q(t, x)$ implies property (Q_r) for $Q(t, x)$.

(ii) If $\tilde{Q}(t, x)$ satisfies property (Q), then $Q(t, x)$ satisfies property (Q) relative to $\tilde{Q}(t, x)$.

Indeed if $\tilde{Q}(t, x)$ has property (Q), then (3) implies $\tilde{z} = (\bar{z}^0, \bar{z}) \in \tilde{Q}(\bar{t}, \bar{x})$, and hence $\bar{z} \in Q(\bar{t}, \bar{x})$ since Q is simply the projection of $\tilde{Q}$ on the z-space. In other words, property (Q) for $\tilde{Q}(t, x)$ implies property (Q_r) for $Q(t, x)$.

The following trivial example shows that property (Q_r) implies property (Q) neither for $Q(t, x)$ nor for $\tilde{Q}(t, x)$.

Indeed, take $A = [0 \leqq t \leqq 1, 0 \leqq x \leqq 1]$, $n = 1$, and let us write z^0, z instead of z^0, z^1. Let us take $\tilde{Q} = [(z^0, z) | z^0 = z^{-1}, 0 < z \leqq 1]$, $Q = [z | 0 < z \leqq 1]$, both $\tilde{Q}$ and Q independent of (t, x). We have here $\operatorname{co} \tilde{Q}(\delta) = [(z^0, z) | z^0 \geqq z^{-1} \text{ for } 0 < z < 1, \text{ or } (z^0 = 1, z = 1)]$, $\operatorname{cl} \operatorname{co} \tilde{Q}(\delta) = E$, $\bigcap_\delta \operatorname{cl} \operatorname{co} \tilde{Q}(\delta) = E$, where $E = [(z^0, z) | z^0 \geqq z^{-1}, 0 < z \leqq 1]$. Thus $(2, 1) \in E$, but $(2, 1) \notin \tilde{Q}$. Analogously, $\operatorname{co} Q(\delta) = [z | 0 < z \leqq 1]$, $\operatorname{cl} \operatorname{co} Q(\delta) = F$, $\bigcap_\delta \operatorname{cl} \operatorname{co} Q(\delta) = F$, where $F = [z | 0 \leqq z \leqq 1]$. Thus $(z = 0) \in F$ but $(z = 0) \notin Q$. Thus, neither $\tilde{Q}$ has property (Q), nor Q has property (Q).

Nevertheless, Q has property (Q_r). Indeed, if $(z^0, z) \in E$, then $z^0 \geqq z^{-1}$, $0 < z \leqq 1$, and thus $z \in Q$.

We give below a few criteria for properties (Q) and (Q_r).

2. Some criteria for properties (Q) and (Q_r).

2.1. We have mentioned in [1] that certain hypotheses, analytic in character, guarantee property (Q) for some particular sets $Q(t, x)$ which were considered in [1] (see, for instance, [1, p. 379, (xvi), and p. 405, (ii)]. Below we shall give other examples of greater generality of criteria for properties (Q) and (Q_r).

Let A be a given set of the tx-space, for every $(t, x) \in A$ let $U(t, x)$ be a given subset of E_m, and let M be the set of all $(t, x, u) \in E_{1+n+m}$ with $(t, x) \in A$, $u \in U(t, x)$. Given two functions $\psi(t, x, u) = (\psi_1, \cdots, \psi_s)$, $G(t, x, u)$ in M with $G \geqq 0$, we say that ψ is of slower growth than G as $|u| \to \infty$ uniformly in A provided, given $\epsilon > 0$, there is some $\bar{u} = \bar{u}(\epsilon) > 0$ such that $(t, x) \in A$, $u \in U(t, x)$, $|u| \geqq \bar{u}$ implies $|\psi(t, x, u)| \leqq \epsilon G(t, x, u)$.

(i) Let A be a given compact subset of the tx-space E_{n+1}, for every $(t, x) \in A$ let $U(t, x)$ be a given subset of the u-space E_m, and let M be the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$. Let $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ be given continuous functions on M, and for every $(t, x) \in A$ let $\tilde{\tilde{Q}}(t, x)$ be the set of all $(z^0, z) \in E_{n+1}$ such that $z^0 \geqq f_0(t, x, u)$, $z = f(t, x, u)$, $u \in U(t, x)$. If $f_0 \geqq 0$, if 1 is of slower growth than f_0 uniformly in A, if f is of slower growth than f_0 uniformly in A, if $U(t, x)$ has property (U) in A, and $\tilde{\tilde{Q}}(t, x)$ is convex for every $(t, x) \in A$, then $\tilde{\tilde{Q}}(t, x)$ has property (Q) in A.

Remark 1. Statement (i) is a modification of [1, p. 405, (ii)] and is a particular case of (ii) below as we shall show in Remark 4.

Remark 2. If $f_0 \geqq -M_0$ for some constant $M_0 > 0$, then statement (i) still holds provided 1 and f are known to be of slower growth than $f_0 + M_0$ uniformly in A. Because of statement (i) and this Remark, the requirement made explicitly in Theorem I of [1, p. 390] that the sets $\tilde{\tilde{Q}}(t, x)$ defined there possess property (Q) in A is now a consequence of the other hypotheses of the same Theorem I.

Remark 3. It is easy to verify that $U(t, x)$ has property (U) in any closed set A if and only if the set M is closed. Thus, the hypothesis of the closure of M very well replaces the property (U) of the sets $U(t, x)$ in statement (i) as well as in all other statements of the present paper.

2.2. Let $f(t, x, u) = (f_1, \cdots, f_n)$, $G(t, x, u)$ be functions defined on M, and let the components f_i of f be divided into two sets: f_i, $i = 1, \cdots, \alpha$, and f_i, $i = \alpha + 1, \cdots, n$, for some $0 \leqq \alpha \leqq n$, so that either set may be empty. Let $z = (z^1, \cdots, z^n) \in E_n$, $\bar{z} = (z^0, z) = (z^0, z^1, \cdots, z^n) \in E_{n+1}$ denote variable points in E_n and E_{n+1}. For every $(t, x) \in A$ let $Q(t, x)$,

$Q_\sigma(t, x)$ denote the sets

$$Q(t, x) = [z \mid z^i = f_i(t, x, u),\ i = 1, \cdots, \alpha,\ z^i \geqq f_i(t, x, u),$$
$$i = \alpha + 1, \cdots, n,\ u \in U(t, x)] \subset E_n,$$

$$(4)\quad Q_\sigma(t, x) = [z = (z^0, z) \mid z^0 \geqq G(t, x, u),\ z^i = f_i(t, x, u),$$
$$i = 1, \cdots, \alpha,\ z^i \geqq f_i(t, x, u),\ i = \alpha + 1, \cdots, n,$$
$$u \in U(t, x)] \subset E_{n+1}.$$

Thus, Q is the projection of Q_σ on the z-space E_n.

(ii) Let A be a given compact subset of the tx-space E_{n+1}, and for every $(t, x) \in A$ let $U(t, x)$ be a given closed subset of the u-space E_m. Let M be the set of all $(t, x, u) \in E_{1+n+m}$ with $(t, x) \in A$, $u \in U(t, x)$. Let $G(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ be given functions defined on M, let $G, f_{\alpha+1}, \cdots, f_n$ be nonnegative and lower semicontinuous on M, and let $f_1, \cdots, f_\alpha$ be continuous on M. Let $1, f_1, \cdots, f_\alpha$ be of slower growth than G (as $\mid u \mid \to \infty$) uniformly on A, and assume that $U(t, x)$ satisfies property (U) in A. If the set $Q_\sigma(t, x)$ is convex for every $(t, x) \in A$, then $Q_\sigma(t, x)$ satisfies property (Q) in A. If only $Q(t, x)$ is known to be convex for every $(t, x) \in A$, then $Q(t, x)$ satisfies property (Q) with respect to $Q_\sigma(t, x)$ in A.

Remark 4. This statement (ii) contains (i) as a particular case, as can be seen by taking $G = f_0$, $\alpha = n$. The proof of (ii) which is given below is a modification of the one in [1, p. 405, (ii)].

Proof of (ii). We have to prove that $\bar{z} = (z^0, z) \in \bigcap_\delta \operatorname{cl} \operatorname{co} Q_\sigma(\bar{t}, \bar{x}, \delta)$ implies $\bar{z} \in Q_\sigma(\bar{t}, \bar{x})$, where Q, Q_σ are defined in (4), and $Q_\sigma(\bar{t}, \bar{x}, \delta)$, $= \bigcup Q_\sigma(t, x)$, where $\bigcup$ ranges over all $(t, x) \in N_\delta(\bar{t}, \bar{x})$. In the second alternative we have to prove that the same assumption implies $z \in Q(\bar{t}, \bar{x})$.

Let $\bar{\bar{z}}$ be a given point $\bar{\bar{z}} = (\bar{z}^0, \bar{z}) \in \bigcap_\delta \operatorname{cl} \operatorname{co} Q_\sigma(\bar{t}, \bar{x}, \delta)$, and let us prove that $\bar{\bar{z}} \in Q_\sigma(\bar{t}, \bar{x})$. For every $\delta > 0$ we have $\bar{\bar{z}} = (\bar{z}^0, \bar{z}) \in \operatorname{cl} \operatorname{co} Q_\sigma(\bar{t}, \bar{x}, \delta)$, and thus for every $\delta > 0$, there are points $\tilde{z} = (z^0, z) \in \operatorname{co} Q_\sigma(\bar{t}, \bar{x}, \delta)$ at a distance as small as we want from $\bar{\bar{z}} = (\bar{z}^0, \bar{z})$. Thus, there is a sequence of numbers $\delta_k > 0$ and of points $\tilde{z}_k = (z_k^0, z_k) \in \operatorname{co} Q_\sigma(\bar{t}, \bar{x}, \delta_k)$ such that $\delta_k \to 0$, $\tilde{z}_k \to \bar{\bar{z}}$ as $k \to \infty$. In other words, for every integer k there is a system of points $(t_k^\gamma, x_k^\gamma) \in N_\delta(\bar{t}, \bar{x})$, $\gamma = 1, \cdots, \nu$, say $\nu = n + 2$, corresponding points $\tilde{z}_k^\gamma = (z_k^{0\gamma}, z_k^\gamma) \in Q_\sigma(t_k^\gamma, x_k^\gamma)$, points $u_k^\gamma \in U(t_k^\gamma, x_k^\gamma)$, and numbers $\lambda_k^\gamma \geqq 0$, $\gamma = 1, \cdots, \nu$, such that

$$(5)\quad \begin{aligned} 1 &= \sum \lambda_k^\gamma, \qquad \tilde{z}_k = \sum \lambda_k^\gamma \tilde{z}_k^\gamma, \\ z_k^0 &= \sum \lambda_k^\gamma z_k^{0\gamma}, \qquad z_k^i = \sum \lambda_k^\gamma z_k^{i\gamma}, \qquad i = 1, \cdots n, \\ z_k^{0\gamma} &\geqq G(t_k^\gamma, x_k^\gamma, u_k^\gamma), \\ z_k^{i\gamma} &= f_i(t_k^\gamma, x_k^\gamma, u_k^\gamma), \qquad i = 1, \cdots, \alpha, \\ z_k^{i\gamma} &\geqq f_i(t_k^\gamma, x_k^\gamma, u_k^\gamma), \qquad i = \alpha + 1, \cdots, n, \\ z_k &= (z_k^1, \cdots, z_k^n), \qquad z_k^\gamma = (z_k^{1\gamma}, \cdots, z_k^{n\gamma}), \end{aligned}$$

where $\gamma = 1, \cdots, \nu$, where $\sum$ ranges over all $\gamma = 1, \cdots, \nu$, and $t_k^\gamma \to \bar{t}$, $x_k^\gamma \to \bar{x}$, $\tilde{z}_k \to \tilde{\bar{z}}$, $z_k^0 \to \bar{z}^0$, $z_k \to \bar{z}$, $z_k^i \to \bar{z}^i$, $i = 1, \cdots, n$, $\gamma = 1, \cdots, \nu$, as $k \to \infty$.

The numbers λ_k^γ are all between 0 and 1; hence we can extract a subsequence, say still $[k]$, such that $\lambda_k^\gamma \to \lambda^\gamma$ as $k \to \infty$, $\gamma = 1, \cdots, \nu$, and hence, $0 \leqq \lambda^\gamma \leqq 1$, $\sum \lambda^\gamma = 1$. The third relation of (5), where $z_k^0 \to \bar{z}^0$, $z_k^0 \geqq 0$, $\bar{z}^0 \geqq 0$, $z_k^{0\gamma} \geqq 0$, shows that for every k at least one of the numbers $z_k^{0\gamma}$ is between 0 and z_k^0 and therefore forms a bounded sequence. We may well assume, by a suitable reindexing, that this occurs for the same index γ. By a suitable extraction and further reindexing we may assume that certain sequences $[z_k^{0\gamma}, k = 1, 2, \cdots]$, $\gamma = 1, \cdots, \nu'$, are bounded, $1 \leqq \nu' \leqq \nu$, while the remaining sequences $[z_k^{0\gamma}, k = 1, 2, \cdots]$, $\gamma = \nu' + 1, \cdots, \nu$, are unbounded, and actually $z_k^{0\gamma} \to +\infty$ as $k \to \infty$, $\gamma = \nu' + 1, \cdots, \nu$. Then the fifth relation of (5) shows that the corresponding sequences $[u_k^\gamma, k = 1, 2, \cdots]$, $\gamma = 1, \cdots, \nu'$, are also bounded sequences because of the assumed property of growth of G. We can further select a subsequence, say still [k], such that $u_k^\gamma \to u^\gamma$, $z_k^{0\gamma} \to z^{0\gamma}$ as $k \to \infty$, $\gamma = 1, \cdots, \nu'$. Now $u_k^\gamma \in U(t_k^\gamma, x_k^\gamma)$ with $t_k^\gamma \to \bar{t}$, $x_k^\gamma \to \bar{x}$ as $k \to \infty$, $\gamma = 1, \cdots, \nu$, and $u_k^\gamma \to u^\gamma$ as $k \to \infty$, $\gamma = 1, \cdots, \nu'$, and consequently, $u^\gamma \in \operatorname{cl} U(\bar{t}, \bar{x}, \sigma)$ for every $\sigma > 0$ and all k sufficiently large. Hence $u^\gamma \in \bigcap_\sigma \operatorname{cl} U(\bar{t}, \bar{x}, \sigma) = U(\bar{t}, \bar{x})$ since U satisfies property (U) at $(\bar{t}, \bar{x})$, and finally $(\bar{t}, \bar{x}, u^\gamma) \in M$, $\gamma = 1, \cdots, \nu'$. Since G is lower semicontinuous and $f_1, \cdots, f_\alpha$ are continuous on M, the fifth and sixth relations of (5) as $k \to \infty$ yield

$$z_k^{0\gamma} \to z^{0\gamma} \geqq G(\bar{t}, \bar{x}, u^\gamma),$$
$$z_k^{i\gamma} \to z^{i\gamma} = f_i(\bar{t}, \bar{x}, u^\gamma), \qquad i = 1, \cdots, \alpha, \quad \gamma = 1, \cdots, \nu'.$$

For $\gamma = \nu' + 1, \cdots, \nu$, we have $z_k^{0\gamma} \to +\infty$ as $k \to \infty$, and the third relation of (5), where all numbers λ_k^γ, $z_k^{0\gamma}$ are nonnegative, shows that $\lambda_k^\gamma \to \lambda^\gamma = 0$ as $k \to \infty$, $\gamma = \nu' + 1, \cdots, \nu$. If we denote by $\sum'$, $\sum''$ sums ranging over $\gamma = 1, \cdots, \nu'$, and $\gamma = \nu' + 1, \cdots, \nu$, respectively, then the first relation of (5) shows now that $1 = \sum' \lambda^\gamma$, since $\lambda^\gamma = 0$ for $\gamma = \nu' + 1, \cdots, \nu$. The third relation of (5), where $z_k^0 \to \bar{z}^0$, and all numbers λ_k^γ, $z_k^{0\gamma}$ are nonnegative, shows that the sequences $[\lambda_k^\gamma z_k^{0\gamma}, k = 1, 2, \cdots]$, $\gamma = \nu' + 1, \cdots, \nu$, are bounded. Thus, we can take the subsequence, say still $[k]$, such that $\lambda_k^\gamma z_k^{0\gamma} \to A_\gamma \geqq 0$ as $k \to \infty$. Then, taking the limit as $k \to \infty$ on the same third relation (5), we obtain

$$z^0 = \sum' \lambda^\gamma z^{0\gamma} + \sum'' A_\gamma \geqq \sum' \lambda^\gamma z^{0\gamma}.$$

Since $z_k^{0\gamma} \to +\infty$ as $k \to +\infty$, and $z_k^{0\gamma} \geqq G(t_k^\gamma, x_k^\gamma, u_k^\gamma)$ for $\gamma = \nu' + 1, \cdots, \nu$, then, for each such γ, either $G(t_k^\gamma, x_k^\gamma, u_k^\gamma)$ is unbounded, and then also u_k^γ is unbounded, and we can take the subsequence so that $|u_k^\gamma| \to +\infty$, $G(t_k^\gamma, x_k^\gamma, u_k^\gamma) \to +\infty$ as $k \to +\infty$; or $G(t_k^\gamma, x_k^\gamma, u_k^\gamma)$ is

bounded, and then u_k^γ is also bounded, and we can take the subsequence so that $u_k^\gamma \to u^\gamma$ as $k \to \infty$, and $G(\bar{t}, \bar{x}, u^\gamma) = z^{0\gamma} \leqq \liminf G(t_k, x_k, u_k^\gamma)$ as $k \to \infty$. Now f_i, $i = 1, \cdots, \alpha$, are of slower growth than G; hence $z_k^{i\gamma} = \epsilon_k^{i\gamma} z_k^{0\gamma}$, $i = 1, \cdots, \alpha$, where $\epsilon_k^{i\gamma} \to 0$ as $k \to \infty$, $\gamma = \nu' + 1, \cdots, \nu$. The fourth relation of (5) then becomes

$$z_k^i = \sum{}' \lambda_k^\gamma z_k^{i\gamma} + \sum{}'' \epsilon_k^{i\gamma} (\lambda_k^\gamma z_k^{0\gamma}),$$

where $\lambda_k^\gamma z_k^\gamma \to A_\gamma \geqq 0$, $\epsilon_k^{i\gamma} \to 0$ as $k \to \infty$, $\gamma = \nu' + 1, \cdots, \nu$, and finally, as $k \to \infty$, we obtain

$$z^i = \sum{}' \lambda^\gamma z^{i\gamma}, \qquad i = 1, \cdots, \alpha.$$

Thus

$$z^{0\gamma} \geqq G(\bar{t}, \bar{x}, u^\gamma), \qquad z^0 \geqq \sum{}' \lambda^\gamma z^{0\gamma},$$

$$z^{i\gamma} = f_i(t, x, u^\gamma), \qquad i = 1, \cdots, \alpha, \quad \gamma = 1, \cdots, \nu',$$

$$1 = \sum{}' \lambda^\gamma, \qquad \bar{z}^i = \sum{}' \lambda^\gamma z^{i\gamma}, \qquad i = 1, \cdots, \alpha.$$

Since $\lambda^\gamma \geqq 0$ for $\gamma = 1, \cdots, \nu'$, $\lambda^\gamma = 0$ for $\gamma = \nu' + 1, \cdots, \nu$, by a suitable reindexing we have $\lambda^\gamma > 0$ for $\gamma = 1, \cdots, \nu^*$, $\lambda^\gamma = 0$ for $\gamma = \nu^* + 1, \cdots, \nu$ with $\nu^* \leqq \nu'$. On the other hand, the relation $1 = \sum{}' \lambda^\gamma$ shows that $\nu^* \geqq 1$. Thus $1 \leqq \nu^* \leqq \nu' \leqq \nu$. If we denote by $\sum{}^*$ any sum ranging over all $\gamma = 1, \cdots, \nu^*$, we have $1 = \sum{}^* \lambda^\gamma$. Since $\lambda_k^\gamma \to \lambda^\gamma > 0$ as $k \to \infty$ for $\gamma = 1, \cdots, \nu^*$, we have also $\lambda_k^\gamma \geqq \lambda^\gamma/2$ for $\gamma = 1, \cdots, \nu^*$ and all k sufficiently large. For $i = \alpha + 1, \cdots, n$ in the fourth relation of (5), or

$$z_k^i = \sum \lambda_k^\gamma z_k^{i\gamma}, \qquad i = \alpha + 1, \cdots, n,$$

we have $z_k^i \geqq 0$, $z_k^{i\gamma} \geqq 0$, $\lambda_k^\gamma \geqq 0$, and hence,

$$z_k^i \geqq \sum{}^* \lambda_k^\gamma z_k^{i\gamma}, \qquad i = \alpha + 1, \cdots, n,$$

where $z_k^i \to \bar{z}^i$ and $\lambda_k^\gamma \geqq \lambda^\gamma/2 > 0$ for k large. This implies that the sequences $[z_k^{i\gamma}, k = 1, 2, \cdots]$, $\gamma = 1, \cdots, \nu^*$, $i = \alpha + 1, \cdots, n$, are bounded. There is, therefore, a subsequence, say still $[k]$, such that $z_k^{i\gamma} \to z^{i\gamma} \geqq 0$ as $k \to \infty$, $i = \alpha + 1, \cdots, n$, and the seventh relation of (5) and the lower semicontinuity of the functions $f_{\alpha+1}, \cdots, f_n$ yield, as $k \to \infty$,

$$z_k^{i\gamma} \to z^{i\gamma} \geqq f_i(\bar{t}, \bar{x}, u^\gamma), \qquad i = \alpha + 1, \cdots, n, \quad \gamma = 1, \cdots, \nu^*.$$

Finally, the fourth relation of (5) yields $\bar{z}^i \geqq \sum{}^* \lambda^\gamma z^{i\gamma}$, $i = \alpha + 1, \cdots, n$. Thus, we have

$$z^{i\gamma} = f_i(\bar{t}, \bar{x}, u^\gamma), \qquad i = 1, \cdots, \alpha, \quad \gamma = 1, \cdots, \nu^*,$$

$$z^{i\gamma} \geqq f_i(\bar{t}, \bar{x}, u^\gamma), \qquad i = \alpha + 1, \cdots, n, \quad \gamma = 1, \cdots, \nu^*,$$

$$
\begin{aligned}
\bar{z}^0 &\geqq \sum{}^* \lambda^\gamma z^{0\gamma}, \\
\bar{z}^i &= \sum{}^* \lambda^\gamma z^{i\gamma}, && i = 1, \cdots, \alpha, \\
\bar{z}^i &\geqq \sum{}^* \lambda^\gamma z^{i\gamma}, && i = \alpha + 1, \cdots, n, \\
1 &= \sum{}^* \lambda^\gamma.
\end{aligned}
$$

These relations show that $\tilde{\bar{z}} = (\bar{z}^0, \bar{z}) = (\bar{z}^0, \bar{z}^1, \cdots, \bar{z}^n) \in \operatorname{co} Q_G(\bar{t}, \bar{x})$, and since Q_G is by hypothesis convex, we have $\tilde{\bar{z}} \in Q_G(\bar{t}, \bar{x})$; hence Q_G possesses property (Q). If $Q(t, x)$ only is known to be convex, then the same relations above show that $\bar{z} = (\bar{z}^1, \cdots, \bar{z}^n) \in \operatorname{co} Q(\bar{t}, \bar{x})$; hence $\bar{z} \in Q(\bar{t}, \bar{x})$, and Q possesses property (Q_r). Statement (ii) is thereby proved.

2.3. For existence theorems for weak solutions we shall need a variant of statement (ii) above. Using the same notations as in §2.2, let $p = (p_1, \cdots, p_\mu)$, $\mu = 1, \cdots, n + 1$, be a new variable satisfying $p \in \Gamma = [p \mid p_j \geqq 0, j = 1, \cdots, \mu, p_1 + \cdots + p_\mu = 1]$, where Γ is the simplex so defined. Let $v = (u^{(1)}, \cdots, u^{(\mu)})$ denote a new variable, where each $u^{(j)}$ is an m-vector required to vary in $U(t, x)$; thus $u^{(j)} \in U(t, x), j = 1, \cdots, \mu$, or equivalently, $v \in [U(t, x)]^\mu$. Let $z = (z^1, \cdots, z^n) \in E_n$, $\tilde{z} = (z^0, z) = (z^0, z^1, \cdots, z^n) \in E_{n+1}$. As in §2.2 let $f(t, x, u) = (f_1, \cdots, f_n)$ $G(t, x, u)$ be functions defined on M, let α be an integer, $\alpha \in \{0, \cdots, n\}$, and, let f^*, G^* be the new functions

$$
\begin{aligned}
f^*(t, x, p, v) &= (f_1{}^*, \cdots, f_n{}^*) = \sum_j p_j f(t, x, u^{(j)}), \\
G^*(t, x, p, v) &= \sum_j p_j G(t, x, u^{(j)}),
\end{aligned} \tag{6}
$$

where $\sum_j$ ranges over all $j = 1, \cdots, \mu$. Then G^* and f^* are defined on the set M^* of all $(t, x, p, v) \in E_{1+n+\mu+\mu m}$ with $(t, x) \in A$, $p \in \Gamma$, $v \in U^\mu(t, x)$. Note that, if G, or a given f_i, is continuous on M, then G^*, $f_i{}^*$ are continuous on M^*. If G, or f_i, is nonnegative and lower semicontinuous on M, then G^*, $f_i{}^*$ are nonnegative and lower semicontinuous on M^*. For every $(t, x) \in A$, now let $R(t, x)$, $R_G(t, x)$ be the sets

$$
\begin{aligned}
R(t, x) &= [(z^1, \cdots, z^n) \mid z^i = f_i{}^*(t, x, p, v), i = 1, \cdots, \alpha, \\
&\quad z^i \geqq f_i{}^*(t, x, p, v), i = \alpha + 1, \cdots, n, p \in \Gamma, v \in U^\mu(t, x)] \subset E_n, \\
R_G(t, x) &= [(z^0, z^1, \cdots, z^n) \mid z^0 \geqq G^*(t, x, p, v), \\
&\quad z^i = f_i{}^*(t, x, p, v), i = 1, \cdots, \alpha, \quad z^i \geqq f_i{}^*(t, x, p, v), \\
&\quad i = \alpha + 1, \cdots, n, p \in \Gamma, v \in U^\mu(t, x)] \subset E_{n+1}.
\end{aligned} \tag{7}
$$

Thus, R is the projection of R_G on the z-space E_n. As it was pointed out in [1], the sets $R(t, x)$, $R_G(t, x)$ are the union of all convex combinations of μ

points of $Q(t, x) \subset E_n$ and $Q_G(t, x) \subset E_{n+1}$. Therefore, if we take $\mu = n + 1$ or $\mu = n + 2$ or larger, the sets $R(t, x)$, $R_G(t, x)$ are certainly convex, $R(t, x) = \operatorname{co} Q(t, x)$, $R_G(t, x) = \operatorname{co} Q_G(t, x)$.

Note that any set $\operatorname{co} R(\bar{t}, \bar{x}, \delta)$ is the union of all points $z \in E_n$ of the form $z = \sum_\gamma \lambda_\gamma z_\gamma$, where $\sum_\gamma$ ranges over all $\gamma = 1, \cdots, \nu$, say, $\nu = n + 1$ or higher, where $\lambda_\gamma \geqq 0$, $\gamma = 1, \cdots, \nu$, $\lambda_1 + \cdots + \lambda_\nu = 1$, and $z_\gamma \in R(t_\gamma, x_\gamma)$; that is, $z_\gamma{}^i = f_i{}^*(t_\gamma, x_\gamma, p_\gamma, v_\gamma)$, $i = 1, \cdots, \alpha$, $z_\gamma{}^i \geqq f_i{}^*(t_\gamma, x_\gamma, p_\gamma, v_\gamma)$, $i = \alpha + 1, \cdots, n$, $v_\gamma \in U^\mu(t_\gamma, x_\gamma)$, $(t_\gamma, x_\gamma) \in N_\delta(\bar{t}, \bar{x})$, $\gamma = 1, \cdots, \nu$. If p_γ and v_γ are denoted by $p_\gamma = (p_{\gamma 1}, \cdots, p_{\gamma\mu})$, $v_\gamma = (u_{\gamma 1}, \cdots, u_{\gamma\mu})$, then $\sum_j p_{\gamma j} = 1, \gamma = 1, \cdots, \nu$, and

$$z_\gamma{}^i = f_i{}^*(t_\gamma, x_\gamma, p_\gamma, v_\gamma) = \sum_j p_{\gamma j} f_i(t_\gamma, x_\gamma, u_{\gamma j}) = \sum_j p_{\gamma j} z_{\gamma j}^i, \qquad i = 1, \cdots, \alpha,$$

$$z_\gamma{}^i \geqq f_i{}^*(t_\gamma, x_\gamma, p_\gamma, v_\gamma) = \sum_j p_{\gamma j} f_i(t_\gamma, x_\gamma, u_{\gamma j}) = \sum_j p_{\gamma j} z_{\gamma j}^i, \qquad i = \alpha + 1, \cdots, n,$$

and then $\operatorname{co} R(t, x, \delta)$ is the union of all points $z = (z^1, \cdots, z^n)$ of the form

$$z^i = \sum_\gamma \sum_j \lambda_\gamma p_{\gamma j} z_{\gamma j}^i, \qquad i = 1, \cdots, \alpha,$$

$$z^i \geqq \sum_\gamma \sum_j \lambda_\gamma p_{\gamma j} z_{\gamma j}^i, \qquad i = \alpha + 1, \cdots, n,$$

$$z_{\gamma j} = (z_{\gamma j}^i, \cdots, z_{\gamma j}^n),$$

with $\lambda_\gamma p_{\gamma j} \geqq 0$, $\sum_\gamma \sum_j \lambda_\gamma p_{\gamma j} = 1$, $z_{\gamma j} = f(t_\gamma, x_\gamma, u_{\gamma j})$, $u_{\gamma j} \in U(t_\gamma, x_\gamma)$, $(t_\gamma, x_\gamma) \in N_\delta(\bar{t}, \bar{x})$, $\gamma = 1, \cdots, \nu$, $j = 1, \cdots, \mu$. In other words, $\operatorname{co} R(\bar{t}, \bar{x}, \delta)$ can be defined in terms of the original functions f instead of f^*, provided we take suitable convex combinations of $\mu\nu$ original points $z_{\gamma j}$. An analogous remark holds for the sets $R_G(t, x)$. These remarks show that statement (ii) above holds not only for the sets Q, Q_G, but also for the sets R, R_G. In other words, we have the following statement.

(iii) Let A be a given compact subset of the tx-space E_{n+1}, and for every $(t, x) \in A$ let $U(t, x)$ be a given closed subset of the u-space E_m. Let M be the set of all $(t, x, u) \in R_{1+n+m}$ with $(t, x) \in A$, $u \in U(t, x)$. Let $G(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ be given functions defined on M, let G, $f_{\alpha+1}, \cdots, f_n$ be nonnegative and lower semicontinuous on M, and let $f_1, \cdots, f_\alpha$ be continuous on M. Let $1, f_1, \cdots, f_\alpha$ be of slower growth than G (as $|u| \to \infty$) uniformly on A, and assume that $U(t, x)$ satisfies property (U) in A. If the set $R_G(t, x)$ is convex for every $(t, x) \in A$, then $R_G(t, x)$ satisfies property (Q) in A. If only $R(t, x)$ is known to be convex for every

$(t, x) \in A$, then $R(t, x)$ satisfies property (Q) with respect to $R_\sigma(t, x)$ in A.

Remark 5. The interest of statement (iii) lies in the fact that the growth property for f_i and G required in (ii) and (iii) does not imply an analogous property for f_i^* and G^*, as we shall show by an example in §5.

3. A closure theorem.

3.1. A sequence of continuous vector functions $x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, is said to be convergent in the ρ-metric toward a continuous vector function $x(t)$, $t_1 \leqq t \leqq t_2$, t_1, t_2 finite, provided $t_{1k} \to t_1$, $t_{2k} \to t_2$ as $k \to \infty$, and $x_k(t) \to x(t)$ as $k \to \infty$ uniformly in $(-\infty, +\infty)$ (where all x_k, x are extended in $(-\infty, +\infty)$ by continuity and constancy outside their interval of definition). We showed in [1, p. 371] that this mode of convergence corresponds to a suitable metrization of the space of all continuous vector functions $x(t)$, $a \leqq t \leqq b$, defined in arbitrary finite intervals of the real line.

Hypotheses. Let A be a closed subset of the tx-space $E_1 \times E_n$, for every $(t, x) \in A$ let $U(t, x)$ be a closed subset of the u-space E_m, let M be the set of all $(t, x, u) \in E_{1+n+m}$ with $(t, x) \in A$, $u \in U(t, x)$, and let $\tilde{f}(t, x, u) = (f_0, f) = (f_0, f_1, \cdots, f_n)$ be a given continuous vector function on M. Let $\tilde{Q}(t, x) = \tilde{f}(t, x, U(t, x)) \subset E_{n+1}$, $Q(t, x) = f(t, x, U(t, x)) \subset E_n$, and let α be a given integer, $\alpha \in \{0, \cdots, n\}$.

Let $x = (y, z)$, $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$, and let A be of the form $A = A_0 \times E_{n-\alpha}$, where A_0 is a closed subset of $E_{\alpha+1}$. Let us assume that $U(t, x), \tilde{f}(t, x, u)$ are independent of z, that is, that $U(t, x') = U(t, x'')$, $\tilde{f}(t, x', u) = \tilde{f}(t, x'', u)$ whenever $x' = (y, z')$, $x'' = (y, z'')$. Then $\tilde{Q}(t, x)$, $Q(t, x)$ also are independent of z. We shall often use the simple notations $U(t, y), \tilde{f}(t, y, u), \tilde{Q}(t, y), Q(t, y)$. Let us assume that $U(t, y)$ has property (U) in A_0.

Let $x_k(t) = (x_k^0, x_k) = (x_k^0, y_k, z_k)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, be a sequence of trajectories in E_{n+1} for which we assume that the α-vector $y_k(t) = (x_k^1, \cdots, x_k^\alpha)$ converges in the ρ-metric toward the absolutely continuous (AC) vector function $y(t)$, $t_1 \leqq t \leqq t_2$, $t_{1k} \to t_1$, $t_{2k} \to t_2$, and that the $(n - \alpha + 1)$-vector $(x_k^0(t), z_k(t)) = (x_k^0, x_k^{\alpha+1}, \cdots, x_k^n)$ converges pointwise for almost all $t \in (t_1, t_2)$ toward a vector $(x^0(t), z(t))$, $t_1 \leqq t \leqq t_2$, with $x_k^0(t_{1k}) \to x^0(t_1)$, $z_k(t_{1k}) \to z(t_1)$, $z_k(t_{2k}) \to z(t_2)$ as $k \to \infty$. Let us assume that $x^0(t)$, $z(t)$ admit of a decomposition $x^0 = X^0 + S^0$, $z = Z + S$, where $(X^0(t), Z(t))$ is an AC vector function in $[t_1, t_2]$, where $dS^0/dt = 0$, $dS/dt = 0$ a.e. in $[t_1, t_2]$, that is, (S^0, S) is singular and possibly discontinuous in $[t_1, t_2]$, and where $X^0(t_1) = x^0(t_1)$, $Z(t_1) = z(t_1)$, $S^0(t_1) = 0$, $S(t_1) = 0$.

Instead of $A = A_0 \times E_{n-\alpha}$, we may well assume A of the form $A = A_0 \times I$, where I is a finite interval of $E_{n-\alpha}$, $I = [a_{\alpha+1}, b_{\alpha+1}] \times \cdots$

$\times [a_n, b_n]$, and each $x_k^i(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, has range in $[a_i, b_i]$, $i = \alpha + 1, \cdots, n$.

CLOSURE THEOREM III. *Under the hypotheses above, if $\tilde{Q}(t, x)$ is convex for every $(t, x) \in A$ and has property (Q) in A, then $\tilde{X}(t) = (X^0(t), y(t), Z(t))$, $t_1 \leqq t \leqq t_2$, is a trajectory in E_{n+1}. If $Q(t, x)$ only is known to be convex for every $(t, x) \in A$ and has property (Q) with respect to $\tilde{Q}(t, x)$ in A, then $X(t) = (y(t), Z(t))$, $t_1 \leqq t \leqq t_2$, is a trajectory in E_n.*

Remark 6. In Closure Theorem III as usual we have said that $\tilde{x}(t) = (x^0, x)$, $t_1 \leqq t \leqq t_2$, is a *trajectory in* E_{n+1}, provided that $\tilde{x}$ is AC in $[t_1, t_2]$ and there is a measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, such that: (a) $(t, x(t)) \in A$ for every $t \in [t_1, t_2]$; (b) $u(t) \in U(t, x(t))$ a.e. in $[t_1, t_2]$; (c) $dx^0/dt = f_0(t, x(t), u(t))$, $dx/dt = f(t, x(t), u(t))$ a.e. in $[t_1, t_2]$. Analogously, $x(t)$, $t_1 \leqq t \leqq t_2$, is said to be a *trajectory in* E_n, provided x is AC in $[t_1, t_2]$, and there is a measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, satisfying (a), (b) above, and (c′) $dx/dt = f(t, x(t), u(t))$ a.e. in $[t_1, t_2]$. Thus, if $\tilde{x}(t) = (x^0, x)$ is a trajectory in E_{n+1}, then $x(t)$, $t_1 \leqq t \leqq t_2$, is a trajectory in E_n. In any case, we say that $u(t)$, $t_1 \leqq t \leqq t_2$, is a *control function*, or a *strategy*, and that u *generates* the trajectory x in E_n (or $\tilde{x}$ in E_{n+1}).

Remark 7. Closure Theorem III and its proof below are simple modifications of [1, pp. 386–389, Closure Theorem II]. (Actually the first part of Theorem III differs only in notations from Theorem II.) As we shall see in §4 (proof of Existence Theorem B), sequences verifying the convergence properties of Closure Theorem III are obtained by applying subsequently Vitali's extraction process on the components x^i, $i = 1, \cdots, \alpha$, and Helly's extraction process on the components x^i, $i = 0$ and $i = \alpha + 1, \cdots, n$. For the last process it is sufficient to know that the sequences $x_k^i(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, are equibounded with equibounded total variations; in particular, this is certainly the case if we know that, say, $f_i \geqq -M_i$, M_i constant, and that the sequences $[x_k^i(t_{1k})]$, $[x_k^i(t_{2k})]$ are bounded ($i = 0$ and $i = \alpha + 1, \cdots, n$).

Remark 8. Closure Theorem III (as well as Closure Theorem II with nonnegative f_0) can be expressed as a lower semicontinuity statement: if the functions f_i, $i = 0$ and $i = \alpha + 1, \cdots, n$, are nonnegative, then

$$X^0(t_2) \leqq \liminf x_k^0(t_{2k}), \tag{8}$$

$$X^i(t_2) \leqq \liminf x^i(t_{2k}), \qquad i = \alpha + 1, \cdots, n, \tag{9}$$

where actually lim inf can be replaced by lim as $k \to \infty$, since the extraction process has been performed already. Under the weaker alternate hypothesis of Closure Theorem III ($Q(t, x)$ convex with property (Q_r)), then only (9) holds. This will be clear from the proof below, and holds even if the hypothesis $f_i \geqq 0$ is replaced by $f_i \geqq -M_i$, M_i constant, $i = 0$ and $i = \alpha + 1, \cdots, n$. In the particular case in which $I[x, u] = \int_{t_1}^{t_2} f_0(t, x, u)\, dt$

(Lagrange problems) and $\alpha = n$, then $x_k \to x$ in the ρ-metric, and Closure Theorem III (under the hypothesis $\tilde{Q}$ convex with property (Q)) (or equivalently Closure Theorem II of [1]) can be expressed as follows: there exists a control function $u(t)$, $t_1 \leqq t \leqq t_2$, such that

$$I[x, u] \leqq \liminf I[x_k, u_k]. \tag{10}$$

Finally, it should be noted that there may be other control functions, say, u, still generating x, and for which (10) does not hold. For instance, for $\alpha = n = 1, f = u(1-u), f_0 = v, m = 2, (u, v) \in U = [v \geqq u, 0 \leqq u \leqq 1]$, then $Q = [z \mid z = u(1-u),\ 0 \leqq u \leqq 1] = [z \mid 0 \leqq z \leqq \frac{1}{4}]$, $\tilde{Q} = [(z^0, z) \mid z^0 \geqq u,\ z = u(1-u),\ 0 \leqq u \leqq 1] = [(z^0, z) \mid z^0 \geqq \frac{1}{2}(1 - (1-4z)^{1/2}),\ 0 \leqq z \leqq \frac{1}{4}]$. Both Q and $\tilde{Q}$ are independent of t, x, are convex, and trivially satisfy property (Q). If we take $u_k(t) = v_k(t) = 0, x_k(t) = 0$, $0 \leqq t \leqq 1, k = 1, 2, \cdots$, then $x(t) = 0, 0 \leqq t \leqq 1$, and x is generated by both $u(t) = v(t) = 0, 0 \leqq t \leqq 1$, and by $\bar{u}(t) = \bar{v}(t) = 1, 0 \leqq t \leqq 1$. Here $I[x, u, v] = I[x_k, u_k, v_k] = 0, k = 1, 2, \cdots, I[x, \bar{u}, \bar{v}] = 1$, and (10) holds for u but not for $\bar{u}$.

3.2. Proof of Closure Theorem III. Let $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, be a given strategy generating the trajectory $\tilde{x}_k = (x_k^0, x_k)$. Then the vector functions

$$\begin{aligned} \phi(t) &= \tilde{X}'(t) = (X^{0\prime}(t), X'(t)) = (X^{0\prime}(t), y'(t), Z'(t)), \\ &\qquad\qquad t_1 \leqq t \leqq t_2, \\ \phi_k(t) &= \tilde{x}_k'(t) = (x_k^{0\prime}(t), x_k'(t)) = (x_k^{0\prime}(t), y_k'(t), z_k'(t)), \\ &\qquad\qquad t_{1k} \leqq t \leqq t_{2k}, \quad k = 1, 2, \cdots, \end{aligned} \tag{11}$$

are defined almost everywhere, are L-integrable, and

$$\phi_k(t) = \tilde{f}(t, x_k(t), u_k(t)) \in \tilde{Q}(t, x_k(t))$$

for almost all $t \in [t_{1k}, t_{2k}]$, $k = 1, 2, \cdots$, and hence also $\phi_k(t) = \tilde{f}(t, y_k(t), u_k(t))$ a.e. in $[t_{1k}, t_{2k}]$, $k = 1, 2, \cdots$. We have to prove that $(t, y(t), Z(t)) \in A$ for every $t \in [t_1, t_2]$, and that there is a measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, such that

$$\tilde{X}'(t) = (X^{0\prime}(t), y'(t), Z'(t)) = \tilde{f}(t, y(t), u(t)), \qquad u(t) \in U(t, x(t)),$$

for almost all $t \in [t_1, t_2]$ (only $u(t) \in U(t, x(t))$ and $X'(t) = (y'(t), Z'(t)) = f(t, y(t), u(t))$ in the weaker alternate hypothesis of Closure Theorem III).

First $t_{1k} \to t_1$, $t_{2k} \to t_2$; hence if $t \in (t_1, t_2)$, or $t_1 < t < t_2$, then $t_{1k} < t < t_{2k}$ for all k sufficiently large, and $(t, y_k(t)) \in A_0$. Since $y_k(t) \to y(t)$ as $k \to \infty$ and A_0 is closed, we conclude that $(t, y(t)) \in A_0$ for every t with $t_1 < t < t_2$, and finally $(t, y(t)) \in A_0$ for every $t_1 \leqq t \leqq t_2$ since

$y(t)$ is continuous at $t = t_1$ and $t = t_2$. Finally, $(t, y(t), Z(t)) \in A_0 \times E_{n-s}$, or $(t, X(t)) \in A$, $t_1 \leqq t \leqq t_2$.

For almost all $t \in [t_1, t_2]$ the derivative $[x^{0\prime}(t), X'(t)] = (x^{0\prime}, y', Z')$ exists and is finite, $S^{0\prime}(t)$, $S'(t)$ exist, $S^{0\prime}(t) = 0$, $S'(t) = 0$, and $x_k^0(t) \to x^0(t)$, $z_k(t) \to z(t)$ as $k \to \infty$. Let t_0 be such a point with $t_1 < t_0 < t_2$. Then, there is a $\sigma > 0$ with $t_1 < t_0 - \sigma < t_0 + \sigma < t_2$, and, for some k_0 and all $k \geqq k_0$, also $t_{1k} < t_0 - \sigma < t_0 + \sigma < t_{2k}$. Let $X_0 = X(t_0) = (y_0, Z_0)$, $y_0 = y(t_0)$, $Z_0 = Z(t_0)$, $z_0 = z(t_0)$, $S_0 = S(t_0)$. We have $S'(t_0) = 0$, hence $z'(t_0)$ exists and $z'(t_0) = Z'(t_0)$. Also, we have $z_k(t_0) \to z(t_0) = z_0$, $x_k^0(t_0) \to x^0(t_0) = x_0^0$.

We have $y_k(t) \to y(t)$ uniformly in $[t_0 - \sigma, t_0 + \sigma]$ and all functions $y(t)$, $y_k(t)$ are continuous in the same interval. Thus, they are equicontinuous in $[t_0 - \sigma, t_0 + \sigma]$. Given $\epsilon > 0$ there is a $\delta > 0$ such that $t, t' \in [t_0 - \sigma, t_0 + \sigma]$, $|t - t'| \leqq \delta$, $k \geqq k_0$ implies $|y(t) - y(t')| \leqq \epsilon/2$, $|y_k(t) - y_k(t')| \leqq \epsilon/2$. We can assume $0 < \delta < \sigma$, $\delta \leqq \epsilon$. For any h, $0 < h \leqq \delta$, let us consider the averages

$$
\begin{aligned}
m_h &= h^{-1} \int_0^h \phi(t_0 + s)\, ds = h^{-1}[\tilde{X}(t_0 + h) - \tilde{X}(t_0)], \\
m_{hk} &= h^{-1} \int_0^h \phi_k(t_0 + s)\, ds = h^{-1}[\tilde{x}_k(t_0 + h) - \tilde{x}_k(t_0)],
\end{aligned} \tag{12}
$$

both of which are $(n + 1)$-vectors, $X = (X^0, y, Z)$, $x_k = (x_k^0, y_k, z_k)$.

Given $\eta > 0$ arbitrary, we can fix h, $0 < h \leqq \delta < \sigma$, so small that

$$
|m_h - \phi(t_0)| \leqq \eta,
$$

$$
|S^0(t_0 + h) - S^0(t_0)| < \eta h/8, \qquad |S(t_0 + h) - S(t_0)| < \eta h/8.
$$

This is possible since $h^{-1} \int_0^h \phi(t_0 + s)\, ds \to \phi(t_0)$ and $[S^0(t_0 + h) - S^0(t_0)]^{h^{-1}} \to 0$, $[S(t_0 + h) - S(t_0)]h^{-1} \to 0$ as $h \to 0+$. Also, we can choose h in such a way that $z_k(t_0 + h) \to z(t_0 + h)$, $x_k^0(t_0 + h) \to x^0(t_0 + h)$ as $k \to \infty$. This is possible since $z_k(t) \to z(t)$, $x_k^0(t) \to x^0(t)$ for almost all t, $t_1 < t < t_2$.

Having so fixed h, let us take $k_1 \geqq k_0$ so large that for $k \geqq k_1$ we have

$$
|y_k(t_0) - y(t_0)|, \quad |y_k(t_0 + h) - y(t_0 + h)| \leqq \min[\eta h/8, \epsilon/2],
$$

$$
|z_k(t_0) - z(t_0)|, \quad |z_k(t_0 + h) - z(t_0 + h)| \leqq \eta h/16,
$$

$$
|x_k^0(t_0) - x^0(t_0)|, \quad |x_k^0(t_0 + h) - x^0(t_0 + h)| \leqq \eta h/16.
$$

This is possible since $y_k(t) \to y(t)$, $z_k(t) \to z(t)$, $x_k^0(t) \to x^0(t)$ both at $t = t_0$ and $t = t_0 + h$ as $k \to \infty$. Then we have

$$
\begin{aligned}
\Delta_1 &= |h^{-1}[y_k(t_0 + h) - y_k(t_0)] - h^{-1}[y(t_0 + h) - y(t_0)]| \\
&\leqq |h^{-1}[y_k(t_0 + h) - y(t_0 + h)]| + |h^{-1}[y_k(t_0) - y(t_0)]| \\
&\leqq h^{-1}(\eta h/8) + h^{-1}(\eta h/8) = \eta/4.
\end{aligned}
$$

Analogously, since $z = Z + S$, we have

$$\begin{aligned}\Delta_2 &= |\,h^{-1}[z_k(t_0 + h) - z_k(t_0)] - h^{-1}[Z(t_0 + h) - Z(t_0)]| \\ &= |\,h^{-1}[z_k(t_0 + h) - z_k(t_0)] - h^{-1}[z(t_0 + h) - z(t_0)] \\ &\quad + h^{-1}[S(t_0 + h) - S(t_0)]| \\ &\leqq |\,h^{-1}[z_k(t_0 + h) - z(t_0 + h)]| + |\,h^{-1}[z_k(t_0) - z(t_0)]| \\ &\quad + |\,h^{-1}[S(t_0 + h) - S(t_0)]| \\ &\leqq h^{-1}(\eta h/16) + h^{-1}(\eta h/16) + h^{-1}(\eta h/8) = \eta/4.\end{aligned}$$

By the same argument and $x^0 = X^0 + S^0$, we have

$$\Delta_3 = |\,h^{-1}[x_k{}^0(t_0 + h) - x_k{}^0(t_0)] - h^{-1}[X^0(t_0 + h) - X^0(t_0)]| \leqq \eta/4.$$

Finally,

$$\begin{aligned}|\,m_{hk} - m_h\,| &= |\,h^{-1}[\tilde{x}_k(t_0 + h) - \tilde{x}_k(t_0)] - h^{-1}[\tilde{X}(t_0 + h) - \tilde{X}(t_0)]| \\ &\leqq \Delta_1 + \Delta_2 + \Delta_3 \leqq \eta/4 + \eta/4 + \eta/4 < \eta.\end{aligned}$$

We conclude that, for the chosen value of h, $0 < h \leqq \delta < \sigma$, and every $k \geqq k_1$, we have

$$(13)\quad |\,m_h - \phi(t_0)| \leqq \eta, \quad |\,m_{hk} - m_h\,| \leqq \eta, \quad |\,y_k(t_0) - y(t_0)| \leqq \epsilon/2.$$

For $0 \leqq s \leqq h$ we now have

$$\begin{aligned}|\,y_k(t_0 + s) - y(t_0)| &\leqq |\,y_k(t_0 + s) - y_k(t_0)| + |\,y_k(t_0) - y(t_0)| \\ &\leqq \epsilon/2 + \epsilon/2 = \epsilon, \\ |(t_0 + s) - t_0\,| &\leqq h \leqq \delta \leqq \epsilon, \\ \tilde{f}(t_0 + s, y_k(t_0 + s), u_k(t_0 + s)) &\in \tilde{Q}(t_0 + s, y_k(t_0 + s)).\end{aligned}$$

Hence, by definition of $\tilde{Q}(t_0\,, y_0\,, 2\epsilon)$, also

$$\phi_k(t_0 + s) = \tilde{f}(t_0 + s, y_k(t_0 + s), u_k(t_0 + s)) \in \tilde{Q}(t_0\,, y_0\,, 2\epsilon)$$

for almost all $s \in [0, h]$. The second integral relation (12) shows that we have also

$$m_{hk} = h^{-1} \int_0^h \phi_k(t_0 + s)\, ds \in \text{cl co}\, \tilde{Q}(t_0\,, y_0\,, 2\epsilon)$$

since the latter is a closed convex set. Finally, by relations (13) we deduce

$$|\,\phi(t_0) - m_{hk}\,| \leqq |\,\phi(t_0) - m_h\,| + |\,m_n - m_{hk}\,| \leqq 2\eta,$$

and hence,

$$\phi(t_0) \in [\text{cl co}\, \tilde{Q}(t_0\,, y_0\,, 2\epsilon)]_{2\eta} \subset E_{n+1}\,.$$

Here by the notation $[H]_{2\eta}$ we denote the set of all points of E_{n+1} at a distance $\leqq 2\eta$ from some point of the set H. Here $\eta > 0$ is an arbitrary number, and the set in brackets is closed. Hence,

$$\phi(t_0) \in \operatorname{cl}\operatorname{co} \tilde{Q}(t_0, y_0, 2\epsilon),$$

and this relation holds for every $\epsilon > 0$, and therefore,

$$(14) \quad \phi(t_0) = (X^{0\prime}(t_0), X'(t_0)) \in \bigcap_\epsilon \operatorname{cl}\operatorname{co} \tilde{Q}(t_0, y_0, 2\epsilon) \subset E_{n+1}.$$

Since $Q(t, x)$ has property (Q) we conclude that

$$\tilde{X}'(t_0) = (X^{0\prime}(t_0), X'(t_0)) \in \tilde{Q}(t_0, y_0) = \tilde{f}(t_0, y_0, U(t_0, y_0)),$$

where $\tilde{f}$ does not depend on z, and where $y_0 = y(t_0)$. Thus, there is some point $\bar{u} = \bar{u}(t_0) \in U(t_0, y(t_0))$ such that

$$\tilde{X}'(t_0) = \tilde{f}(t_0, y(t_0), \bar{u}).$$

This holds for almost all $t_0 \in [t_1, t_2]$.

We shall now make use of the implicit function theorem in the following form. Let S be a measure space, Y a Hausdorff space, and X a metrizable space which is the union of a countable number of compact metrizable subsets. Let $F: X \to Y$ be a continuous function, and $g: S \to Y$ a measurable function, such that $g(S) \subset F(X)$. Then there exists a measurable function $w: S \to X$ such that $F[w(s)] = g(s)$ for all $s \in S$ (see [10, Theorem 1]). Let S be the set of all $t \in [t_1, t_2]$ for which $\tilde{X}'(t) = \tilde{f}(t, y(t), u)$ for some $u \in U(t, y(t))$. As we have proved above S is measurable and meas $S = t_2 - t_1$. Let $X = M_0$ be the set of all (t, y, u) with $(t, y) \in A_0$, $u \in U(t, y)$. Then M_0 is a closed subset of $E_{1+\alpha+m}$ since $U(t, y)$ has property (U) in the closed set A_0 (Remark 3), and M_0 is the union of the countably many compact subsets $M_k = [(t, y, u) \in M_0 \mid |t| + |y| + |u| \leqq k]$, $k = 1, 2, \cdots$. Let $Y = E_{2+\alpha+n}$ and let $g: S \to Y$ be defined by $g: t \to (t, y(t), \tilde{X}'(t))$, and let $F: M_0 \to Y$ be defined by F: $(t, y, u) \to (t, y, \tilde{f}(t, y, u))$. Then g is measurable and F is continuous, and $g(S) \subset F(M_0)$. There exists, therefore, a measurable $w: S \to M_0$ such that $g = Fw$, or $w: t \to (t, y(t), u(t))$ with $\tilde{X}'(t) = \tilde{f}(t, y(t), u(t))$ for all $t \in S$, and $u(t)$ is measurable in S. By defining $u(t)$ arbitrarily in $[t_1, t_2] - S$, we conclude that there is a measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, with $\tilde{X}'(t) = \tilde{f}(t, y(t), u(t))$, $u(t) \in U(t, y(t))$ a.e. in $[t_1, t_2]$. The first part of Closure Theorem III is thereby proved. If we know only that $Q(t, x)$ is convex and has property (Q) with respect to $\tilde{Q}(t, x)$, then from (14) we conclude that

$$X'(t_0) \in Q(t_0, y_0) = f(t_0, y_0, U(t_0, y_0)).$$

By the same argument above we then conclude that there is a measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, with $X'(t) = f(t, y(t), u(t))$, $u(t) \in U(t, y(t))$ a.e. in $[t_1, t_2]$. The statement of Closure Theorem III is thereby proved.

If we assume $f_i \geqq 0$, $i = 0$ and $i = \alpha + 1, \cdots, n$, we conclude that all functions $x_k{}^i(t)$, $x^i(t)$ are monotone nondecreasing, and so are the functions $X^i(t)$, $S^i(t)$. Since $S^i(t_1) = 0$ we conclude that the same $S^i(t)$ are nonnegative, and

$$X^i(t_2) = x^i(t_2) - S^i(t_2) \leqq x^i(t_2) = \lim x_k{}^i(t_{2k}), \qquad i = 0 \quad \text{and} \quad i = \alpha + 1, \cdots, n. \tag{15}$$

If we assume only $f_1 \geqq -M_i$, M_i a constant, then $f_i + M_i \geqq 0$, and then all functions $x_k{}^i(t) + M_i(t - t_{1k})$, $x^i(t) + M_i(t - t_1)$, $X^0(t) + M_i(t - t_1)$, $S^i(t)$ are monotone nondecreasing; hence $S^i(t) \geqq 0$ as before, and relation (15) holds without changes. Also Remark 8 is thereby proved.

4. Existence theorems for usual solutions.

4.1. We consider Mayer-type optimization problems. We are concerned therefore with a differential system

$$\frac{dx}{dt} = f(t, x(t), u(t)),$$

with boundary conditions

$$(t_1, x(t_1), t_2, x(t_2)) \in B \subset E_{2n+2},$$

with a functional to be minimized:

$$I[x, u] = e(t_1, x(t_1), t_2, x(t_2)),$$

and constraints:

$$(t, x(t)) \in A, \qquad u(t) \in U(t, x(t)).$$

Let us define the terms more precisely.

Let $x = (x^1, \cdots, x^n) \in E_n$, $u = (u^1, \cdots, u^m) \in E_m$, and let x denote the space variable and u the control variable. We shall denote by A a given subset of the tx-space $E_1 \times E_n$, and for every $(t, x) \in A$ let $U(t, x)$ be a given subset of the u-space E_m. Let M denote the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$. Let $G(t, x, u)$ be a real-valued function and $f(t, x, u) = (f_1, \cdots, f_n)$ be a vector-valued function, both defined on M. Let B be a given subset of the $t_1x_1t_2x_2$-space E_{2n+2}, $x_1 = (x_1^1, \cdots, x_1^n)$, $x_2 = (x_2^1, \cdots, x_2^n)$, and let $e(t_1, x_1, t_2, x_2)$ be a given real-valued function defined on B.

We shall say that a pair of functions $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, is *admissible* provided: (a) $x(t)$ is absolutely continuous (AC) in $[t_1, t_2]$; (b) $u(t)$ is measurable in $[t_1, t_2]$; (c) $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$; (d) $u(t) \in U(t, x(t))$ a.e. in $[t_1, t_2]$; (e) $dx/dt = f(t, x(t), u(t))$ a.e. in $[t_1, t_2]$, (f) $(t_1, x(t_1), t_2, x(t_2)) \in B$. The point $(t_1, x(t_1), t_2, x(t_2))$ will be denoted

briefly by $\eta(x)$, and we seek the minimum of the functional $I[x, u] = e[\eta(x)]$ in the class of all admissible pairs x, u.

4.2.

EXISTENCE THEOREM A. *Let α, n, $0 \leqq \alpha \leqq n$, be given integers, and for $x = (x^1, \cdots, x^n)$ let y, z denote $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$, so that $x = (y, z)$. Let A_0 be a compact subset of the ty-space $E_{\alpha+1}$, let I be a finite closed interval of the z-space $E_{n-\alpha}$, $I = [a_{\alpha+1}, b_{\alpha+1}] \times \cdots \times [a_n, b_n]$, and thus $A = A_0 \times I$ is a compact subset of the tx-space E_{n+1}. For every $(t, y) \in A_0$ let $U(t, y)$ be a given closed subset of the u-space E_m satisfying property (U) in A_0. Let M_0 be the set of all (t, y, u) with $(t, y) \in A_0$, $u \in U(t, y)$, and $M = M_0 \times I$ is then the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, y)$, $x = (y, z)$. Let $f(t, y, u) = (f_1, \cdots, f_n)$, $H(t, y, u)$ be functions defined on M_0, and assume that $f_1, \cdots, f_\alpha$ are continuous on M_0, and that H, $f_{\alpha+1}, \cdots, f_n$ are nonnegative and lower semicontinuous on M_0. Let us assume that for every $i = 1, \cdots, \alpha$, the following growth condition holds:*

(γ_i) Given $\epsilon > 0$ there is a constant $M_{i\epsilon} \geqq 0$ such that $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon H(t, y, u)$ for all $(t, y, u) \in M_0$.

For every $(t, y) \in A_0$ let $Q_H(t, y)$ be the set of all $\tilde{z} = (z^0, z^1, \cdots, z^n) \in E_{n+1}$ defined by

$$Q_H(t, y) = [\tilde{z} \mid z^0 \geqq H(t, y, u),\ z^i = f_i(t, y, u),\ i = 1, \cdots, \alpha,\ z^i \geqq f_i(t, y, u),\ i = \alpha + 1, \cdots, n,\ u \in U(t, y)] \subset E_{n+1}, \tag{16}$$

and assume that $Q_H(t, y)$ is convex for every $(t, y) \in A_0$ and satisfies property (Q) in A_0. For every $(t, x, u) \in M$, $x = (y, z)$, we shall write $f(t, x, u) = f(t, y, u)$, $H(t, x, u) = H(t, y, u)$, $U(t, x) = U(t, y)$, $Q_H(t, x) = Q_H(t, y)$. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}, $x_1 = (x_1^1, \cdots, x_1^n)$, $x_2 = (x_2^1, \cdots, x_2^n)$, and assume that B is independent of $x_2^{\alpha+1}, \cdots, x_2^n$; hence B is of the form $B = B_0 \times E_{n-\alpha}$, where B_0 is a closed subset of $E_{n+2+\alpha}$. Let $e(t_1, x_1, t_2, x_2)$ be a real-valued continuous function defined on B, which is monotone nondecreasing with respect to each variable $x_2^{\alpha+1}, \cdots, x_2^n$. Let Ω be the class of all admissible pairs x, u (defined as in §4.1 in relation to the sets A, $U(t, x)$, M, B and functions f and e), for which $H(t, y(t), u(t))$ is L-integrable in $[t_1, t_2]$ and

$$\int_{t_1}^{t_2} H(t, y(t), u(t))\, dt \leqq M_1 \tag{17}$$

for some constant $M_1 \geqq 0$, and assume that Ω is not empty.

Then the functional $I[x, u] = e[\eta(x)]$ has an absolute minimum in Ω (and the optimal pair satisfies (17)).

The hypothesis $H \geqq 0, f_i \geqq 0$, $i = \alpha + 1, \cdots, n$, can be replaced by the weaker requirement $H \geqq -K_0, f_i \geqq -K_i$, $i = \alpha + 1, \cdots, n$, for some

constants K_0, K_i, provided we replace (γ_i) by the analogous requirement $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon[H(t, y, u) + K_0]$.

For any of the indices $i = 1, \cdots, \alpha$, condition (γ_i) can be disregarded, provided we impose a further restriction on Ω, namely, that Ω is made up of only those pairs x, u for which (17) holds, and also

$$\int_{t_1}^{t_2} \left| \frac{dx^i}{dt} \right|^{p_i} dt \leqq M_i \tag{18}$$

for some constants $M_i \geqq 0$ and $p_i > 1$, and at least all $i = 1, \cdots, \alpha$ for which (γ_i) does not hold.

The condition that $Q_H(t, y)$ (if convex) satisfies property (Q) is certainly satisfied if we know that $1, f_1, \cdots, f_\alpha$ are of slower growth than H uniformly in A_0, and then conditions (γ_i), $i = 1, \cdots, \alpha$, are all satisfied.

Finally, H may be one of the functions $f_{\alpha+1}, \cdots, f_n$, say $H = f_n$, and then $0 \leqq \alpha \leqq n - 1$, and a relation (17) is satisfied by all admissible pairs x, u with $M_1 = b_n - a_n$. In this situation, if we disregard (17), then instead of Q_H we may consider the simpler sets $Q(t, y)$ of all $z = (z^1, \cdots, z^n) \in E_n$ defined by

$$\begin{aligned} Q(t, y) = [z \mid z^i = f_i(t, y, u), i = 1, \cdots, \alpha, z^i \geqq f_i(t, y, u), \\ i = \alpha + 1, \cdots, n, u \in U(t, y)] \subset E_n . \end{aligned} \tag{19}$$

Then we shall replace the requirements concerning $Q_H(t, y)$ by analogous requirements concerning $Q(t, y)$. In other words, we shall require that $Q(t, y)$ is convex for every $(t, y) \in A_0$ and that $Q(t, y)$ satisfies property (Q) in A_0.

Remark 9. In Existence Theorem A, if the set A is of the form $A = A_0' \times E_{n-\alpha}$, where A_0' is only a closed subset of the y-space E_α, then conditions should be added to guarantee that, for any minimizing sequence of admissible pairs x, u satisfying (17), the trajectories are contained in some compact set $A_0 \times I$ as required in Theorem A. (For examples of such conditions, see [1] and [7].) Having stated above Existence Theorem I and a number of its variants, we proceed now with a proof of the theorem.

Proof of Existence Theorem A. Let us denote by i the infimum of the functional $I[x, u] = e[\eta(x)]$ in Ω. Since $\eta(x) \in B \cap (A \times A)$, A compact and B closed, then $B \cap (A \times A)$ is compact, and the continuous function e is bounded there. Thus, i is finite. Let $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, be a minimizing sequence for $I[x, u]$ in Ω. Then $(t, x_k(t)) \in A$ for $t \in [t_{1k}, t_{2k}]$, and if we write $x_k(t) = (y_k, z_k)$, then $(t, y_k(t)) \in A_0$, $z_k(t) \in I$ for $t \in [t_{1k}, t_{2k}]$. Also, $u_k(t) \in U(t, y_k(t))$ a.e. in $[t_{1k}, t_{2k}]$, $\int_{t_{1k}}^{t_{2k}} H(t, x_k(t), u_k(t))\, dt \leqq M_1$, $k = 1, 2, \cdots$, and $e[\eta(x_k)] \to i$ as $k \to \infty$. Since A is compact, the sequences $[t_{1k}]$, $[t_{2k}]$ are bounded, and we have

$\bar{a} \leqq t_{1k} \leqq t_{2k} \leqq \bar{b}$ for all k and some constants $\bar{a}$ and $\bar{b}$. Also, $a_i \leqq x_k^i(t) \leqq b_i$ for all $t \in [t_{1k}, t_{2k}]$, $k = 1, 2, \cdots$, $i = \alpha + 1, \cdots, n$.

Let x^0 denote a new variable satisfying $dx^0/dt = H(t, x, u)$ and $x^0(t_1) = 0$. Let $x_k^0(t) = \int_{t_{1k}}^{t} H(\tau, x_k(\tau), u_k(\tau))\, d\tau$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, and then $a_0 = 0 \leqq x_k^0(t) \leqq M_1 = b_0$ for the same t and k. We denote by A' the compact set $A' = A \times [a_0, b_0] \subset E_{n+1}$.

We shall prove that the functions $x_k^i(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, $i = 1, 2, \cdots, \alpha$, are equiabsolutely continuous. Namely, we shall prove that their derivatives dx_k^i/dt are equiabsolutely integrable in $[t_{1k}, t_{2k}]$. Given $\epsilon > 0$, let $\sigma = \epsilon/2M_1$ and let E denote any measurable subset of $[t_{1k}, t_{2k}]$. Then $|f_i(t, y, u)| \leqq M_{i\sigma} + \sigma H(t, y, u)$, $i = 1, \cdots, \alpha$, for all $(y, t, u) \in M_0$ and hence,

$$\begin{aligned}\int_E |dx_k^i/dt|\, dt &= \int_E |f_i(t, y_k(t), u_k(t))|\, dt \\ &\leqq \int_E [M_{i\sigma} + \sigma Ht, y_k(t), u_k(t)]\, dt \\ &\leqq M_{i\sigma} \text{ meas } E + \sigma M_1 \\ &= M_{i\sigma} \text{ meas } E + \epsilon/2.\end{aligned}$$

Thus, for meas $E \leqq \delta = \epsilon/2M_{i\sigma}$, we have $\int_E |dx_k^i/dt|\, dt \leqq \epsilon/2 + \epsilon/2 = \epsilon$, $i = 1, \cdots, \alpha$, and this proves our assertion. Thus, the functions $x_k^i(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, $i = 1, \cdots, \alpha$, are equibounded and equiabsolutely continuous, and the sequences $[t_{1k}]$, $[t_{2k}]$ are bounded.

First, we extract a subsequence, say still $[k]$, such that $t_{1k} \to t_1$, $t_{2k} \to t_2$, and hence $\bar{a} \leqq t_1 \leqq t_2 \leqq \bar{b}$. Then we consider the same functions $x_k^i(t)$, $i = 1, \cdots, \alpha$, as defined in the whole interval $[\bar{a}, \bar{b}]$ by continuity and constancy outside $[t_{1k}, t_{2k}]$. By Ascoli's theorem there is a further subsequence, say still $[k]$, such that $x_k^i(t) \to x^i(t)$ as $k \to \infty$ uniformly in $[\bar{a}, \bar{b}]$, and the limit functions $x^i(t)$ are not only continuous but also AC in $[\bar{a}, \bar{b}]$ and constant outside $[t_1, t_2]$. We shall restrict $x^i(t)$ to the interval $[t_1, t_2]$. Because of the uniform continuity we have $x_k^i(t_{1k}) = x_{k1}^i \to x_1^i = x^i(t_1)$, $x_k^i(t_{2k}) = x_{k2}^i \to x_2^i = x^i(t_2)$ as $k \to \infty$, $i = 1, \cdots, \alpha$. Since $y = (x^1, \cdots, x^\alpha)$, we can write $y_k(t_{1k}) \to y(t_1) = (x_1^1, \cdots, x_1^\alpha)$, $y_k(t_{2k}) \to y(t_2) = (x_2^1, \cdots, x_2^\alpha)$ as $k \to \infty$.

We shall now consider the $n - \alpha + 1$ sequences of scalar functions

$$x_k^0(t) = \int_{t_{1k}}^{t} H(\tau, y_k(\tau), u_k(\tau))\, d\tau,$$

$$x_k^i(t) = x_{k1}^i + \int_{t_{1k}}^{t} f_i(\tau, y_k(\tau), u_k(\tau))\, d\tau, \qquad i = \alpha + 1, \cdots, n,$$

$$t_{1k} \leqq t \leqq t_{2k}, \quad k = 1, 2, \cdots,$$

where $H \geqq 0, f_i \geqq 0$, and hence these functions are monotone nondecreasing and also equibounded as mentioned before, We shall extend these functions in the fixed interval $[\bar{a}, \bar{b}]$ with $\bar{a} \leqq t_{1k} \leqq t_{2k} \leqq \bar{b}$, $t_{1k} \to t_1$, $t_{2k} \to t_2$, by continuity and constancy outside $[t_{1k}, t_{2k}]$. By Helly's theorem in $[\bar{a}, \bar{b}]$ we shall now extract a further subsequence, say still $[k]$, such that $x_k^i(t) \to x^i(t)$ as $k \to \infty$ for every $t \in [\bar{a}, \bar{b}]$, $i = 0$ and $i = \alpha + 1, \cdots, n$. We may well extract the subsequence in such a way that the limits also exist: $x_1^i = \lim x_k^i(t_{1k})$, $x_2^i = \lim x_k^i(t_{2k})$ as $k \to \infty$, $i = 0$ and $i = \alpha + 1, \cdots, n$, and then $x_1^0 = 0$.

Note that the functions $x^i(t)$, $\bar{a} \leqq t \leqq \bar{b}$, $i = 0$ and $i = \alpha + 1, \cdots, n$, are monotone nondecreasing in $[\bar{a}, \bar{b}]$, possibly discontinuous, constant in $[a, t_1)$ and in $(t_2, b]$. Let us prove that $x_1^i \leqq x^i(t_1 + 0)$, $x^i(t_2 - 0) \leqq x_2^i$. Indeed, if we take any $\bar{t}$ with $t_1 < \bar{t} < t_2$ and any $\epsilon > 0$, there is some $\bar{k}$ such that $x_k^i(\bar{t}) \leqq x^i(\bar{t}) + \epsilon$, $t_{1k} < \bar{t}$ for every $k \geqq \bar{k}$, and hence $x_k^i(t_{1k}) \leqq x_k^i(\bar{t}) \leqq x^i(\bar{t}) + \epsilon$, and finally, as $k \to \infty$, also $x_1^i \leqq x^i(\bar{t}) + \epsilon$. Here ϵ is arbitrary, hence $x_1^i \leqq x^i(\bar{t})$, and as $\bar{t} \to t_1 + 0$, we obtain $x_1^i \leqq x^i(t_1 + 0)$. Analogously, we can prove that $x^i(t_2 - 0) \leqq x_2^i$.

We shall restrict $x^i(t)$ to the interval $[t_1, t_2]$ and modify $x^i(t)$ at t_1 and t_2, if needed, by taking $x^i(t_1) = x_1^i = \lim x_k^i(t_{1k})$, $x^i(t_2) = x_2^i = \lim x_k^i(t_{2k})$, $i = 0$ and $i = \alpha + 1, \cdots, n$, so that, in particular, $x^0(t_1) = 0$. By doing this the functions $x^i(t)$, $t_1 \leqq t \leqq t_2$, are still monotone nondecreasing in $[t_1, t_2]$, and we still have $\lim x_k^i(t) = x^i(t)$ for all $t \in (t_1, t_2)$, $i = 0$ and $i = \alpha + 1, \cdots, n$. Since $z = (x^{\alpha+1}, \cdots, x^n)$, we have thus $z_k(t_{1k}) \to z(t_1) = (x_1^{\alpha+1}, \cdots, x_1^n)$, $z_k(t_{2k}) \to z(t_2) = (x_2^{\alpha+1}, \cdots, x_2^n)$ as well as $x_k^0(t_{1k}) \to x^0(t_1) = 0$, $x^0(t_{2k}) \to x^0(t_2) = x_2^0 \leqq M_1$. We have obviously $a_i \leqq x^i(t) \leqq b_i$ for all $t_1 \leqq t \leqq t_2$ and $i = 0$ and $i = \alpha + 1, \cdots, n$.

Note that $(t, y_k(t)) \in A_0$ for $t_{1k} \leqq t \leqq t_{2k}$ implies $(t, y(t)) \in A_0$ for every $t \in (t_1, t_2)$, and by continuity of y we conclude that $(t, y(t)) \in A_0$ for every $t_1 \leqq t \leqq t_2$. Thus, $(t, y(t), z(t)) \in A = A_0 \times I$, and $(t, x^0(t), y(t), z(t)) \in A' = A \times [a_0, b_0]$ for all $t_1 \leqq t \leqq t_2$.

We shall now decompose $x^i(t)$, $t_1 \leqq t \leqq t_2$, $i = 0$ and $i = \alpha + 1, \cdots, n$, into an AC part $X^i(t)$ and a singular part $S^i(t)$, $x^i(t) = X^i(t) + S^i(t)$, with $X^i(t_1) = x^i(t_1) = x_1^i$, $S^i(t_1) = 0$, X^i, S^i both nondecreasing, hence $S^i(t) \geqq 0$, and $x_1^i = X_1^i = X^i(t_1) \leqq X^i(t_2) = X_2^i \leqq x_2^i(t_2)$, $i = 0$ and $i = \alpha + 1, \cdots, n$.

Let $\tilde{u} = (v^0, u^1, \cdots, u^m, v^{\alpha+1}, \cdots, v^n) = (v^0, u, v)$ denote an auxiliary control μ-vector, $\mu = m + n - \alpha + 1$, let $\tilde{U}(t, y)$ be the set of all $\tilde{u} \in E_\mu$ with $u = (u^1, \cdots, u^m) \in U(t, y)$, $v^0 \geqq H(t, y, u)$, $v^i \geqq f_i(t, y, u)$, $i = \alpha + 1, \cdots, n$. Let $\tilde{\tilde{f}}(t, y, u), \tilde{f}(t, y, u)$ denote the $(n + 1)$- and n-vector functions

$$\tilde{\tilde{f}}(t, y, u) = (\tilde{f}_0, f_1, \cdots, f_\alpha, \tilde{f}_{\alpha+1}, \cdots, \tilde{f}_n),$$

$$\tilde{f}(t, y, u) = (f_1, \cdots, f_\alpha, \tilde{f}_{\alpha+1}, \cdots, \tilde{f}_n),$$

with $\tilde{f}_0 = v^0, \tilde{f}_i = v^i, i = \alpha + 1, \cdots, n$. Finally, we consider the auxiliary differential equations

$$d\tilde{x}/dt = \tilde{\tilde{f}}(t, y, u), \qquad x = (x^0, x^1, \cdots, x^n) = (x^0, x) = (x^0, y, z)$$

or

$$dx^0/dt = v^0, \qquad dx^i/dt = f_i(t, y, u), \qquad i = 1, \cdots, \alpha,$$
$$dx^i/dt = v^i, \qquad i = \alpha + 1, \cdots, n,$$

with constraints $\tilde{u}(t) \in \tilde{U}(t, y(t))$ or

$$v^0(t) \geqq H(t, y(t), u(t)), \qquad u(t) \in U(t, y(t)),$$
$$v^i(t) \geqq f_i(t, y(t), u(t)), \qquad i = \alpha + 1, \cdots, n.$$

Note that $\tilde{\tilde{f}}(t, y, \tilde{U}(t, y)) = Q_H(t, y), \tilde{f}(t, y, \tilde{U}(t, y)) = Q(t, y)$, where Q_H, Q are the sets (16) and (19), and Q_H satisfies property (Q) in A_0. Let us prove that $\tilde{U}(t, y)$ satisfies property (U) in A_0. By Remark 3 it is enough to prove that the set $\tilde{M}_0$ of all $(t, y, v^0, u, v) \in E_{2+n+m}$ satisfying $(t, y) \in A_0, u \in U(t, y), v^0 \geqq H, v^i \geqq f_i, i = \alpha + 1, \cdots, n$, is closed. Indeed, having already denoted by M_0 the set of all (t, y, u) with $(t, y) \in A_0, u \in U(t, y)$, we see that $\tilde{M}_0$ is the cylindrical set of all (t, y, v^0, u, v) with $(t, y, u) \in M_0$ and $v^0 \geqq H(t, y, u), v^i \geqq f_i(t, y, u), i = \alpha + 1, \cdots, n$. Here M_0 is closed because A_0 is compact and $U(t, y)$ has property (U) on A_0. On the other hand, for every integer k, the part $\tilde{M}_{0k}$ of $\tilde{M}_0$ contained in the slab $|u| \leqq k, |v^0| \leqq k, |v^i| \leqq k, i = \alpha + 1, \cdots, n$, of E_{2+n+m} is the set of all (t, y, u, v^0, v) with (t, y, u) in the compact set $[(t, y, u) \in M_0, |u| \leqq k]$, and $v^0 \geqq H(t, y, u) \geqq 0, v^i \geqq f_i(t, y, u) \geqq 0$, $i = \alpha + 1, \cdots, n$, and this set $\tilde{M}_{0k}$ is compact since H, f_i are nonnegative and lower semicontinuous. Thus $\tilde{M}_0$ is closed, and $\tilde{U}(t, y)$ has property (U) on A_0. Thus, by Closure Theorem III (first part, with $A', \tilde{x}, \tilde{U}, \tilde{\tilde{f}}$ replacing $A, x, U, \tilde{f}$), there is a measurable function $\tilde{u}(t) = (v^0, u^1, \cdots, u^m, v^{\alpha+1}, \cdots, v^n) = (v^0, u, v), t_1 \leqq t \leqq t_2$, with $\tilde{u}(t) \in \tilde{U}(t, y(t))$, hence $u(t) \in U(t, y(t))$, such that

$$\frac{dX^0}{dt} = v^0 \geqq H(t, y(t), u(t)) \geqq 0,$$

$$\frac{dx^i}{dt} = f_i(t, y(t), u(t)), \qquad i = 1, \cdots, \alpha,$$

$$\frac{dX^i}{dt} = v^i \geqq f_i(t, y(t), u(t)) \geqq 0, \qquad i = \alpha + 1, \cdots, n,$$

a.e. in $[t_1, t_2]$.

We shall take

$$Z^0(t) = \int_{t_1}^{t} H(\tau, y(\tau), u(\tau))\, d\tau,$$

$$Z^i(t) = x_1{}^i + \int_{t_1}^{t} f_i(\tau, y(\tau), u(\tau))\, d\tau, \quad i = \alpha + 1, \cdots, n,$$

where $t_1 \leqq t \leqq t_2$, $x_1{}^i = x^i(t_1)$, and thus $x_1{}^i = Z^i(t_1) \leqq Z_2{}^i = Z_i(t_2) \leqq X^i(t_2) \leqq x_2{}^i$, $x_1{}^0 = 0$, $i = 0$ and $i = \alpha + 1, \cdots, n$. Let us consider now the pair $\tilde{X} = (Z^0, X), u$, or

$$\begin{aligned} \tilde{X}(t) &= (Z^0(t), y(t), Z(t)) \\ &= (Z^0, x^1, \cdots, x^\alpha, Z^{\alpha+1}, \cdots, Z^n), \quad u(t), \quad t_1 \leqq t \leqq t_2. \end{aligned} \tag{20}$$

This pair obviously satisfies properties (a) and (b) of §4.1 with A', $\tilde{x}$ replacing A, x. Also, it satisfies property (c) since $A' = A_0 \times I \times [a_0, b_0]$, and $[x_1{}^i, Z_2{}^i] = [x_2{}^i, Z^i(t_2)] \subset [x_1{}^i, x_2{}^i] \subset [a_i, b_i]$, $i = 0$ and $i = \alpha + 1, \cdots, n$. In particular, for $i = 0$,

$$0 \leqq \int_{t_1}^{t_2} H(t, y(t), u(t))\, dt = Z^0(t_2) \leqq X^0(t_2) \leqq x^0(t_2) \leqq b_0 = M_1.$$

Obviously the pair (20) also satisfies properties (d) and (e) of §4.1. Finally, B is closed and B is independent of $x_2{}^{\alpha+1}, \cdots, x_2{}^n$; hence,

$$\eta(x_k) = (t_{1k}, y_k(t_{1k}), z_k(t_{1k}), t_{2k}, y_k(t_{2k}), z_k(t_{2k})) \in B$$

yields, as $k \to \infty$,

$$(t_1, y(t_1), Z(t_1), t_2, y(t_2), Z(t_2)) \in B,$$

and also

$$\eta(X) = (t_1, y(t_1), Z(t_1), t_2, y(t_2), Z(t_2)) \in B.$$

We conclude that the pair (20) satisfies properties (a)–(f) of §4.1 and, therefore, is an admissible pair. Hence X, u belongs to Ω, and $e[\eta(X)] \geqq i$. On the other hand, e is monotone nondecreasing in $x_2{}^{\alpha+1}, \cdots, x_2{}^n$, and hence

$$\begin{aligned} e[\eta(X)] &= e(t_1, x_1{}^1, \cdots, x_1{}^n, t_2, x_2{}^1, \cdots, x_2{}^\alpha, Z_2{}^{\alpha+1}, \cdots, Z_2{}^n) \\ &\leqq e(t_1, x_1{}^1, \cdots, x_1{}^n, t_2, x_2{}^1, \cdots, x_2{}^\alpha, x_2{}^{\alpha+1}, \cdots, x_2{}^n) \\ &= e(t_1, x(t_1), t_2, x(t_2)) = \lim e(\eta(x_k)) = i, \end{aligned}$$

or $e[\eta(X)] \leqq i$. Thus, $e[\eta(X)] = i$, and the main part of Existence Theorem A is thereby proved.

If for any of the indices $i = 1, \cdots, \alpha$ the corresponding relation (γ_i) does not hold, but Ω is restricted to only those admissible pairs which satisfy (17) and corresponding relations (18), then the equiabsolute integrability of the functions $dx_k{}^i/dt$ can be proved by the classical argu-

ment based on the Hölder inequality. The remaining part of the proof of Theorem A is unchanged.

If $1, f_1, \cdots, f_\alpha$ are of slower growth than H, then all conditions (γ_i), $i = 1, \cdots, \alpha$, are necessarily satisfied as we shall prove in §4.3 below, and the set $Q_H(t, y)$, which is assumed to be convex, necessarily satisfies property (Q) in A_0 as proved in §2.2 (ii). Then $Q_H(t, x)$ trivially satisfies property (Q) in A.

Let us assume $H = f_n$. Then for any admissible pair x, u we have

$$0 \leqq \int_{t_1}^{t_2} H(t, y(t), u(t))\, dt = x^n(t_2) - x^n(t_1) \leqq b_n - a_n ;$$

that is, a relation (17) is necessarily satisfied with $M_1 = b_n - a_n$. If, in the present situation, we disregard (17), then we can repeat the proof of Theorem A disregarding the auxiliary variable x^0. We shall apply again Closure Theorem III (first part, with $n + 1$ replaced by n) and we prove that $X = (y, Z)$ is a trajectory in E_n.

Under the weaker hypothesis $H \geqq -K_0, f_i \geqq -K_i, i = \alpha + 1, \cdots, n$, the proof of Theorem A remains the same since we already noticed that Closure Theorem III still holds under this weaker hypothesis. Note that now $x^i(t) + K_i(t - t_1)$ is monotone nondecreasing in $[t_1, t_2]$; hence the same holds for the AC part, say, $X^i(t) + K_i(t - t_1)$, and for the singular part, say, $S^i(t)$; hence, $S^i(t) \geqq 0$ as before, $i = 0$ and $i = \alpha + 1, \cdots, n$, and the argument is the same.

Existence Theorem A is thereby proved together with the variants stated at the end of the theorem.

Remark 10. Condition (γ_i) for $H \geqq 0$ stated in Theorem A, as well as the modified form for $H \geqq -K_0$ stated at the end of the same Theorem A, can well be replaced by the following more general hypothesis:

(γ_i') Given $\epsilon > 0$ there is a nonnegative locally L-integrable function $M_{i\epsilon}(t)$, t real, such that $|f_i(t, y, u)| \leqq M_{i\epsilon}(t) + \epsilon H(t, y, u)$ for all $(t, y, u) \in M_0$.

Obviously (γ_i') for $\epsilon = 1$ implies that $H(t, y, u) \geqq -M_{i\epsilon}(t)$ for all $(t, y, u) \in M_0$. The proofs are the same. This remark applies also to the remaining theorems of the present paper.

4.3. Here we denote by A any compact subset of the tx-space $E_1 \times E_n$, and for every $(t, x) \in A$ we denote by $U(t, x)$ any closed subset of the u-space E_m. Let M be the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$, and $G(t, x, u) \geqq 0$ and let $f_1(t, x, u), \cdots, f_\alpha(t, x, u)$ be given functions defined on M.

The condition "(α) $1, f_1, \cdots, f_\alpha$ are of slower growth than $G \geqq 0$ on A" should be compared with the more general condition "(β) $G \geqq 0$

and given $\epsilon > 0$, there is some constant M_ϵ such that $|f_i(t, x, u)| \leqq M_\epsilon + \epsilon G(t, x, u)$ for all $(t, x, u) \in M$, $i = 1, \cdots, \alpha$."

Let us prove that (α) implies (β). Indeed, given $\epsilon > 0$ there is some $\bar{u} = \bar{u}(\epsilon) \geqq 0$ such that $(t, x) \in A$, $u \in U(t, x)$, $|u| \geqq \bar{u}(\epsilon)$ implies $|f_i(t, x, u)| \leqq \epsilon G(t, x, u)$, while $(t, x) \in A$, $u \in U(t, x)$, $|u| \leqq \bar{u}(\epsilon)$, together with the continuity of f_i on the compact set $S = [(t, x, u) \mid (t, x) \in A, u \in U(t, x), |u| \leqq \bar{u}(\epsilon)]$ implies $|f_i(t, x, u)| \leqq M_{i\epsilon}$ for some constant $M_{i\epsilon}$. If M_ϵ denotes the largest of the constants $M_{i\epsilon}$, $i = 1, 2, \cdots, \alpha$, then $|f_i(t, x, u)| \leqq M_\epsilon + \epsilon G(t, x, u)$ for all $(t, x, u) \in H$. The compactness of the set M is a consequence of the property (U) for $U(t, x)$ (see Remark 3) and the compactness of A. We have proved that if A is compact, $U(t, x)$ satisfies property (U), and $f_1, \cdots, f_\alpha$ are continuous, then (α) implies (β).

Let us prove that there are functions f_i, G satisfying (β) but not (α). Indeed, take

$$f_i(t, x, u) = t^2 u + 1, \qquad G(t, x, u) = tu^2$$

for $(t, x) \in A = [0 \leqq t \leqq 1, 0 \leqq x \leqq 1]$, $u \in U = (-\infty, +\infty)$, $M = A \times U$. Neither 1 nor f_i is of slower growth than G uniformly in A. On the other hand, given $\epsilon > 0$, let $M_\epsilon = 1 + \epsilon^{-1}$. For $|u| \geqq \epsilon^{-1}$ we have $|f_i| \leqq t^2|u| + 1 \leqq |u|^{-1} tu^2 + 1 \leqq M_\epsilon + \epsilon G$; for $|u| \leqq \epsilon^{-1}$ we have $|f_i| = t^2|u| + 1 \leqq \epsilon^{-1} + 1 = M_\epsilon \leqq M_\epsilon + \epsilon G$. Thus, the functions f_i, G, defined above satisfy (β) but not (α).

4.4. With the notations of §4.3 we consider here the sets

$$Q(t, x) = [(z^1, \cdots, z^n) \mid z^i = f_i(t, x, u),\ i = 1, \cdots, \alpha,$$
$$z^i \geqq f_i(t, x, u),\ i = \alpha + 1, \cdots, n,\ u \in U(t, x)] \subset E_n,$$

$$Q_G(t, x) = [(z^0, z^1, \cdots, z^n) \mid z^0 \geqq G(t, x, u), z^i = f_i(t, x, u),\ i = 1, \cdots, \alpha,$$
$$z^i \geqq f_i(t, x, u),\ i = \alpha + 1, \cdots, n,\ u \in U(t, x)] \subset E_{n+1}.$$

While condition (α) is strong enough to guarantee that the sets $\tilde{Q}(t, x)$, if convex, have property (Q), and that the sets $Q(t, x)$, if convex, have property (Q) with respect to $Q_G(t, x)$ as proved in §2 (i), instead condition (β) does not imply these properties. This can be seen by the following example. Let $n = 1$, $\alpha = 1$, $G(t, x, u) = f_1(t, x, u) = (1 + u^2)^{-1}$, $u \in U = E_1$. Property (β) holds trivially with $M_\epsilon = 1$, since $0 \leqq f_1 \leqq 1 = M_\epsilon \leqq M_\epsilon + \epsilon G$. Here the sets $Q = Q(t, x)$, $Q_G = Q_G(t, x)$ do not depend on (t, x), and hence $Q_G(\delta) = Q_G(t, x, \delta) = Q_G$ for all (t, x) and $\delta > 0$, and the same holds for Q. Here Q is the half-open segment $[z^1 \mid 0 < z^1 \leqq 1]$, Q_G is the set $[(z^0, z^1) \mid 0 < z^1 \leqq 1, z^0 \geqq z^1]$, and cl co $Q_G(\delta) =$ cl co Q_G is the set $[(z^0, z^1) \mid 0 \leqq z^1 \leqq 1, z^0 \geqq z^1]$. Thus, the point $(z^0 = 0, z^1 = 0)$

$\in \bigcap_\delta \operatorname{cl}\operatorname{co} Q_G(\delta)$ and $(z^1 = 1) \in \bigcap_\delta \operatorname{cl}\operatorname{co} Q(\delta)$, but neither $(z^0 = 0, z^1 = 0)$ belongs to Q_G, nor $z^1 = 1$ belongs to Q. Thus, neither Q_G has property (Q), nor Q has property (Q), nor Q has property (Q_r). Analogously, for $n = 1$, $\alpha = 0$, $G = f_1 = (1 + u^2)^{-1}$, then $Q = [z^1 \mid z^1 > 0]$, $Q_G = [(z^0, z^1) \mid z^0 > 0, z^1 > 0]$; hence $\operatorname{cl}\operatorname{co} Q_G = [(z^0, z^1) \mid z^0 \geqq 0, z^1 \geqq 0]$, $\operatorname{cl}\operatorname{co} Q = [z^1 \mid z^1 \geqq 0]$. The situation concerning the points $(z^0 = 0, z^1 = 0)$ and $z^1 = 0$ is now the same as before.

4.5. With the usual notation $x = (y, z)$, $y \in E_\alpha$, $z \in E_{n-\alpha}$, let A_0 denote a given closed subset of the ty-space $E_1 \times E_\alpha$, and let A be the closed set $A = A_0 \times E_{n-\alpha} \subset E_{n+1}$. For every $(t, y) \in A_0$ let $U(t, y)$ be a given closed subset of the u-space E_m, let M_0 be the set of all (t, y, u) with $(t, y) \in A_0$, $u \in U(t, y)$, and then $M = M_0 \times E_{n-\alpha}$ is the set of all (t, x, u) with $(t, x) \in A$, $x = (y, z)$, $u \in U(t, y)$. Let $f_1(t, y, u), \cdots, f_\alpha(t, y, u)$ be continuous functions defined on M_0 and for every $(t, x, u) \in M$, $x = (y, z)$, let $f(t, x, u) = f(t, y, u)$, $U(t, x) = U(t, y)$. Let C be a given compact subset of A, let N be the set of all (t, x, u), $x = (y, z)$, with $(t, x) \in C$, $u \in U(t, y)$, and let $G(t, x, u)$ be a given nonnegative lower semicontinuous function defined on N. Let C_0 be the projection of C on the space $E_1 \times E_\alpha$; hence C_0 is compact and $C_0 \subset A_0$. For every $(t, y) \in C_0$ we shall denote by $C(t, y)$ the set of all $z \in E_{n-\alpha}$ with $(t, y, z) \in C$. Then

$$C_0 \subset A_0 \subset E_1 \times E_\alpha,$$

$$C = [(t, y, z) \mid (t, y) \in C_0,\ z \in C(t, y)] \subset A = A_0 \times E_{n-\alpha}.$$

Obviously, $C(t, y)$ is a compact subset of $E_{n-\alpha}$ for every $(t, y) \in C_0$.

We shall now define the nonnegative function $H(t, y, u)$ by taking

$$H(t, y, u) = \inf_{z \in C(t,y)} G(t, y, z, u) \tag{21}$$

for every $(t, y) \in C_0$ and $u \in U(t, y)$. Thus $H \geqq 0$ is defined on the set N_0 of all (t, y, u) with $(t, y) \in C_0$, $u \in U(t, y)$.

(i) If $G(t, x, u) \geqq 0$ is lower semicontinuous on N, then $H(t, y, u)$ is lower semicontinuous on N_0.

Proof. First G is lower semicontinuous on N and so is its restriction on the (compact) set $C(t, y)$. Hence, we can replace min for inf in (21). Now let (t_0, y_0, u_0) be a point of N_0, and let $l = \liminf H(t, y, u)$, where lim inf is taken as $(t, y, u) \to (t_0, y_0, u_0)$ with $(t, y, u) \in N_0$. Then there is a sequence (t_k, y_k, u_k), $k = 1, 2, \cdots$, of points of N_0 with $H(t_k, y_k, u_k) \to l$, and we have to prove that $H(t_0, y_0, u_0) \leqq l$. Here $0 \leqq l \leqq +\infty$, $t_k \to t_0$, $y_k \to y_0$, $u_k \to u_0$, $u_k \in U(t_k, y_k)$. If $l = +\infty$ the statement is trivial. Assume $0 \leqq l < +\infty$. By the definition of H (with inf replaced by min), for every k there is some point $z_k \in C(t_k, y_k)$ such that $H(t_k, y_k, u_k) = G(t_k, y_k, z_k, u_k)$, $k = 1, 2, \cdots$. The set C is compact, hence closed,

and thus $C(t, y)$ satisfies property (U) on C_0 by force of Remark 3. Also, $(t_k, y_k, z_k) \in C$; hence $[z_k]$ is a bounded sequence, and we can extract a convergent subsequence, say still $[z_k]$, with $z_k \to z_0 \in E_{n-\alpha}$. Now for every $\delta > 0$, we have $z_k \in C(t_0, y_0, \delta)$ for k sufficiently large; hence $z_0 \in \text{cl}\, C(t_0, y_0, \delta)$ for every $\delta > 0$ and by property (U), also,

$$z_0 \in \bigcap_\delta \text{cl}\, C(t_0, y_0, \delta) = C(t_0, y_0)$$

or $(t_0, y_0, z_0) \in C$. Then $t_k \to t_0$, $y_k \to y_0$, $u_k \to u_0$, $z_k \to z_0$ as $k \to \infty$, and hence,

$$G(t_0, y_0, x_0, u_0) \leqq \liminf G(t_k, y_k, z_k, u_k) = \liminf H(t_k, y_k, u_k) = l,$$

$$H(t_0, y_0, u_0) = \inf G(t_0, y_0, z, u_0) \leqq G(t_0, y_0, z_0, u_0) \leqq l,$$

where inf is taken for $z \in C(t_0, y_0)$. The statement is thereby proved.

(ii) If $1, f_i(t, y, u)$, $i = 1, 2, \cdots, \alpha$, are of slower growth than $C(t, x, u)$ on C, then $1, f_i$, $i = 1, 2, \cdots, \alpha$, are of slower growth than $H(t, y, u)$ on C_0.

Proof. Indeed, given $\epsilon > 0$, there is $\bar{u} = \bar{u}(\epsilon) \geqq 0$ such that $1 \leqq \epsilon G(t, y, z, u)$, $|f_i(t, y, u)| \leqq \epsilon G(t, y, z, u)$ for all $|u| \geqq \bar{u}(\epsilon)$, and hence,

$$\epsilon H(t, y, u) = \inf [\epsilon G(t, y, z, u)] \geqq 1,$$

$$\epsilon H(t, y, u) = \inf [\epsilon G(t, y, z, u)] \geqq |f_i(t, y, u)|, \quad i = 1, \cdots, \alpha,$$

for all $(t, y) \in C_0$, $u \in U(t, y)$, and where inf is taken in $C(t, y)$.

(iii) If for every $\epsilon > 0$ there is a constant $M_\epsilon \geqq 0$ such that $|f_i(t, y, u)| \leqq M_\epsilon + \epsilon G(t, y, z, u)$ on N, then we have also $|f_i(t, y, u)| \leqq M_\epsilon + \epsilon H(t, y, u)$ on N_0.

Proof. Indeed,

$$M_\epsilon + \epsilon H(t, y, u) = \inf [m_\epsilon + \epsilon G(t, y, z, u)] \geqq |f_i(t, y, u)|, \quad i = 1, \cdots, \alpha.$$

4.6. In [7] McShane has given an existence statement for usual solutions [7, (13.12)], in which a relation of the form (17) is assumed to be satisfied by the elements of a minimizing sequence, but the same relation need not be satisfied by the optimal pairs. We shall prove below an existence theorem of the same type (Theorem B). In Remark 9 we shall show that Theorem B includes McShane's statement. An analogous situation will be shown to occur for weak solutions in §5.

EXISTENCE THEOREM B. *Let α, n, $0 \leqq \alpha \leqq n$, be given integers, and for every $x = (x^1, \cdots, x^n)$, let $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$, so that $x = (y, z)$. Let A_0 be a closed subset of the ty-space $E_{\alpha+1}$, and then $A = A_0 \times E_{n-\alpha}$ is a closed subset of the tx-space E_{n+1}. For every $(t, y) \in A_0$ let $U(t, y)$ be a closed subset of the u-space E_m satisfying property (U) in A_0. Let $f(t, y, u) = (f_1, \cdots, f_n)$ be a given vector function defined on the set M_0*

of all (t, y, u) *with* $(t, y) \in A_0$, $u \in U(t, y)$. *Assume* $f_{\alpha+1}, \cdots, f_n$ *nonnegative and lower semicontinuous on* M_0, *and* $f_1, \cdots, f_\alpha$ *continuous on* M_0. *Let* B *be a given closed subset of the* $t_1x_1t_2x_2$*-space* E_{2n+2}, $x_1 = (x_1^1, \cdots, x_1^n)$, $x_2 = (x_2^1, \cdots, x_2^n)$, *and assume that* B *is independent of* $x_2^{\alpha+1}, \cdots, x_2^n$. *Let* $e(t_1, x_1, t_2, x_2)$ *be a given continuous real-valued function on* B, *which is monotone nondecreasing in each of the variables* $x_2^{\alpha+1}, \cdots, x_2^n$. *Let* C *be a given compact subset of* A, let C_0 *be the projection of* C *on* A_0, *and let* $H(t, y, u)$ *be a given nonnegative and lower semicontinuous function on the set* N_0 *of all* (t, y, u) *with* $(t, y) \in C_0$, $u \in U(t, y)$. *Let* $Q_H(t, y)$, $Q(t, y)$ *be the sets defined by*

$$Q_H(t, y) = [(z^0, z) \mid z^0 \geqq H(t, y, u), z^i = f_i(t, y, u), i = 1, \cdots, \alpha, \quad z^i \geqq f_i(t, y, u), i = \alpha + 1, \cdots, n, u \in U(t, y)] \subset E_{n+1}, \tag{22}$$

$$Q(t, y) = [z \mid z^i = f_i(t, y, u), i = 1, \cdots, \alpha, \quad z^i \geqq f_i(t, y, u), i = \alpha + 1, \cdots, n, u \in U(t, y)] \subset E_n, \tag{23}$$

and assume that $Q(t, y)$ *satisfies property* (Q) *with respect to* $Q_H(t, y)$ *in* C_0, *and that* $Q(t, y)$ *is convex for every* $(t, y) \in C_0$. *Finally, assume that for every* $i = 1, \cdots, \alpha$ *the following growth condition is satisfied*:

(γ_i) *For every* $\epsilon > 0$ *there is a constant* $M_{i\epsilon} \geqq 0$ *such that* $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon H(t, y, u)$ *for all* $(t, y, u) \in N_0$.

Let Ω *be the class of all admissible pairs* x, u (*defined as in* §4.1 *in relation to the sets* A, $U(t, y)$, B *and function* f), *and assume* Ω *to be not empty. Assume that there is a sequence* $x_k(t), u_k(t), t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, *of admissible pairs* (*elements of* Ω) *such that*: (a) $(t, x_k(t)) \in C$ *for all* $t \in [t_{1k}, t_{2k}]$, $k = 1, 2, \cdots$; (b) $\int_{t_{1k}}^{t_{2k}} H(t, y_k(t), u_k(t))\, dt \leqq M_1$ *for some constant* $M_1 \geqq 0$ *and all* $k = 1, 2, \cdots$; (c) $I[x_k, u_k] = e[\eta(x_k)] \to i$ *as* $k \to \infty$, *where* i *is the infimum of* $I[x, u]$ *in* Ω.

Then the functional $I[x, u] = e[\eta(x)]$ *has an absolute minimum in* Ω.

If we know that $Q_H(t, y)$ is convex for every $(t, y) \in C_0$ and satisfies property (Q) in C_0, then the optimal pairs satisfy

$$\int_{t_1}^{t_2} H(t, y(t), u(t))\, dt \leqq M_1.$$

As in §4.2, any of the requirements (γ_i) and (b) may be replaced by a corresponding restriction (18) on the class Ω.

Also, as in §4.2, the requirement that Q possesses property (Q) with respect to Q_H is certainly satisfied if we know that Q is convex and that 1, $f_1, \cdots, f_\alpha$ are of slower growth than H uniformly in C_0. In this situation, all conditions (γ_i), $i = 1, \cdots, \alpha$, are certainly satisfied.

As in §4.2, the requirement $H \geqq 0$, $f_i \geqq 0$, $i = \alpha + 1, \cdots, n$, can be

softened into $H \geqq -K_0, f_i \geqq -K_i$, $i = \alpha + 1, \cdots, n$, for some constants K_0, K_i, provided in (γ_i) we require $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon[H(t, y, u) + K_0]$.

Finally, if H is one of the functions $f_{\alpha+1}, \cdots, f_n$, say, $H = f_n$, then $0 \leqq \alpha \leqq n - 1$, and a relation as in (b) is satisfied by all admissible pairs with $M_1 = b_n - a_n$. We can then disregard requirement (b) and require as in §4.2 that $Q(t, y)$ is convex and satisfies condition (Q) in C_0.

Proof of Existence Theorem B. The same proof of Theorem A holds with obvious changes, and C_0 replacing A_0. We shall then apply the second part of Closure Theorem III instead of the first one in order to conclude that $X(t) = (y(t), Z(t))$ is a trajectory in E_n. Then $X(t)$ is proved to be optimal by the same argument as for Theorem A.

Remark 11. McShane's statement for usual solutions [7, (13.12)], differs essentially from Theorem B because of the following two points: (i) the role of our function $H(t, y, u)$ nonnegative and lower semicontinuous on N_0 is taken in [7] by a function $G(t, x, u)$ nonnegative and continuous on the set N of all (t, x, u) with $(t, x) \in C$, $u \in U(t, y)$, $x = (y, z)$; (ii) instead of the hypothesis "Q is convex, Q has property (Q_r), and conditions (γ_i) hold, $i = 1, \cdots, \alpha$," there is in [7] the hypothesis that Q is convex and $1, f_1, \cdots, f_\alpha$ are of slower growth than G uniformly in C. If we introduce the auxiliary function H as in §4.5, then the hypothesis that G is continuous implies certainly that $H(t, y, u)$ is lower semicontinuous, and the hypothesis that $1, f_1, \cdots, f_\alpha$ are of slower growth than G implies that the same functions are of slower growth than H (§4.5, (ii)), that all conditions (γ_i) are satisfied (§4.5, (iii)), and Q has certainly property (Q_r) (§2.2 (ii)). This shows that McShane's statement [7, (13.12)], is essentially contained in Theorem B.

Example. Take $n = 1, m = 1, f_1 = t^2u, G = tu^2, 0 \leqq t \leqq 1, 0 \leqq x \leqq 1$, $U = [-\infty < u < +\infty]$; then condition (β) is satisfied; that is, $G \geqq 0$, and for every $\epsilon > 0$ there is a constant M_ϵ with $|f_1| \leqq M_\epsilon + \epsilon G$, as we have shown in §4.3. Also we have $H = G$. On the other hand, the sets Q, Q_G are here $Q(t, x) = [z \mid z^1 = t^2u, -\infty < u < -\infty]$, and $Q_G(t, x) = [(z^0, z^1) \mid z^0 \geqq tu^2, z^1 = t^2u, -\infty < u < +\infty]$. If A, say, is the set $A = [0, 1]^2$, then $Q(0, x) = [z^1 = 0]$, $Q_G(0, x) = [z^0 \geqq 0, z^1 = 0]$, while for $0 < t \leqq 1$, $Q_G(t, x) = [z^0 \geqq t^{-3}(z^1)^2, -\infty < z^1 < +\infty]$, $Q(t, x) = [-\infty < z^1 < +\infty]$. Thus, the set $Q_G(t, x)$ possesses property (Q) in A, and the set $Q(t, x)$ possesses property (Q) with respect to $Q_G(t, x)$ in A. Let us take $t_1 = 0$, $0 \leqq t_2 \leqq 1$, $x(0) = 0$, $x(1) = 1$, let $B = [t_1 = 0, x_1 = 0, 0 \leqq t_2 \leqq 1, x_2 = 1]$, and take $e = t_2$ (problem of minimum time to transfer x from 0 to 1). Let Ω be the class of all admissible pairs x, u such that $\int_0^1 tu^2\,dt \leqq M_0$ and take M_0 sufficiently large that Ω is not empty. Then there is an absolute minimum for $e[\eta(x)] = t_2$ in Ω by force of Theorem A. Here 1 is not of slower growth than G uniformly in A.

4.7. We list below only two corollaries of Theorem A for Lagrange problems. Here we have a functional $I[x, u]$ of the form

$$I[x, u] = \int_{t_1}^{t_2} f_0(t, x, u)\, dt, \tag{24}$$

a system of ordinary differential equations $dx/dt = f(t, x, u)$ with $x = (x^1, \cdots, x^n), f = (f_1, \cdots, f_n)$, with boundary conditions of the form $(t_1, x(t_1), t_2, x(t_2)) \in B \subset E_{2n+2}$, and unilateral constraints on the space variable and the control variable of the usual form $(t, x(t)) \in A$, $u \in U(t, x)$. Thus the class Ω of admissible pairs $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, is now defined by the requirements: (a) $x(t)$ AC in $[t_1, t_2]$; (b) $u(t)$ measurable in $[t_1, t_2]$; (c) $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$; (d) $u(t) \in U(t, x(t))$ a.e. in $[t_1, t_2]$; (e) $dx/dt = f(t, x(t), u(t))$ a.e. in $[t_1, t_2]$; (f) $(t_1, x(t_1), t_2, x(t_2)) \in B$; (g) $f_0(t, x(t), u(t))$ L-integrable in $[t_1, t_2]$.

With the use of the variable x^0 satisfying $x^0(t_1) = 0$ and $dx^0/dt = f_0(t, x, u)$, the new vector variable is $\tilde{x} = (x^0, x^1, \cdots, x^n)$, and $I[x, u] = e = x_2^0$; thus e is monotone nondecreasing in $x_2^0, f_0, f_1, \cdots, f_n$ do not depend on x^0 and B is now replaced by $\tilde{B} = (x_1^0 = 0) \times E_1 \times B$, which is independent of x_2^0.

COROLLARY 1 (for Lagrange problems and usual solutions). *Let α, n, $0 \leqq \alpha \leqq n$, be given integers, and for every $x = (x^1, \cdots, x^n)$ let y, z denote $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$, so that $x = (y, z)$. Let A_0 be a compact subset of the ty-space $E_{\alpha+1}$, let I be a closed finite interval of the z-space $E_{n-\alpha}$, so that $A = A_0 \times I \subset E_{n+1}$ is also compact. For every $(t, y) \in A_0$ let $U(t, y)$ be a given closed subset of the u-space E_m satisfying property* (U) *in A_0. Let $f_0(t, y, u)$, $f(t, y, u) = (f_1, \cdots, f_n)$ be given functions defined on the set M_0 of all (t, y, u) with $(t, y) \in A_0$, $u \in U(t, y)$. Assume $f_0, f_{\alpha+1}, \cdots, f_n$ nonnegative and lower semicontinuous on M_0, and $f_1, \cdots, f_\alpha$ continuous on M. Assume*: (i) *that the set $Q_{f_0}(t, y) = [(z^0, \cdots, z^n) \mid z^0 \geqq f_0(t, y, u), z^i = f_i(t, y, u), i = 1, \cdots, \alpha, z^i \geqq f_i(t, y, u), i = \alpha + 1, \cdots, n, u \in U(t, y)]$ is convex for every $(t, y) \in A_0$*; (ii) *the set $Q_{f_0}(t, y)$ satisfies property* (Q) *in A_0*; (iii) *for every $\epsilon > 0$ there is a constant $M_\epsilon \geqq 0$ such that $|f_i(t, y, u)| \leqq M_\epsilon + \epsilon f_0(t, y, u)$ for all $(t, y, u) \in M_0$, $i = 1, \cdots, \alpha$. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2} and assume that B is independent of $x_2^{\alpha+1}, \cdots, x_2^n$. Let Ω_0 be the class of all admissible pairs x, u as defined above, and assume that Ω_0 is not empty.*

Then the functional $I[x, u] = \int_{t_1}^{t_2} f_0(t, y(t), u(t))\, dt$ has an absolute minimum in Ω.

Note that both conditions (ii) and (iii) are certainly satisfied if (α) 1, $f_1, \cdots, f_\alpha$ are of slower growth than f_0 uniformly in A_0.

COROLLARY 2 (for Lagrange problems and usual solutions). *Let A be a compact subset of the x-space E_n. For every $(t, x) \in A$ let $U(t, x)$ be a closed*

subset of the u-space E_m satisfying property (U) *in A. Let* $f_0(t, x, u), f(t, x, u) = (f_1, \cdots, f_n)$ *be given functions defined on the set M of all* (t, x, u) *with* $(t, x) \in A$, $u \in U(t, x)$. *Assume* f_0 *nonnegative and lower semicontinuous, and* $f_1, \cdots, f_\alpha$ *all continuous on M. Assume*: (i) *that the set* $Q_{f_0}(t, x) = [(z^0, \cdots, z^n) \mid z^0 \geqq f_0(t, x, u), z^i = f_i(t, x, u), i = 1, \cdots, n, u \in U(t, x)]$ *is convex for every* $(t, x) \in A$; (ii) *the set* $Q_{f_0}(t, x)$ *satisfies property* (Q) *in A*; (iii) *for every* $\epsilon > 0$ *there is a constant* $M_\epsilon > 0$ *such that* $|f_i(t, x, u)| \leqq M_\epsilon + \epsilon f_0(t, x, u)$ *for all* $(t, x, u) \in M$, $i = 1, \cdots, n$. *Let B be a closed subset of the* $t_1x_1t_2x_2$*-space* E_{2n+2}. *Let* Ω_0 *be the class of all admissible pairs x, u as defined above, and assume that* Ω_0 *is not empty.*

Then the functional $I[x, u] = \int_{t_1}^{t_2} f_0(t, x, u)\, dt$ *has an absolute minimum in* Ω_0.

Note that both conditions (ii) and (iii) are certainly satisfied if (α) $1, f_1, \cdots, f_n$ are of slower growth than f_0 uniformly in A.

In Corollary 1 the hypothesis $f_0 \geqq 0, f_i \geqq 0, i = \alpha + 1, \cdots, n$, can be softened into $f_0 \geqq -K_0, f_i \geqq -K_i, i = \alpha + 1, \cdots, n$, for some constants K_0, K_i, provided in (iii) we require $|f_i(t, y, u)| \leqq M_\epsilon + \epsilon[f_0(t, y, u) + K_0]$, $i = 1, \cdots, \alpha$. An analogous remark holds for Corollary 2 where $\alpha = n$.

Remark 12. Corollary 1 contains as a particular case Corollary 2 (for $\alpha = n$) and this in turn contains as a particular case Existence Theorem 1 of [1, p. 390]. These statements are thus particular cases of Existence Theorem A of the present paper. Note that in Theorem 1 of [1, p. 390] we required the strong growth condition: (ϕ) there are constants $C, D \geqq 0$ and a continuous function $\Phi(\zeta)$, $0 \leqq \zeta < +\infty$, with $\Phi(\zeta)/\zeta \to +\infty$ as $\zeta \to +\infty$, such that $|f(t, x, u)| \leqq C + D|u|, f_0(t, x, u) \geqq \Phi(|u|)$ for all $(t, x, u) \in M$. This condition (ϕ) certainly implies (α) of Corollary 2, and then both conditions (ii) and (iii) are satisfied. As mentioned in Remark 2 of §2.1, the hypothesis that Q satisfies property (Q) of Theorem 1 in [1, p. 390] has been proved here to be a consequence of the other hypotheses of the same Theorem 1. Note that condition (ϕ) certainly implies $f_0 \geqq -K_0$ for some constant K_0.

5. Existence theorems for weak solutions. Instead of the problem sketched at the beginning of §4, we consider here the differential system

$$\frac{dx}{dt} = f^*(t, x(t), p(t), v(t)) = \sum p_j(t) f(t, x(t), u^{(j)}(t)), \tag{25}$$

with the same boundary conditions

$$(t_1, x(t_1), t_2, x(t_2)) \in B \subset E_{2n+2},$$

the same functional to be minimized:

$$I[x, u] = e(t_1, x(t_1), t_2, x(t_2)),$$

and constraints:

$$(t, x(t)) \in A, \quad u^{(j)}(t) \in U(t, x(t)), \quad j = 1, \cdots, \mu, \tag{26}$$
$$p_j(t) \geqq 0, \quad \sum p_j(t) = 1,$$

where $\sum$ ranges over all $j = 1, \cdots, \mu$, and (p, v) denotes the new control variable $(p, v) = (p_1, \cdots, p_\mu, u^{(1)}, \cdots, u^{(\mu)})$. The constraints (26) can be written in the usual form

$$(t, x(t)) \in A, \quad (p(t), v(t)) \in V(t, x(t)),$$

where now $V(t, x) = \Gamma \times [U(t, x)]^\mu$, and Γ is the simplex $[p_j \geqq 0, j = 1, \cdots, \mu, \sum p_j = 1]$. In other words, we have a new Mayer-type problem with an extended system of control variables. The systems $x(t)$, $p(t)$, $v(t)$ relative to the new problems are called "weak solutions" or "generalized solutions" of the original problem. As in §4, we shall now define the terms more precisely.

5.1. Let $x = (x^1, \cdots, x^n) \in E_n$, $u = (u^1, \cdots, u^m) \in E_m$, let A be a given subset of the tx-space E_{n+1}, and for every $(t, x) \in A$ let $U(t, x)$ be a given subset of the u-space E_m. Let M denote the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$. Let $G(t, x, u)$ be a real-valued function and $f(t, x, u) = (f_1, \cdots, f_n)$ a vector-valued function both defined on M. Let $p = (p_1, \cdots, p_\mu)$ and let Γ be the simplex $[p_j \geqq 0, j = 1, \cdots, \mu, \sum p_j = 1]$, where $\sum$ ranges over all $j = 1, \cdots, \mu$. Let $v = (u^{(1)}, \cdots, u^{(\mu)})$ with $u^{(j)} \in U(t, x)$, and let f^*, G^* be the new functions

$$G^*(t, x, p, v) = \sum p_j G(t, x, u^{(j)}),$$
$$f^*(t, x, p, v) = (f_1{}^*, \cdots, f_n{}^*) = \sum p_j f(t, x, u^{(j)}),$$

which are defined on the set M^* of all (t, x, p, v) with $(t, x) \in A, p \in \Gamma$, $v \in [U(t, x)]^\mu$. Let B be a given subset of the $t_1x_1t_2x_2$-space E_{2n+2}.

We shall say that a system of functions $x(t)$, $p(t)$, $v(t)$, $t_1 \leqq t \leqq t_2$, is *admissible* provided: (i) $x(t) = (x^1, \cdots, x^n)$ is AC in $[t_1, t_2]$; (ii) $(p(t), v(t)) = (p_1, \cdots, p_\mu, u^{(1)}, \cdots, u^{(\mu)})$ is measurable in $[t_1, t_2]$; (iii) $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$; (iv) $(p(t), v(t)) \in V(t, x(t)) = \Gamma \times [U(t, x(t))]^\mu$ a.e. in $[t_1, t_2]$; (v) $dx/dt = f^*(t, x(t), p(t), v(t))$ a.e. in $[t_1, t_2]$; (vi) $(t_1, x(t_1), t_2, x(t_2)) \in B$. As in §4.1 we consider the functional $I[x, p, v] = e[\eta(x)] = e(t_1, x(t_1), t_2, x(t_2))$.

We say that $[p(t), v(t)]$ is a *generalized strategy*, that $p(t)$ is a *probability distribution*, and that $x(t)$ is a *generalized trajectory*. We shall also say, as mentioned above, that $x(t), p(t), v(t)$ is a *weak*, or *generalized solution*.

Given any integer α, $0 \leqq \alpha \leqq n$, we shall denote by $R(t, x)$, $R_\alpha(t, x)$ the sets:

$$R(t,x) = [(z^1, \cdots, z^n) \mid z^i = f_i^*(t,x,p,v),\ i = 1, \cdots, \alpha,$$
$$z^i \geqq f_i^*(t,x,p,v),\ i = \alpha+1, \cdots, n,\ p \in \Gamma,\ v \in U^\mu(t,x)],$$
$$R_G(t,x) = [(z^0, z^1, \cdots, z^n) \mid z^0 \geqq G^*(t,x,p,v),\ z^i = f_i^*(t,x,p,v),$$
$$i = 1, \cdots, \alpha,\ z^i \geqq f_i^*(t,x,p,v),\ i = \alpha+1, \cdots, n,\ p \in \Gamma,$$
$$v \in U^\mu(t,x)].$$

As we have shown in [1, p. 415], the sets $R(t,x)$, $R_G(t,x)$ are the sets of all points of E_n, or E_{n+1}, which are convex combinations of all possible systems of μ points of $Q(t,x)$, $Q_G(t,x)$, respectively. Thus, by taking $\mu \geqq n+1$, or $\mu \geqq n+2$, the sets R, R_G are the convex hulls of the sets Q, Q_G, respectively, and thus R, R_G are certainly convex.

5.2. Condition (α) of §4.3, that is, (α) $1, f_1, \cdots, f_\alpha$ of slower growth than $G \geqq 0$, does *not* imply an analogous situation for $1, f_1^*, \cdots, f_\alpha^*, G^*$. Indeed, for instance, take $n = 1, m = 1, G = u^2, f_1 = u, u \in U = [-\infty < u < +\infty]$, $\mu = 2$, so that $G^* = p_1(u^{(1)})^2 + p_2(u^{(2)})^2, f_1^* = p_1u^{(1)} + p_2u^{(2)}$, with $p_1 \geqq 0$, $p_2 \geqq 0$, $p_1 + p_2 = 1$, $-\infty < u^{(1)}, u^{(2)} < +\infty$. Then, for $u^{(1)} = 0, u^{(2)} = k \geqq 0$, $p_1 = 1$, $p_2 = 0$, we have $G^* = 0, f_1^* = 0$. Thus, $G^* = 0$ though $|u^{(1)}| + |u^{(2)}| = k \to +\infty$, that is, 1 is not of slower growth than G^* uniformly in A when $(p,v) \to +\infty$, and this last relation is meant as the norm $|u^{(1)}| + |u^{(2)}| \to +\infty$ (or the same with any analogous norm). Nevertheless, condition (β) of §4.3, that is, (β) $G \geqq 0$, and for every $\epsilon > 0$ there is some M_ϵ such that $|f_i| \leqq M_\epsilon + \epsilon G$, $i = 1, \cdots, \alpha$, *does* imply an analogous situation on f_i^*, G_i^*, since

$$|f_i^*(t,x,p,v)| = |\sum_j p_j f_i(t,x,p,u^{(j)})|$$
$$\leqq \sum_j p_j |f_i(t,x,p,u^{(j)})|$$
$$\leqq \sum_j p_j (M_\epsilon + \epsilon G(t,x,p,u^{(j)}))$$
$$= M_\epsilon + \epsilon G^*.$$

If $G, f_{\alpha+1}, \cdots, f_n$ are nonnegative and lower semicontinuous, and condition (α) of §4.3 holds, that is, $1, f_1, \cdots, f_\alpha$ are of slower growth than G, then $f_1^*, \cdots, f_\alpha^*$ may not verify an analogous growth property with respect to G^*; nevertheless, the set $R_G(t,x)$ (now certainly convex) satisfies property (Q) as proved in §2.2 (iii), and the set $R(t,x)$ (now convex) satisfies property (Q) with respect to $R_G(t,x)$. On the other hand, the weaker condition (β) of §4.3 ($|f_i| \leqq M_\epsilon + \epsilon G$, $i = 1, \cdots, \alpha$) implies an analogous property for f_i^*, G^*, and this property can be used, as in the proof of Theorem A, to show that the corresponding coordinates x_k^i of

the trajectories $x_k(t)$ of a minimizing sequence $x_k(t)$, $p_k(t)$, $u_k(t)$, $k = 1, 2, \cdots$, of generalized systems, are equiabsolutely continuous.

We are now in a position to state Existence Theorems A*, B* for weak optimal solutions, which are analogous to Theorems A, B of §4.2 and §4.6 for usual solutions. With the remarks above, Theorems A*, B* are essentially corollaries of Theorems A, B. As mentioned, we always assume below that μ has been chosen large enough so that all sets R, R_σ, or analogous ones, are convex.

5.3.

EXISTENCE THEOREM A* (for weak solutions). *Let α, n, $0 \leqq \alpha \leqq n$, be given integers, and for $x = (x^1, \cdots, x^n)$ let y, z denote $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$, so that $x = (y, z)$. Let A_0 be a compact subset of the ty-space $E_{\alpha+1}$, let I be a finite closed interval of the z-space $E_{n-\alpha}$, $I = [a_{\alpha+1}, b_{\alpha+1}] \times \cdots \times [a_n, b_n]$, and thus $A = A_0 \times I$ is a compact subset of the tx-space E_{n+1}. For every $(t, y) \in A_0$ let $U(t, y)$ be a given closed subset of the u-space E_m satisfying property (U) in A_0. Let M_0 be the set of all (t, y, u) with $(t, y) \in A_0$, $u \in U(t, y)$; $M = M_0 \times I$ is then the set of all (t, x, u) with $(t, x) \in A, u \in U(t, y), x = (y, z)$. Let $f(t, y, u) = (f_1, \cdots, f_n)$, $H(t, y, u)$ be functions defined on M_0, and assume that $f_1, \cdots, f_\alpha$ are continuous on M_0, and that $H, f_{\alpha+1}, \cdots, f_n$ are nonnegative and lower semicontinuous on M_0. Let us assume that for every $i = 1, \cdots, \alpha$ the following growth condition holds:*

(γ_i) There is a constant $M_{i\epsilon} \geqq 0$ such that $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon H(t, y, u)$ for all $(t, y, u) \in M_0$.

For every $(t, y) \in A_0$ let $R_H(t, y)$ be the set of all $z = (z^0, z^1, \cdots, z^n) \in E_{n+1}$ defined by

$$R_H(t, y) = [\tilde{z} \mid z^0 \geqq H^*(t, y, p, v), z^i = f_i^*(t, y, p, v), i = 1, \cdots, \alpha, \quad z^i \geqq f_i^*(t, y, p, v), i = \alpha + 1, \cdots, n, p \in \Gamma, \quad v \in U^\mu(t, y)] \subset E_{n+1}, \tag{27}$$

and assume that the (convex) set $R_H(t, y)$ satisfies property (Q) in A_0. For every $(t, x, u) \in M$, $x = (y, z)$, we shall write $f(t, x, u) = f(t, y, u)$, etc. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}, $x_1 = (x_1^1, \cdots, x_1^n)$, $x_2 = (x_2^1, \cdots, x_2^n)$, and assume that B is independent of $x_2^{\alpha+1}, \cdots, x_2^n$. Let $e(t_1, x_1, t_2, x_2)$ be a real-valued continuous function defined on B, which is monotone nondecreasing with respect to each variable $x_2^{\alpha+1}, \cdots, x_2^n$. Let Ω^* be the class of all admissible systems x, p, v for which $H^*(t, y(t), p(t), v(t))$ is L-integrable in $[t_1, t_2]$ and

$$\int_{t_1}^{t_2} H^*(t, y(t), p(t), v(t))\, dt \leqq M_1 \tag{28}$$

for some constant $M_1 \geqq 0$, and assume that Ω^* is not empty.

Then the functional $I[x, u] = e[\eta(x)]$ *has an absolute minimum in* Ω^* *(and the optimal generalized system satisfies* (17)).

The requirement $H \geqq 0, f_i \geqq 0, i = \alpha + 1, \cdots, n$, can be replaced by the weaker one $H \geqq -K_0, f_i \geqq -K_i, i = \alpha + 1, \cdots, n$, for some constants K_0, K_i, provided we replace (γ_i) by the analogous requirement $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon[H(t, y, u) + K_0]$.

For any of the indices $i = 1, \cdots, \alpha$, condition (γ_i) can be disregarded, provided we impose a further restriction on Ω^*, namely, that Ω^* is made up of only those systems x, p, v for which (28) holds, and also

$$\int_{t_1}^{t_2} \left| \frac{dx^i}{dt} \right|^{p_i} dt \leqq M_i \tag{29}$$

for some constants $M_i \geqq 0$ and $p_i > 1$ and at least all $i = 1, \cdots, \alpha$, for which (γ_i) does not hold.

The condition that $R_H(t, y)$ satisfies property (Q) is certainly satisfied if $1, f_1, \cdots, f_\alpha$ are of slower growth than H uniformly in A_0. In this situation conditions (γ_i), $i = 1, \cdots, \alpha$, are all necessarily satisfied.

Finally, H may be one of the functions $f_{\alpha+1}, \cdots, f_n$, say, $H = f_n$, and then $0 \leqq \alpha \leqq n - 1$, and relation (28) is satisfied by all admissible systems x, p, v with $M_1 = b_n - a_n$. In this situation (28) can be disregarded and instead of R_H we may consider the simpler set $R(t, y)$ of all $z = (z^1, \cdots, z^n) \in E_n$ defined by

$$\begin{aligned} R(t, y) = [z \mid z^i = f_i^*(t, y, p, v), i = 1, \cdots, \alpha, z^i \geqq f_i^*(t, y, p, v), \\ i = \alpha + 1, \cdots, n, p \in \Gamma, v \in U^\mu(t, y)] \subset E_n. \end{aligned} \tag{30}$$

Then we shall replace the requirement concerning R_H by an analogous requirement on $R(t, y)$. In other words, we shall require that the (convex) set $R(t, y)$ satisfies property (Q) in A_0.

Existence Theorem B* (for weak solutions). *Let* $\alpha, n, 0 \leqq \alpha \leqq n$, *be given integers, and for every* $x = (x^1, \cdots, x^n)$ *let* $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$, *so that* $x = (y, z)$. *Let* A_0 *be a closed subset of the ty-space* $E_{\alpha+1}$, *and then* $A = A_0 \times E_{n-\alpha}$ *is a closed subset of the tx-space* E_{n+1}. *For every* $(t, y) \in A_0$ *let* $U(t, y)$ *be a closed subset of the u-space* E_m *satisfying property* (U) *in* A_0. *Let* $f(t, y, u) = (f_1, \cdots, f_n)$ *be a given vector function defined on the set* M_0 *of all* (t, y, u) *with* $(t, y) \in A_0, u \in U(t, y)$. *Assume* $f_{\alpha+1}, \cdots, f_n$ *nonnegative and lower semicontinuous on* M_0, *and* $f_1, \cdots, f_\alpha$ *continuous on* M_0. *Let* B *be a given closed subset of the* $t_1x_1t_2x_2$*-space* E_{2n+2}, $x_1 = (x_1^1, \cdots, x_1^n)$, $x_2 = (x_2^1, \cdots, x_2^n)$, *and assume that* B *is independent of* $x_2^{\alpha+1}, \cdots, x_2^n$. *Let* $e(t_1, x_1, t_2, x_2)$ *be a given continuous real-valued function on* B, *which is monotone nondecreasing in each of the variables* $x_2^{\alpha+1}, \cdots, x_2^n$. *Let* C *be a given compact subset of* A, *let* C_0 *be the projection of* C *on* A_0,

and let $H(t, y, u)$ *be a given nonnegative and lower semicontinuous function on the set* N_0 *of all* (t, y, u) *with* $(t, y) \in C_0$, $u \in U(t, y)$. *Let* $R_H(t, y)$, $R(t, y)$ *be the sets defined by relations* (27) *and* (30), *and assume that* $R(t, y)$ *satisfies property* (Q) *with respect to* $R_H(t, y)$ *in* C_0, *and that* $R(t, y)$ *is convex for every* $(t, y) \in C_0$. *Finally, assume that for every* $i = 1, \cdots, \alpha$, *the following growth condition is satisfied*:

(γ_i) *For every* $\epsilon > 0$ *there is a constant* $M_{i\epsilon} \geqq 0$ *such that* $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon H(t, y, u)$ *for all* $(t, y, u) \in N_0$.

Let Ω^* *be the class of all admissible systems* x, p, v *and assume* Ω^* *to be not empty. Assume that there is a sequence* $x_k(t)$, $p_k(t)$, $v_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, *of admissible systems* (*elements of* Ω^*) *such that*: (a) $(t, x_k(t)) \in C$ *for all* $t \in [t_{1k}, t_{2k}]$, $k = 1, 2, \cdots$; (b) $\int_{t_{1k}}^{t_{2k}} H^*(t, y_k(t), p_k(t), v_k(t))\, dt \leqq M_1$ *for some constant* $M_1 \geqq 0$ *and all* $k = 1, 2, \cdots$; (c) $I[x_k, p_k, v_k] = e[\eta(x_k)] \to i$ *as* $k \to \infty$, *where* i *is the infimum of* $I[x, p, u]$ *in* Ω^*.

Then the functional $I[x, p, v] = e[\eta(x)]$ *has an absolute minimum in* Ω^*.

If we know that the (convex) set $R_H(t, y)$ satisfies property (Q) in C_0, then the optimal system satisfies $\int_{t_1}^{t_2} H^*(t, y(t), p(t), v(t))\, dt \leqq M_1$.

As in the previous theorem, the requirement $H \geqq 0$, $f_i \geqq 0$, $i = \alpha + 1, \cdots, n$, can be replaced by the weaker requirement $H \geqq -K_0$, $f_i \geqq -K_i$, $i = \alpha + 1, \cdots, n$, for some constants K_0, K_i, provided in (γ_i) we require $|f_i(t, y, u)| \leqq M_{i\epsilon} + \epsilon[H(t, y, u) + K_0]$.

Analogously, any of the requirements (γ_i) may be replaced by a corresponding restriction (29) on the class Ω^*.

The requirement that the (convex) set R possesses property (Q) with respect to R_H is certainly satisfied if $1, f_1, \cdots, f_\alpha$ are of slower growth than H uniformly in C_0. In this situation, all conditions (γ_i), $i = 1, \cdots, \alpha$, are certainly satisfied.

Finally, if H is one of the functions $f_{\alpha+1}, \cdots, f_n$, say $H = f_n$, then $0 \leqq \alpha \leqq n - 1$, and a relation as in (b) is satisfied by all admissible systems, with $M_1 = b_n - a_n$. We shall then disregard requirement (b), and require as usual that the (convex) set $R(t, y)$ satisfies condition (Q) in C_0.

Remark 13. By considerations similar to those of Remark 11 we can easily see that Existence Theorem B* for weak solutions contains McShane's Existence Theorem 2.7 of [7]. Also, as in §4.7 we can deduce from Theorem A* existence statements relative to weak solutions of Lagrange problems, which improve the corresponding statements of [1, pp. 427–428].

Remark 14. In [7] criteria are given, involving Pontryagin's principle, to insure that an optimal generalized solution x, p, v is actually a usual solution, that is, all p_j are zero but one which is equal to one.

REFERENCES

[1] L. Cesari, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints, I and II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–430.

[2] ———, *Existence theorems for optimal solutions in Pontryagin and Lagrange problems*, this Journal, 3 (1965), pp. 475–498.

[3] P. Brunovsky, *On the necessity of a certain convexity condition for lower closure of control problems*, this Journal, 6 (1968), pp. 174–185.

[4] A. F. Filippov, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[5] R. V. Gamkrelidze, *On sliding optimal states*, Soviet Math. Dokl., 3 (1962), pp. 559–561.

[6] J. LaPalm, *Existence theorems for problems of optimal control and the calculus of variations with exceptional sets*, Doctoral thesis, University of Michigan, Ann Arbor, 1967.

[7] E. J. McShane, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

[8] T. Nishiura, *On an existence theorem for optimal control*, this Journal, 5 (1967), pp. 532–544.

[9] A. Lasota and C. Olech, *On the closedness of the set of trajectories of a control system*, Bull. Acad. Polon. Sci., Sér. Sci. Math. Astronom. Phys., 14 (1966), pp. 615–621.

[10] E. J. McShane and R. B. Warfield, *On Fillippov's implicit functio ns lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

INVARIANT SUBSPACES AND THE CONTROLLABILITY AND OBSERVABILITY OF LINEAR DYNAMICAL SYSTEMS*

D. A. FORD† AND C. D. JOHNSON‡

Summary. The geometric mechanism of uncontrollability and unobservability for a particular class of scalar input–scalar output linear dynamical systems is characterized in terms of the invariant subspaces of the system matrix **A** and its transpose.

1. Introduction. An important class of linear dynamical systems can be described by the vector-matrix differential equations[1]

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + u(t)\mathbf{f}, \tag{1a}$$

$$y(t) = \langle \mathbf{h}, \mathbf{x}(t) \rangle, \tag{1b}$$

where $\mathbf{x}$ is a real n-vector, the system *state vector*; $\mathbf{A}$ is a real, constant, $n \times n$ matrix; $\mathbf{f}$ is a real constant n-vector; $u(t)$, the system *control*, or *input*, is a scalar function of time, $\mathbf{h}$ is a constant n-vector and $y(t)$ is the (scalar) system *output*.

Two basic notions in the study of dynamical systems are the concepts of complete controllability and the dual concept of complete observability. These are described in the following definitions.

DEFINITION 1. The dynamical system (1) is said to be *completely controllable* (c.c.) if and only if, for each pair of finite states $(\mathbf{x}_0, \mathbf{x}_T)$, there exist a finite interval $[t_0, T]$ and a control $u(t) = u(t; \mathbf{x}_0, \mathbf{x}_T, t_0, T)$, $t_0 \leqq t \leqq T$, such that the corresponding solution $\mathbf{x}(t)$ of (1a) satisfies $\mathbf{x}(t_0) = \mathbf{x}_0$ and $\mathbf{x}(T) = \mathbf{x}_T$.

In order to define complete observability in a like manner, it is convenient to consider, together with (1), the homogeneous (unforced) equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}. \tag{2}$$

DEFINITION 2. The state $\mathbf{x}$ of the dynamical system (1) is said to be *completely observable* (c.o.) if and only if, for each finite output $y(t)$ which satisfies (1b) and (2) on a positive interval $t_0 \leqq t \leqq T < \infty$, there corresponds a unique initial state $\mathbf{x}(t_0)$.

---

* Received by the editors May 5, 1967, and in revised form April 1, 1968. This research was conducted at the University of Alabama Research Institute, Huntsville, Alabama, and was supported in part by the National Aeronautics and Space Administration under Grant NsG-381.

† Department of Mathematics, Emory University, Atlanta, Georgia 30322.

‡ Department of Electrical Engineering, University of Alabama in Huntsville, Huntsville, Alabama 35807.

[1] $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner product of $\mathbf{x}$ and $\mathbf{y}$.

The general concepts of controllability and observability, as defined above, were originally introduced by Kalman [1], and have since been studied in some detail [2]–[7]. In [2] it was shown that, in special cases, the failure of (1) to be completely controllable can be ascribed to certain geometric conditions which exist between the vector $\mathbf{f}$ and the one- and $(n-1)$-dimensional invariant subspaces of the matrix $\mathbf{A}$. Similar results were obtained for the case of observability involving the vector $\mathbf{h}$ and the invariant subspaces of the transpose of $\mathbf{A}$. In the present note, the results obtained in [2] are generalized to include invariant subspaces of arbitrary dimension. By this means, it is possible to give a general characterization of the geometric mechanism of uncontrollability and/or unobservability for the system (1). In particular, explicit geometric descriptions are given for the set $F(\mathbf{A})$ of all vectors $\mathbf{f}$ for which (1) is not completely controllable, and the set $H(\mathbf{A})$ of all vectors $\mathbf{h}$ for which the system (1) is not completely observable.

2. Invariant subspaces. Controllability and observability of the class of linear dynamical systems described in §1 can be characterized in several ways. The particular characterizations given by the following well-known theorems [3], [4] will play a central role in the discussion which follows.

THEOREM 1. *The linear dynamical system* (1) *fails to be completely controllable if and only if there exists a real polynomial* $p(s)$ *of degree less than* n *such that* $p(\mathbf{A})\mathbf{f} = \mathbf{0}$.

THEOREM 2. *The linear dynamical system* (1) *fails to be completely observable if and only if there exists a real polynomial* $q(s)$ *of degree less than* n *such that* $q(\mathbf{A}')\mathbf{h} = \mathbf{0}$.[2]

Theorems 1 and 2 are often stated in the following alternative forms.

THEOREM 1a. *A necessary and sufficient condition that* (1) *be completely controllable is that the matrix* $\mathbf{K}$, *whose columns are the vectors* $\mathbf{f}, \mathbf{Af}, \mathbf{A}^2\mathbf{f}, \cdots, \mathbf{A}^{n-1}\mathbf{f}$, *has maximal*[3] *rank* n.

THEOREM 2a. *A necessary and sufficient condition that the system* (1) *is completely observable is that the matrix* $\mathbf{\Omega}$, *whose columns are the vectors* $\mathbf{h}, \mathbf{A}'\mathbf{h}, (\mathbf{A}')^2\mathbf{h}, \cdots, (\mathbf{A}')^{n-1}\mathbf{h}$, *has maximal rank* n.

In accomplishing the aim of characterizing the sets $F(\mathbf{A})$ and $H(\mathbf{A})$, strong use will be made of the observation in [2], and elsewhere, that the set $F(\mathbf{A})$ is the union of all proper invariant subspaces of the matrix $\mathbf{A}$, and the set $H(\mathbf{A})$ is the union of all proper invariant subspaces of the transpose of $\mathbf{A}$. In [2], it was shown that if λ is a real eigenvalue of $\mathbf{A}$, and if $\mathfrak{h}'$ is a *row*

[2] The transpose of the matrix $\mathbf{A}$ is denoted by $\mathbf{A}'$.

[3] It is remarked that Kalman and Gamkrelidze used the condition $\det \mathbf{K} \neq 0$ in independent and (almost) simultaneous researches published in 1957 [10], [11] (see also [12]).

eigenvector associated with λ (i.e., $\beta'\mathbf{A} = \lambda\beta'$), then $F(\mathbf{A})$ contains the $(n-1)$-dimensional hyperplane orthogonal to β. It is shown in the present paper that if the system (1) is to be controllable for any vector $\mathbf{f}$ at all, then each real eigenvalue, whether repeated or not, must yield only *one* eigenline. (That is, the row eigenvector β' associated with a given eigenvalue λ must be unique to within a constant (scalar) multiplier.) To see why this is the case, suppose that a repeated real eigenvalue λ has two row eigenvectors (β_1', β_2') which are linearly independent. It follows that (β_1, β_2) span a two-dimensional linear subspace S of E^n, every vector of which is a row eigenvector associated with λ. It was observed above that the $(n-1)$-dimensional hyperplane orthogonal to any one of the vectors in S is a subset of $F(\mathbf{A})$, but the union of all such hyperplanes, and consequently $F(\mathbf{A})$, would fill up all of E^n. Thus no vector $\mathbf{f}$ could exist for which the system (1) is controllable. The situation is the same with regard to complex eigenvalues with the exception that the one-dimensional "eigenlines" in the argument above are replaced by the real, two-dimensional, "eigenplanes" associated with each pair of complex conjugate eigenvalues.

Before proceeding to the details of the characterizations of $F(\mathbf{A})$ and $H(\mathbf{A})$, some well-known definitions and results from the theory of matrices will be stated.

DEFINITION 3. A subspace S of E^n is called $\mathbf{A}$-*invariant* if $\mathbf{Ax}$ belongs to S whenever $\mathbf{x}$ belongs to S.

DEFINITION 4. The *minimal* $\mathbf{A}$-*polynomial* $p(s)$ *of a vector* $\mathbf{x}$ is the polynomial of least degree such that $p(\mathbf{A})\mathbf{x} = \mathbf{0}$. The *minimal* $\mathbf{A}$-*polynomial* $p(s)$ *of an* $\mathbf{A}$-*invariant subspace* S is the polynomial of least degree such that $p(\mathbf{A})\mathbf{x} = \mathbf{0}$ for every vector $\mathbf{x}$ in S.

DEFINITION 5. If $\mathbf{x}$ is any real n-vector, then the subspace spanned by the vectors

$$\mathbf{x}, \mathbf{Ax}, \mathbf{A}^2\mathbf{x}, \cdots$$

is called an $\mathbf{A}$-*cyclic subspace of* E^n *generated by the vector* $\mathbf{x}$. Any subspace so obtained is called an $\mathbf{A}$-*cyclic subspace of* E^n.

3. Characterization of the set $F(\mathbf{A})$. Theorem 1 indicates that the criterion for the system (1) to be completely controllable is that the minimal $\mathbf{A}$-polynomial of the vector $\mathbf{f}$ has degree n; that is, that E^n be $\mathbf{A}$-cyclic with $\mathbf{f}$ as generator.[4] This implies, in particular, that there cannot exist a polynomial $q(s)$ of degree *less* than n such that $q(\mathbf{A}) = \mathbf{0}$. Thus if (1) is completely controllable for *some* $\mathbf{f}$, the characteristic polynomial $c(s)$ of $\mathbf{A}$ is the (unique) monic polynomial of least degree such that $c(\mathbf{A}) = \mathbf{0}$. A matrix

[4] Kalman [13], [14] has recently shown this criterion to be equivalent to the abstract algebraic condition that $\mathbf{f}$ generates a *module* over the polynomial ring.

having this latter property is said to *cyclic*.[5] Since the matrices $\mathbf{A}$ and $\mathbf{A}'$ are similar, and hence have the same characteristic polynomial, these same observations apply also to the case of complete observability as characterized in Theorem 2. In the remainder of this paper, we make the following basic assumptions.

ASSUMPTION 1. The matrix $\mathbf{A}$ is cyclic.

ASSUMPTION 2. The characteristic polynomial $c(s)$ of $\mathbf{A}$ has a factorization into irreducible (real) divisors in the form

$$c(s) = p_1(s)^{k_1}p_2(s)^{k_2} \cdots p_m(s)^{k_m}, \tag{3}$$

where the polynomials $p_i(s)$, $i = 1, 2, \cdots, m$, are mutually distinct and, being irreducible over the field of real numbers, are either linear or quadratic.

As a result of Assumption 1, the following well-known properties, which will be useful in the sequel, hold (for a proof, see [8]).

THEOREM 3. *Suppose* $\mathbf{A}$ *is a real,* $n \times n$ *cyclic matrix,* S *is an* $\mathbf{A}$*-invariant subspace of* E^n*, and* $p(s)$ *is its minimal* $\mathbf{A}$*-polynomial. Then* S *is* $\mathbf{A}$*-cyclic, the degree of* $p(s)$ *is equal to the dimension of* S*, and* S *is the null space of* $p(\mathbf{A})$.

We are now in a position to state our main result on controllability.

THEOREM 4. *Suppose that* $c(s)$ *has the factorization* (3) *and define polynomials* $q_i(s)$ *by*

$$q_i(s)p_i(s) = c(s), \qquad i = 1, 2, \cdots, m. \tag{4}$$

Then, the set $F(\mathbf{A})$ *of all vectors* $\mathbf{f}$ *for which* (1) *is not completely controllable is the union of all the null spaces* T_i *of the matrices* $q_i(\mathbf{A})$*. Moreover,* T_i *has dimension* $n - 1$ *or* $n - 2$ *according as* $p_i(s)$ *is linear or quadratic.*

Proof. Suppose $\mathbf{f}$ is an n-vector, and for some i, $q_i(\mathbf{A})\mathbf{f} = \mathbf{0}$. The degree of $q_i(s)$ is less than n, so $\mathbf{f} \in F(\mathbf{A})$.

If $\mathbf{f} \in F(\mathbf{A})$, then there is a proper divisor $p(s)$ of $c(s)$ such that $p(\mathbf{A})\mathbf{f} = \mathbf{0}$. Then $p(s)$ divides $q_i(s)$ for some i, so $\mathbf{f} \in T_i$.

Finally, by Theorem 3, each subspace T_i has dimension equal to the degree of its respective minimal $\mathbf{A}$-polynomial $q_i(s)$, which is $n - 1$ or $n - 2$ according as $p_i(s)$ is linear or quadratic.

4. Characterization of the set $H(\mathbf{A})$. Since a matrix and its transpose are similar, and therefore have the same characteristic polynomial, all the results of §3 apply also to the case of complete observability if $\mathbf{A}$ is replaced by $\mathbf{A}'$. Thus our main result on observability may be stated at once.

THEOREM 5. *Suppose that* $c(s)$ *has the factorization* (3) *and define polynomials* $q_i(s)$, $i = 1, 2, \cdots, m$, *by* $q_i(s)p_i(s) = c(s)$. *Then the set* $H(\mathbf{A})$ *of*

[5] A matrix having this property is sometimes referred to as a "nonderogatory" matrix. A test to determine whether or not a matrix is cyclic is described in the Appendix.

all vectors $\mathbf{h}$ *for which* (1) *is not completely observable is the union of all the null spaces* U_i *of the matrices* $q_i(\mathbf{A}')$. *Moreover, the space* U_i *has dimension* $n-1$ *or* $n-2$ *according as* $p_i(s)$ *is linear or quadratic.*

5. A duality theorem. It is possible, by means of the theorem proved below, to characterize $F(\mathbf{A})$ in terms of the polynomials $p_i(s)$ rather than the polynomials $q_i(s)$ of Theorem 4. If n is larger than four, the $p_i(s)$ are of smaller degree than the $q_i(s)$, thus indicating the desirability of such a characterization. The theorem establishes a duality between the $\mathbf{A}$-invariant spaces and those spaces invariant with respect to the transpose of $\mathbf{A}$.

THEOREM 6. *Suppose that* $p(s)$ *is a divisor of* $c(s)$. *Let* S *be the null space of* $p(\mathbf{A})$. *If* $S^{\perp}$ *is the orthogonal complement of* S, *then* $S^{\perp}$ *is the null space of* $q(\mathbf{A}')$, *where* $q(s)p(s) = c(s)$.

Proof. Since $p(\mathbf{A}') = [p(\mathbf{A})]'$, $S^{\perp}$ is the range of $p(\mathbf{A}')$. Thus if $\mathbf{x} \in S^{\perp}$, there is an n-vector $\mathbf{y}$ such that $p(\mathbf{A}')\mathbf{y} = \mathbf{x}$, so that $q(\mathbf{A}')\mathbf{x} = q(\mathbf{A}')p(\mathbf{A}')\mathbf{y} = c(\mathbf{A}')\mathbf{y} = \mathbf{0}$. Thus $S^{\perp}$ is a subspace of the null space of $q(\mathbf{A}')$.

If k is the degree of $p(s)$, then, by Theorem 3, the dimension of S is k, so the dimension of $S^{\perp}$ is $n-k$. The null space of $q(\mathbf{A}')$, by Theorem 3, has minimal polynomial $q(s)$, and consequently its dimension is equal to the degree of $q(s)$, which is $n-k$. Thus $S^{\perp}$ is the null space of $q(\mathbf{A}')$.

Theorems 4, 5 and 6 can be combined to obtain a characterization of the sets $F(\mathbf{A})$ and $H(\mathbf{A})$ in terms of the null spaces $S_i^{\perp}$ of the matrices $p_i(\mathbf{A}')$, and the null spaces S_i of the matrices $p_i(\mathbf{A})$. By this means we obtain the following theorem.

THEOREM 7. *The set* $F(\mathbf{A})$ *is the set of all vectors* $\mathbf{f}$ *which lie orthogonal to one of the spaces* $S_i^{\perp}$. *The set* $H(\mathbf{A})$ *is the set of all vectors* $\mathbf{h}$ *which lie orthogonal to one of the spaces* S_i.

Appendix. A test for cyclic matrices. It was observed in [2] that the linear dynamical system (1) is always uncontrollable and unobservable, irrespective of the choice of the vectors $\mathbf{f}$ and $\mathbf{h}$, if and only if the matrix $\mathbf{A}$ is *not* cyclic;[6] that is, if and only if $\mathbf{A}$ satisfies a polynomial equation of degree *less* than n. It may be shown that this latter condition implies the existence of an irreducible factor $p(s)$ of the characteristic polynomial $c(s)$ of $\mathbf{A}$ such that the dimension of the null space of $p(\mathbf{A})$ is greater than the degree of the polynomial $p(s)$. This cannot happen unless $p(s)$ is a factor of $c(s)$ with multiplicity greater than one.

Thus, in order to determine whether or not $\mathbf{A}$ is cyclic, one need only check the ranks of the matrices $p(\mathbf{A})$, where $p(s)$ is a *repeated* irreducible factor (linear or quadratic) of the characteristic equation $c(s)$ of $\mathbf{A}$. This procedure appears to be simpler than finding the minimal polynomial of $\mathbf{A}$, or than

[6] A matrix which is not cyclic is sometimes called a "derogatory" matrix.

reducing **A** to one of the various canonical forms which are similarity invariants.

Acknowledgment. The authors wish to acknowledge the many helpful suggestions furnished by the referees.

REFERENCES

[1] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat · Mexicana (2), 5 (1960), pp. 102–199.

[2] C. D. JOHNSON, *Invariant hyperplanes for linear dynamical systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 113–116.

[3] R. E. KALMAN, *On the general theory of control systems*, Proc. 1st International Congress on Automatic Control, vol. 1, Butterworths, London, 1961, pp. 481–492.

[4] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.

[5] E. KREINDLER AND P. E. SARACHIK, *On the concepts of controllability and observability in linear systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 129–136.

[6] E. G. GILBERT, *Controllability and observability in multivariable control systems*, this Journal, 1 (1963), pp. 128–151.

[7] H. HERMES, *Controllability and the singular problem*, this Journal, 2 (1964), pp. 241–260.

[8] F. R. GANTMACHER, *The Theory of Matrices*, vol. 1, Chelsea, New York, 1959.

[9] G. BIRKHOFF AND S. MACLANE, *A Survey of Modern Algebra*, rev. ed., Macmillan, New York, 1953.

[10] R. E. KALMAN, *Optimal nonlinear control of saturating systems by intermittent action*, IRE Wascon Convention Record, vol. 1, Part 4, 1957, pp. 120–135.

[11] R. V. GAMKRELIDZE, *On the theory of optimal processes in linear systems*, Dokl. Akad. Nauk SSSR (N.S.), 116 (1957), pp. 9–11.

[12] ———, *The theory of time optimal processes in linear systems*, Izv. Akad. Nauk SSSR Ser. Mat., 22 (1958), pp. 449–474.

[13] R. E. KALMAN, *Algebraic theory of linear systems*, Arch. Automatyki i Telemechaniki (Warsaw), 11 (1966), pp. 119–126. Also, Proc. Third Allerton Conference on Circuit and Systems Theory, University of Illinois, Urbana, 1965, pp. 563–577.

[14] ———, *Algebraic aspects of the theory of dynamical systems*, Proc. Conference on Differential Equations and Dynamical Systems, J. K. Hale and J. P. LaSalle, eds., Academic Press, New York, 1967, pp. 133–146.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

L_2-STABILITY OF TIME-VARYING SYSTEMS—CONSTRUCTION OF MULTIPLIERS WITH PRESCRIBED PHASE CHARACTERISTICS*

M. I. FREEDMAN†

Abstract. A system consisting of a linear element $G(s)$ and a time-varying gain $n(\cdot)$ is considered. It is assumed that this system is stable for all constant gains in the sector $[0, k]$ (i.e., for $n(t) \equiv l$, $0 \leqq l \leqq k$). It is then shown that the system is stable for all $n(\cdot)$ in that sector satisfying

$$\left| \frac{1}{t} \int_0^t \left| \frac{d}{d\tau} \left[\log \left(\frac{n(\tau)}{1 - n(\tau)/k} \right) \right] \pm 2\sigma \right| d\tau - 2\sigma \right| \leqq \frac{K}{t}$$

for some $K > 0$ and all $t > 0$.

Here σ is a constant determined by an equation involving (inversely) the derivative of the phase function $\arg \{G(i\omega) + 1/k\}$. The proof follows what are by now established lines in employing a "multiplier" operator. However, a method is used to eliminate any "multiplier" dependence from the final results, so that these results are explicit and geometric.

1. Introduction. Many of the early frequency domain stability criteria afforded simple geometric interpretations. These criteria assured stability on the basis of the plot of a Nyquist curve and its position with respect to some other line or geometric figure. (Consider for instance the Popov criterion or the circle criterion.)

Some of the more recent work has resulted in theorems of a less geometric character. Brockett and Forys [1] considered a feedback system with time-invariant element $G(s)$ and time-varying element $n(\cdot)$, $0 \leqq n(\cdot) \leqq k$ (see Fig. 1) and concluded that if there exists some $Z(s)$ of the form $Z(s) = \sum_{i=1}^{n} \alpha_i/(s + \beta_i)$ with $\alpha_i \geqq 0$ and $\beta_i \geqq \sigma$ and with $Z(s)^{\pm 1}[G(s) + 1/k]$ positive real, then the system would be stable for all $n(\cdot)$ satisfying $\dot{n}(t)/n(t) \leqq 2\sigma(1 - n(t)/k)$ for $t \geqq 0$.

In the context of integral equations, a result of Falb and Zames [2] showed that for a system consisting of a convolution operator $\mathbf{G}$ and a monotone nonlinearity $f(\cdot)$ satisfying $0 \leqq \sigma f(\sigma) \leqq k\sigma^2$, stability can be proven if there exists an operator $\mathbf{Z}$: $L_2[0, \infty) \to L_2[0, \infty)$ defined by

$$(\mathbf{Z}x)(t) = x(t) + \int_0^t z_1(t - \tau)x(\tau)\, d\tau,$$

where

---

* Received by the editors March 29, 1968, and in revised form July 5, 1968.

† National Aeronautics and Space Administration, Electronics Research Center, Cambridge, Massachusetts 02139.

$$z_1(\cdot) \in L_1(0, \infty), \quad \int_0^\infty |z_1(t)|\, dt < 1$$

and

$$\mathrm{Re}\left\{[Z(i\omega) + \alpha i\omega]\left[G(i\omega) + \frac{1}{k}\right]\right\} \geqq \delta > 0$$

for some $\alpha > 0$ and all real ω.

Both of the preceding results depend on the existence of a "multiplier" operator which, when combined with $[G(s) + 1/k]$, yields a positive operator. This idea, expressed as a factorization property of $G(s)$, first appeared in Zames [3a]. The usefulness of the multiplier approach as portrayed in [1], [2] and [3a] is limited by the absence of any explicit method for finding suitable multipliers.

In Freedman and Zames [4] the stability of a system involving a linear time-invariant element $G(s)$ and a time-varying gain $n(\cdot)$ was considered (as in Fig. 1). The method of proof in [4] involved introduction of a multiplier much as in [2] above. However, a constructive process was developed which allowed for removal of this multiplier from the final results, so that these results were geometric in nature.

It is this author's contention that [4] contains the core of a procedure which can be used to initiate a program aimed at returning more closely to the geometric character of the earlier results. More explicitly it is felt that the ideas in [4] can be utilized to remove the multiplier from many of the more recent stability results, thus yielding criteria depending only on $G(s)$ and its properties.

For the purpose of this paper the result of Brockett and Forys [1] mentioned above was considered and a criterion was developed which is free from dependence on any multiplier.

To describe this criterion more fully consider the feedback system represented by Fig. 1, and assume it is stable for all constant gains in sector

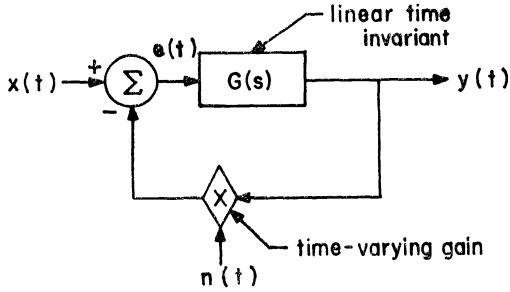


FIG. 1. *A feedback system*

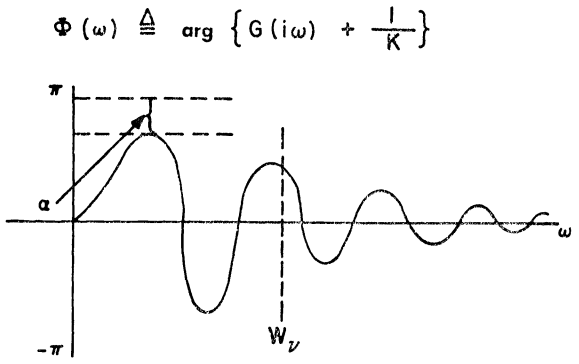


FIG. 2. *Plot of the phase function* $\Phi(\omega)$

$[0, k]$. Denote by $\Phi(\omega)$ the phase function $\arg\{G(i\omega) + 1/k\}$. Then by the criterion to be presented here the system will be stable if there is a $K > 0$ such that

$$\left| \frac{1}{t} \int_0^t \left| \frac{\dot{n}(\tau)}{n(\tau)\,(1 - n(\tau)/k)} \pm 2\sigma \right| d\tau - 2\sigma \right| \leqq \frac{K}{t}$$

for all $t > 0$ (and so more weakly if $\dot{n}(t)/n(t) \leqq 2\sigma(1 - n(t)/k)$ or $\dot{n}(t)/n(t) \geqq -2\sigma(1 - n(t)/k)$), where σ is determined by an equation involving (see Fig. 2):

1. the magnitude of the closest approach of $\Phi(\omega)$ to $\pm\pi$;

2. a "cutoff frequency W_ν", i.e., a frequency after which the values of $\Phi(\omega)$ are of no importance in this theory;

3. the magnitude of the square integral of the derivative of $\Phi(\omega)$ over $[-W_\nu, W_\nu]$.

One sees therefore that not only does the geometric set

$$N = \{G(i\omega)| \; -\infty < \omega < \infty\}$$

come into play here, but also in some sense the "angular" rate at which this set is traced out as ω varies (as represented by $\Phi'(\omega)$). In fact it will appear that such rates are closely related to the existence and properties of certain multipliers.

The crucial lemma on which the theory presented here rests is Lemma 1 in §4, which shows that a multiplier operator with any prescribed phase function can be constructed provided only that its phase function and its derivative satisfy suitable integrability conditions. This result appeared in [4, §4, Lemma 2] and the reader is referred to that paper for the proof. However, a short sketch of that proof is included here.

2. The main problem and its solution.

DEFINITION. Let $L_p[0, \infty)$, where $p = 1, 2, \cdots, \infty$, be the linear space of real-valued measurable functions $x(\cdot)$ on $[0, \infty)$ with the property that

$$\int_0^\infty |x(t)|^p \, dt < \infty \quad \text{if} \quad 1 \leqq p < \infty,$$

or $x(\cdot)$ is essentially bounded if $p = \infty$. Let $L_2[0, \infty)$ be normed with the norm

$$\| x(\cdot) \| = \left\{ \int_0^\infty |x(t)|^2 \, dt \right\}^{1/2}.$$

The spaces $L_p(-\infty, \infty)$ on the interval $(-\infty, \infty)$ are similarly defined.

The definition of the extended space L_{2e} is introduced next. (For a more complete discussion of such spaces, see [3b].)

DEFINITION. Let L_{2e} be the space of those real-valued measurable functions $x(\cdot)$ on $[0, \infty)$ satisfying

$$\int_0^T |x(t)|^2 \, dt < \infty \quad \text{for all} \quad T \geqq 0.$$

2.1. Feedback equations and stability. The feedback system of Fig. 1 will be represented for all $t \geqq 0$ by the integral equation

$$e(t) = x(t) - n(t) \left[\int_0^t e(\tau) g(t - \tau) \, d\tau \right] \tag{1}$$

or, alternatively, the pair:

$$\begin{aligned} e(t) &= x(t) - n(t)\, y(t), \\ y(t) &= \int_0^t e(\tau) g(t - \tau) \, d\tau, \end{aligned} \tag{2}$$

in which the following assumptions are made:

ASSUMPTION 1. $x(\cdot)$ is in $L_2[0, \infty)$. (The function $x(\cdot)$ represents the combined effects of an input and of possible nonzero initial conditions.)

ASSUMPTION 2. $g(\cdot)$ is in $L_1[0, \infty)$.

ASSUMPTION 3. $n(\cdot)$ is a real-valued function, absolutely continuous on $[0, \infty)$. (Since $n(\cdot)$ is absolutely continuous, its derivative $\dot{n}(\cdot)$ exists almost everywhere, and

$$n(b) - n(a) = \int_a^t \dot{n}(t) \, dt$$

for any nonnegative real a and b (see Hobson [5, §406, pp. 592–593]).)

ASSUMPTION 4. $e(\cdot)$ (and also $y(\cdot)$) is in L_{2e} (i.e., existence of solutions in L_{2e} for $L_2[0, \infty)$-inputs is being assumed[1]).

DEFINITION. Feedback system (1) will be termed *L_2-stable* if for any pair $(x(\cdot), e(\cdot))$ for which (1) (or (2)) and the related Assumptions 1–4 hold, then $e(\cdot)$ is in $L_2[0, \infty)$, with $\| e(\cdot)\| \leqq \text{const.} \| x(\cdot)\|$.

This notion of stability is natural in the setting of integral equations. In the context of differential equations it implies asymptotic stability ($\lim_{t\to\infty} y(t) = 0$) and with additional minor assumptions can also be used to show bounded-input, bounded-output stability.

2.2. The main stability theorem. This section contains the main stability results. A few definitions and remarks will provide the setting.

DEFINITION. For any $k > 0$, let

$$W_k = \left\{ f(\cdot) \in L_1[0, \infty) \mid F(i\omega) + \frac{1}{l} \neq 0 \right.$$
$$\left. \text{for } -\infty < \omega < \infty \text{ and all } l,\ 0 < l < k \right\},$$

where $F(s)$, the Laplace transform of $f(\cdot)$, is the complex-valued function with domain $\{s \mid \text{Re}\{s\} \geqq 0\}$ defined, as usual, by the integral

$$F(s) = \int_0^\infty e^{-st} f(t)\, dt;$$

i.e., $f(\cdot) \in W_k$ if and only if the set $\{F(i\omega) \mid \omega \in (-\infty, \infty)\}$ does not cut the negative real axis from $-\infty$ up to and including the point $-1/k$.

Remark 1. The statement $g(\cdot) \in W_k$ may be interpreted via the "principle of the argument" to be equivalent to the statement: The equation $G(s) = -1/k$ has no complex roots in $\text{Re}\{s\} \geqq 0$. Also the classical Nyquist criterion assures that $g(\cdot)$ in W_k is a necessary and sufficient condition for feedback system (1) to be L_2-stable for all constant gains between 0 and k, i.e., for $n(t) \equiv l$, where l is a constant $0 \leqq l \leqq k$.

Remark 2 *and some special notation.* Given $g(\cdot) \in W_k$, define the phase function:

$$\text{(a)} \qquad \Phi(\omega) \triangleq \arg\{G(i\omega) + 1/k\}.$$

Then since $G(i\omega) + 1/k$ does not cut the negative real axis as ω passes from $-\infty$ to $+\infty$, it follows that $\Phi(\omega)$ is uniquely defined for all ω and takes values only in $(-\pi, \pi)$. Further, since $g(\cdot) \in L_1[0, \infty)$ the Riemann-

[1] The problem of existence of L_{2e} solutions will not be discussed in this paper, except to say that such questions may be settled favorably by various minor additional hypotheses.

Lebesgue lemma assures that $\lim_{|\omega|\to\infty} G(i\omega) = 0$ and so

$$\lim_{|\omega|\to\infty} \arg \{G(i\omega) + 1/k\} = 0$$

also. Therefore a simple continuity argument shows that the function $G(i\omega) + 1/k$ must have a "closest angular approach" α to the negative real axis; that is, letting:

(b) $\alpha \triangleq \min (\pi - |\Phi(\omega)|)$,

then $\alpha > 0$.

Next let:

(c) $\nu = (\pi - \alpha)/3$

and define the cutoff frequency:

(d) $W_\nu \triangleq \min \{W/|\Phi(\omega)| \leqq \nu \text{ for } |\omega| \geqq W\}$.

Here the existence of W_ν is assured by an argument employing the continuity of $\Phi(\omega)$ and the fact that $\lim_{|\omega|\to\infty} \Phi(\omega) = 0$.

The importance of W_ν is that it represents a frequency value beyond which information about the phase function $\Phi(\omega)$ need not be utilized in the theory to follow.

In the remainder of this paper, the notation introduced in (a) through (d) above will be used freely.

The main result may now be stated.

THEOREM 1. *Suppose* (1) (*or equivalently* (2)) *and the related Assumptions* 1–4 *hold for a pair* $(x(\cdot), e(\cdot))$. *Let* $k > 0$ *be given and assume that*:

(i) $g(\cdot) \in W_k$ (*i.e., the system is* L_2*-stable for* $n(t) \equiv l$ *for any* l, $0 \leqq l \leqq k$),

(ii) $0 < \inf n(t) \leqq n(t) \leqq \sup \leqq n(t) < k$.

Then:

(a) *if* $|\Phi(\omega)| \leqq \pi/2$ *for* $-\infty < \omega < \infty$, *the system is* L_2*-stable by a simple positivity argument.* (*Note that* $|\Phi(\omega)| \leqq \pi/2$ *for* $-\infty < \omega < \infty$ *implies* $\operatorname{Re}\{G(i\omega)\} \geqq -1/k$ *for* $-\infty < \omega < \infty$ *and the desired result follows from the basic Popov theorem.*)

On the other hand:

(b) *if* $\Phi(\omega) > \pi/2$ *for some* ω, *the cutoff frequency* W_ν *is strictly positive and one may define*

$$\sigma_* \triangleq \frac{(3\pi\alpha/16)^2}{\int_{-W_\nu}^{+W_\nu} |\Phi'(\omega)|^2 \, d\omega},$$

where $\Phi'(\omega)$ *denotes the derivative of the phase function* $\Phi(\omega)$.

Suppose then that there exists a constant $K > 0$ *and a constant* σ *in* $(0, \sigma_*)$ *such that*

$$\left| \frac{1}{t} \int_0^t \left| \frac{\dot{n}(\tau)}{n(\tau)(1 - n(\tau)/k)} - 2\sigma \right| d\tau - 2\sigma \right| \leqq \frac{K}{t} \tag{3}$$

for all $t > 0$ *or else that*

$$\left| \frac{1}{t} \int_0^t \left| \frac{\dot{n}(\tau)}{n(\tau)(1 - n(\tau)/k)} + 2\sigma \right| d\tau - 2\sigma \right| \leqq \frac{K}{t} \tag{4}$$

for all $t > 0$. *Then* $e(\cdot)$ *is in* $L_2[0, \infty)$ *and, in fact,* $\| e(\cdot) \| \leqq$ const. $\| x(\cdot) \|$. *Therefore system* (1) *is* L_2*-stable.*

Remark 3. Theorem 1 is an immediate consequence of Theorem 2 in §3, Lemma 3 and Remark 4 in §4, and Corollary 3 in §5.

COROLLARY 1. *Under the assumptions and notation of Theorem* 1 *a sufficient condition on* $n(\cdot)$ *for* (3) *to hold, and hence for* L_2*-stability, is that*

$$\frac{\dot{n}(t)}{n(t)} \leqq 2\sigma \left(1 - \frac{n(t)}{k}\right)$$

for all $t > 0$. *Similarly, the inequality*

$$\frac{\dot{n}(t)}{n(t)} \geqq -2\sigma \left(1 - \frac{n(t)}{k}\right)$$

is a sufficient condition for (4) *to hold and so also to ensure* L_2*-stability.* (See R. W. Brockett and L. J. Forys [1] for a multiplier result along these lines.)

Proof of Corollary 1. The proof is immediate.

COROLLARY 2. *Let the hypotheses and notation of Theorem* 1 *hold except in case* (b) *redefine* σ_* *as follows*:

$$\sigma_* \triangleq \left(\frac{3\pi\alpha}{16\sqrt{2W_\nu}}\right)^2 \frac{1}{\max\limits_{-W_\nu \leqq \omega \leqq W_\nu} |\Phi'(\omega)|^2}.$$

Then with this definition of σ_* *the conclusions of Theorem* 1 *and also of Corollary* 1 *remain valid, i.e., system* (1) *is* L_2*-stable.*

Proof of Corollary 2. This choice of σ_* is less than or equal to the σ_* of the original statement.

3. A theorem on multipliers. The following theorem is essentially proved in Zames and Freedman [4] and the reader is recommended to that paper for the proof. The existence of a multiplier operator $\mathbf{Z}$ satisfying certain properties with respect to $\mathbf{G} + 1/k$ is hypothesized and a stability conclusion is then drawn for all time-varying gains $n(\cdot)$ suitably restricted. The

result in the form presented below is rather similar to a result in Brockett and Forys [1]; however, it differs in one sense in that the form of the multiplier $\mathbf{Z}$ remains unspecified in this paper. Moreover, the key fact to keep in mind here is that this theorem represents an intermediate stage in the developments leading to Theorem 1 stated in the previous section, and Theorem 1 contains no explicit mention of the multiplier.

For what follows the definition of some special operator spaces (actually they are Banach algebras) will be of use.

DEFINITION. Let $\mathcal{L}_{\sigma_0}$ be the class of operators $\mathbf{H}: L_{2e} \to L_{2e}$ satisfying

$$(\mathbf{H}x)(t) = h_0\, x(t) + \int_0^t x(\tau) h_1(t - \tau)\, d\tau$$

(for all $x(\cdot) \in L_{2e}[0, \infty)$ and all $t \geqq 0$), where h_0 is a real constant and $h_1(\cdot)$ is a real-valued measurable function on $[0, \infty)$ satisfying $h_1(t) \exp(\sigma_0 t) \in L_1[0, \infty)$.

For $\mathbf{H} \in \mathcal{L}_{\sigma_0}$ the Laplace transform of $\mathbf{H}$ is given by

$$H(s) = h_0 + \int_0^\infty h_1(t) \exp(-st)\, dt$$

(for all complex s with $\operatorname{Re}\{s\} \geqq -\sigma_0$).

THEOREM 2. *Let* (1) (*or equivalently* (2)) *and the related Assumptions* 1–4 *hold for a pair* $(x(\cdot), e(\cdot))$. *Let* $k > 0$ *be given and assume that*

$$0 < \inf n(t) \leqq n(t) \leqq \sup n(t) < k.$$

Further, let the following two assumptions be made:

(i) *There is a constant* $\sigma > 0$ *and an operator* $\mathbf{Z} \in \mathcal{L}_\sigma$ *satisfying, for all real* ω,

(a) $\operatorname{Re}\{Z(i\omega - \sigma)\} \geqq 0$,

(b) $\operatorname{Re}\{Z(i\omega)[G(i\omega) + 1/k]\} \geqq \delta$

for some positive constant δ.

(ii) *The function* $n^*(\cdot)$ *defined by* $n^*(t) \triangleq n(t)(1 - n(t)/k)^{-1}$ *for all* $t \geqq 0$, *or the reciprocal of this function, may be factored into a product* $n_1(\cdot) \cdot n_2(\cdot)$ *of two absolutely continuous functions on* $[0, \infty)$, *where* $n_1(t) \exp(-2\sigma t)$ *is monotone nonincreasing while* $n_2(t)$ *is monotone nondecreasing in* t, *and* $0 < \inf n_1(t) \leqq \sup n_2(t) < \infty$ (*and so also* $0 < \inf n_2(t) \leqq \sup n_2(t) < \infty$).

Then the system (1) *is* L_2*-stable, i.e.,* $e(\cdot)$ *is in* $L_2[0, \infty)$ *and* $\| e(\cdot) \| \leqq \text{const.} \| x(\cdot) \|$.

Proof of Theorem 2. See Zames and Freedman [4, Lemma 1]. The factorization property (imposed on the time-varying gain) hypothesized in [4] is slightly different from the one made in this present paper, but the proof

under the conditions considered here follows the same general arguments and is, in fact, a bit easier. For that reason the reader is referred to [4].

4. Multipliers with prescribed phase characteristics. In this section three lemmas will be presented which together settle the question of the existence of a suitable multiplier for Theorem 2. Such a multiplier will be seen always to exist and, in fact, an associated range of σ values (as required for the hypothesis of Theorem 2) may be obtained directly from data concerning the phase of $G(i\omega) + 1/k$ without recourse to construction of $\mathbf{Z}$ in any given case.

The foundation of the results to follow in this section is Lemma 1 below, which assures the existence of an operator $\mathbf{Z}$ in $\mathfrak{L}_\sigma$ with any prescribed phase function $\Phi_0(\omega) = \arg\{Z(i\omega)\}$, provided only that the function $\Phi_0(\omega)$ and its derivative $\Phi_0'(\omega)$ satisfy certain integrability conditions.

LEMMA 1 (Operators with prescribed phase). *If*:

(i) $\Phi_0(\omega)$ *is a real-valued continuous a. e. differentiable odd function of* ω *for* ω *in* $(-\infty, \infty)$,

(ii) $\Phi_0(\omega)$ *and* $\Phi_0'(\omega)$ *are in* $L_2(-\infty, \infty)$,

then:

(a) *there is a function* $\lambda(\cdot)$ *in* $L_1(-\infty, \infty)$ *with* $\lambda(t) = 0$ *for* $t < 0$ *and with a Laplace transform* $\Lambda(s)$ *satisfying* $\operatorname{Im}\{\Lambda(i\omega)\} = \Phi_0(\omega)$;

(b) *there is a* $y(\cdot)$ *in* $L_1(-\infty, \infty)$ *with* $y(t) = 0$ *for* $t < 0$, *and with a Laplace transform* $Y(s)$ *satisfying* $1 + Y(s) = \exp[\Lambda(s)]$ *for* $\operatorname{Re}\{s\} \geqq 0$;

(c) *if* $-\pi < \Phi_0(\omega) < \pi$, *there is a* $y(\cdot) \in L_1(-\infty, \infty)$ *with* $y(t) = 0$ *for* $t < 0$, $1 + Y(s) \neq 0$ *in* $\operatorname{Re}\{s\} \geqq 0$ (*so* $1 + Y(s)$ *is minimum phase*) *and* $\arg\{1 + Y(i\omega)\} = \Phi_0(\omega)$.

Outline of proof. For the complete proof of this lemma the reader is once again referred to Zames and Freedman [4]. However a brief outline of the main ideas is presented here.

(i) Let $\phi_0(t)$ denote the inverse limit-in-the-mean Fourier transform of $i\Phi_0(\omega)$ and define

$$\lambda(t) = \begin{cases} 2\phi_0(t) & \text{for } t \geqq 0, \\ 0 & \text{for } t < 0. \end{cases}$$

Then it may be deduced from (i) and (ii) of Lemma 1 that $\lambda(\cdot)$ is in $L_1(-\infty, \infty)$ and that the Laplace transform of $\lambda(\cdot)$ denoted $\Lambda(s)$ satisfies $\operatorname{Im}\{\Lambda(i\omega)\} = \Phi_0(\omega)$, and so (a) above holds.

(ii) Next, for each $n = 1, 2, \cdots$ let

$$y_n(t) = \lambda(t) + \frac{(\lambda * \lambda)(t)}{2!} + \cdots + \frac{\overbrace{(\lambda * \cdots * \lambda)}^{n}(t)}{n!},$$

where $*$ represents convolution; that is,

$$(x_1 * x_2)(t) \triangleq \int_{-\infty}^{\infty} x_1(t-\tau)x_2(\tau)\,d\tau.$$

Then, for each n, $y_n(\cdot)$ is in $L_1[0, \infty)$ and the Laplace transform of $y_n(\cdot)$, $Y_n(s)$, satisfies

$$Y_n(s) = \Lambda(s) + \frac{\Lambda^2(s)}{2!} + \cdots + \frac{\Lambda^n(s)}{n!}.$$

(iii) Finally it may be shown that $y_n(\cdot)$ converges in L_1-norm to a function $y(\cdot)$ in $L_1[0, \infty)$ and the Laplace transform of this function $y(\cdot)$, denoted $Y(s)$, must satisfy

$$Y(s) = \exp[\Lambda(s)] - 1 \quad \text{for} \quad \operatorname{Re}\{s\} \geqq 0.$$

From this fact (b) and (c) of Lemma 1 follow.

In applying Lemma 1 to the problem of constructing a multiplier $\mathbf{Z}$ which will meet the hypotheses of Theorem 2, it is clear that the larger σ may be chosen (with a corresponding $\mathbf{Z} \in \mathfrak{L}_\sigma$ assured), the less restrictive will be the conditions on $n(\cdot)$. A measure of how large a σ is possible will be developed via the following lemma. This lemma in its face concerns the rate of convergence of the values of the harmonic function $\operatorname{Im}\{\Lambda(s)\}$ defined on a half-plane $\operatorname{Re}\{s\} \geqq 0$ as s approaches $\operatorname{Re}\{s\} = 0$ along ordinate lines.

LEMMA 2. *Under the same notation and assumptions as in the previous theorem, it follows that for any* $\sigma > 0$,

$$\sup_{-\infty<\omega<\infty} |\operatorname{Im}\{\Lambda(i\omega+\sigma)\} - \Phi_0(\omega)| \leqq \frac{16\sqrt{\sigma}}{3\pi}\left[\int_{-\infty}^{\infty} |\Phi_0'(\omega)|^2\,d\omega\right]^{1/2},$$

and so if $-\pi < \Phi_0(\omega) < \pi$ *for all* ω *and if the principal value of* $\arg\{\cdot\}$ *is taken, then*

$$\sup_{-\infty<\omega<\infty} |\arg\{1 + Y(i\omega+\sigma)\} - \Phi_0(\omega)| \leqq \frac{16\sqrt{\sigma}}{3\pi}\left[\int_{-\infty}^{\infty} |\Phi_0'(\omega)|^2\,d\omega\right]^{1/2}.$$

Proof of Lemma 2. (The proof presented here follows closely the lines of the singular integral theory of Titchmarsh [6, §116, pp. 28–29].) For all s with $\operatorname{Re}\{s\} \geqq 0$ the Laplace transform

$$\Lambda(s) \triangleq \int_0^{\infty} \exp(-st)\lambda(t)\,dt. \tag{5}$$

Now let, for the moment,

$$q(t) = \begin{cases} \exp(-st) & \text{for} \quad t \geqq 0, \\ 0 & \text{for} \quad t < 0. \end{cases}$$

So, for any s with $\operatorname{Re}\{s\} > 0$ the Fourier transform of $q(\cdot)$ is $1/(i\omega + s)$. Applying Parseval's theorem to (5) results in

$$\Lambda(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{i\omega + s}\,\overline{\Lambda(i\omega)}\,d\omega. \tag{6}$$

The Parseval theorem applies as $\lambda(\cdot)$ is also in $L_2[0, \infty)$. However, applying Cauchy's theorem to the analytic function $\Lambda(s)$ yields

$$0 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{i\omega + s}\,\Lambda(i\omega)\,d\omega \quad \text{for} \quad \operatorname{Re}\{s\} > 0 \tag{7}$$

since $\Lambda(z) \to 0$ as $|z| \to \infty$ in $\operatorname{Re}\{z\} \geqq 0$ implies that

$$\int_{D_R} \frac{1}{z + s}\,\Lambda(z)\,dz \to 0 \quad \text{as} \quad R \to \infty,$$

where D_R is the semicircle $\{z \mid |z| = R, \operatorname{Re}\{z\} \geqq 0\}$.

Now

$$\Phi_0(\omega) = \operatorname{Im}\{\Lambda(i\omega)\} = (\Lambda(i\omega) - \overline{\Lambda(i\omega)})/2i$$

as seen in the previous lemma, and so, subtracting (6) from (7) gives

$$\Lambda(s) = -\frac{i}{\pi}\int_{-\infty}^{\infty} \Phi_0(\omega)\,\frac{1}{s + i\omega}\,d\omega \quad \text{for} \quad \operatorname{Re}\{s\} > 0. \tag{8}$$

Next, for convenience of notation, letting $s = \sigma + iy$, $\bar{s} = \sigma - iy$ and $\operatorname{Im}\{\Lambda(s)\} = V(\sigma, y)$, one obtains

$$V(\sigma, y) = \frac{1}{2i}\left[-\frac{i}{\pi}\int_{-\infty}^{\infty} \frac{\Phi_0(\omega)}{s + i\omega} + \frac{\Phi_0(\omega)}{\bar{s} - i\omega}\right] d\omega \quad \text{for} \quad \operatorname{Re}\{s\} > 0$$

and so

$$V(\sigma, y) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\sigma}{\sigma^2 + (y + \omega)^2}\,\Phi_0(\omega)\,d\omega \quad \text{for} \quad \sigma > 0,$$

or, replacing ω by $-\omega$ and noting that $\Phi_0(\omega) = -\Phi_0(-\omega)$, we have

$$V(\sigma, y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\sigma}{\sigma^2 + (y - \omega)^2}\,\Phi_0(\omega)\,d\omega \quad \text{for} \quad \sigma > 0. \tag{9}$$

Next note that for any $y \in (-\infty, \infty)$,

$$\frac{1}{\pi}\int_{-y}^{\infty} \frac{\sigma}{\sigma^2 + (y - \omega)^2}\,d\omega = \frac{1}{\pi}\int_{-\infty}^{y} \frac{\sigma}{\sigma^2 + (y - \omega)^2}\,d\omega = \frac{1}{2}.$$

Therefore,

$$\left|\frac{1}{\pi}\int_{y}^{\infty} \frac{\sigma}{\sigma^2 + (y - \omega)^2}\,\Phi_0(\omega)\,d\omega - \frac{\Phi_0(y)}{2}\right| = \left|\frac{1}{\pi}\int_{y}^{\infty} \frac{\sigma}{\sigma^2 + (y - \omega)^2}\,[\Phi_0(\omega) - \Phi_0(y)]\,d\omega\right|. \tag{10}$$

Now for any $\sigma > 0$ and any ω and y with $\omega \neq y$,

$$\frac{\sigma}{\sigma^2 + (y-\omega)^2} \leqq \frac{1}{\sigma} \quad \text{and} \quad \frac{\sigma}{\sigma^2 + (y-\omega)^2} \leqq \frac{\sigma}{(y-\omega)^2}.$$

Using these relations to bound (10) results in

$$\begin{aligned} &\left| \frac{1}{\pi} \int_y^\infty \frac{\sigma}{\sigma^2 + (y-\omega)^2} [\Phi_0(\omega) - \Phi_0(y)]\, d\omega \right| \\ &\qquad \leqq \frac{1}{\pi\sigma} \int_y^{y+\sigma} |\Phi_0(\omega) - \Phi_0(y)|\, d\omega + \frac{\sigma}{\pi} \int_{-y+\sigma}^\infty \frac{|\Phi_0(\omega) - \Phi_0(y)|}{(y-\omega)^2}\, d\omega. \end{aligned} \tag{11}$$

The first integral on the right-hand side of (11) satisfies the following chain of inequalities:

$$\begin{aligned} &\frac{1}{\pi\sigma} \int_y^{y+\sigma} |\Phi_0(\omega) - \Phi_0(y)|\, d\omega \\ &\qquad \leqq \frac{1}{\pi\sigma} \int_y^{y+\sigma} \int_y^\omega |\Phi_0'(t)|\, dt\, d\omega \\ &\qquad \leqq \frac{1}{\pi\sigma} \int_y^{y+\sigma} \left(\left[\int_y^\omega |\Phi_0'(t)|^2\, dt \right]^{1/2} (\omega - y)^{1/2} \right) d\omega \\ &\qquad \leqq \left(\int_{-\infty}^\infty |\Phi_0'(t)|^2\, dt \right)^{1/2} \frac{1}{\pi\sigma} \cdot \frac{2\sigma^{3/2}}{3} \\ &\qquad = \frac{2}{3\pi} \sqrt{\sigma} \left(\int_{-\infty}^\infty |\Phi_0'(t)|^2\, dt \right)^{1/2}. \end{aligned} \tag{12}$$

The second integral on the right-hand side of (11) also satisfies a chain of inequalities, as follows:

$$\begin{aligned} \frac{\sigma}{\pi} \int_{y+\sigma}^\infty \frac{|\Phi_0(\omega) - \Phi_0(y)|}{(\omega - y)^2}\, d\omega &= \frac{\sigma}{\pi} \int_\sigma^\infty \frac{|\Phi_0(y+t) - \Phi_0(y)|}{t^2}\, dt \\ &\leqq \frac{\sigma}{\pi} \int_\sigma^\infty \frac{1}{t^2} \left[\int_y^{y+t} |\Phi_0'(\xi)|\, d\xi \right] dt \\ &\leqq \left(\int_{-\infty}^\infty |\Phi_0'(\xi)|^2\, d\xi \right)^{1/2} \cdot \frac{\sigma}{\pi} \cdot \int_\sigma^\infty t^{-3/2}\, dt \\ &\leqq \left(\int_{-\infty}^\infty |\Phi_0'(\xi)|^2\, d\xi \right)^{1/2} \cdot \frac{\sigma}{\pi} \cdot \frac{2}{\sqrt{\sigma}} \\ &= \left(\int_{-\infty}^\infty |\Phi_0'(\xi)|^2\, d\xi \right)^{1/2} \cdot \frac{2\sqrt{\sigma}}{\pi}. \end{aligned} \tag{13}$$

Combining (11), (12) and (13) with (10) one finds that

$$\left|\frac{1}{\pi}\int_y^\infty \frac{\sigma}{\sigma^2+(y-\omega)^2}\Phi_0(\omega)\,d\omega - \frac{\Phi_0(y)}{2}\right| \leqq \frac{8\sqrt{\sigma}}{3\pi}\left(\int_{-\infty}^{\infty}|\Phi_0'(\xi)|^2\,d\xi\right)^{1/2}. \tag{14}$$

Now by a completely analogous procedure, one also has

$$\left|\frac{1}{\pi}\int_{-\infty}^y \frac{\sigma}{\sigma^2+(y-\omega)^2}\Phi_0(\omega)\,dw - \frac{\Phi_0(y)}{2}\right| \leqq \frac{8\sqrt{\sigma}}{3\pi}\left(\int_{-\infty}^{\infty}|\Phi_0'(\xi)|^2\,d\xi\right)^{1/2}. \tag{15}$$

Finally (14) and (15) together with (9) yield

$$\sup_{-\infty<y<\infty}|V(\sigma,y)-\Phi_0(y)| \leqq \frac{16\sqrt{\sigma}}{3\pi}\left(\int_{-\infty}^{\infty}|\Phi_0'(\xi)|^2\,d\xi\right)^{1/2}.$$

This completes the proof.

The following lemma completes the "constructive" process for the **Z** hypothesized in Theorem 2 and also gives a range of applicable σ values.

LEMMA 3. *Let the hypotheses and notation of Theorem* 1 *covering* $g(\cdot)$ *hold, i.e.,* $g(\cdot)\in W_k$, $\Phi(\omega)\triangleq \arg\{G(i\omega)+1/k\}$, $\alpha\triangleq\min(\pi-|\Phi(\omega)|)>0$, $\nu\triangleq\pi-\alpha/3$ *and* $W_\nu=\inf\{W_\nu\,|\,|\Phi(\omega)|\leqq\nu$ *for* $|\omega|\geqq W\}$. *Assume* $W_\nu>0$ *and let*

$$\sigma_*\triangleq\left(\frac{3\pi\alpha}{16}\right)^2\cdot\frac{1}{\left[\int_{-W_\nu}^{W_\nu}|\Phi'(\omega)|^2\,d\omega\right]^{1/2}}.$$

Then for any σ *in* $[0,\sigma^*)$ *there is a* $y(\cdot)$ *in* $L_1[0,\infty)$, *with Laplace transform* $Y(s)$, *and a* $\delta>0$, *such that*:

(i) $\operatorname{Re}\{1+Y(i\omega)\}\geqq\delta>0$,

(ii) $\operatorname{Re}\left\{[1+Y(i\omega+\sigma)]\left[G(i\omega)+\frac{1}{k}\right]\right\}\geqq\delta>0.$

Remark 4. If $\mathbf{Z}\in\mathfrak{L}_\sigma$ is defined by

$$(\mathbf{Z}x)(t)=x(t)+\int_0^t y(t-\tau)e^{-\sigma(t-\tau)}x(\tau)\,d\tau,$$

then the following conditions are equivalent to (i) and (ii) above:

(i′) $\operatorname{Re}\{Z(i\omega-\sigma)\}\geqq\delta>0$,

(ii′) $\operatorname{Re}\left\{Z(i\omega)\cdot\left[G(i\omega)+\frac{1}{k}\right]\right\}\geqq\delta>0.$

Hence condition (i) of Theorem 2 is satisfied for all σ in $(0,\sigma_*)$.

The hypothesis $g(\cdot) \in W_k$ assures that $\arg\{G(i\omega) + 1/k\}$ lies in the interval $(-\pi, \pi)$. If conclusions (i) and (ii) of Lemma 3 are to be fulfilled, then it is clearly necessary that $\arg\{1 + Y(i\omega)\}$ lie in $(-\pi/2, \pi/2)$ and also that $\arg\{[1 + Y(i\omega + \sigma)][G(i\omega) + 1/k]\}$ lie in $(-\pi/2, \pi/2)$. That these two conditions are also sufficient for the validity of (i) and (ii) can be seen from the following remark.

Remark 5. If $F(\omega)$ is a continuous complex-valued function on $(-\infty, \infty)$, $F(\omega) \neq 0$, $|\arg\{F(\omega)\}| < \pi/2$ and $\lim_{|\omega|\to\infty} F(\omega)$ exists and is a real constant greater than zero, then there is a constant $\delta > 0$ with the property that $\operatorname{Re}\{F(\omega)\} \geqq \delta$.

Recalling that

$$\arg\{[1 + Y(i\omega + \sigma)][G(i\omega) + 1/k]\} = \arg\{1 + Y(i\omega + \sigma)\} + \arg\{G(i\omega) + 1/k\},$$

an initial attempt at constructing $y(\cdot)$ might be to employ Lemma 1 in order to find a $y(\cdot) \in L_1(0, \infty)$ with

$$\arg\{1 + Y(i\omega)\} = -\tfrac{1}{2}\arg\{G(i\omega) + 1/k\}.$$

With such a choice for $y(\cdot)$, both $\arg\{1 + Y(i\omega)\}$ and $\arg\{1 + Y(i\omega)\} + \arg\{G(i\omega) + 1/k\}$ $(= -\tfrac{1}{2}\arg\{G(i\omega) + 1/k\})$ would lie in $(-\pi/2, \pi/2)$. The construction which will be adopted here must differ from the choice of phase function suggested above for large ω in order to meet the conditions required for application of Lemmas 1 and 2, namely, that the phase function chosen and its derivative must both have finite square integrals.

Proof of Lemma 3. For each $\epsilon > 0$ choose $l_\epsilon(\omega)$ a continuous, a.e., differentiable real-valued function defined on (W_ν, ∞) and satisfying:

$$\text{(a)} \qquad l_\epsilon(W_\nu) = -\frac{\Phi(W_\nu)}{2} = -\frac{\arg\{G(i\omega) + 1/k\}}{2},$$

$$\text{(b)} \qquad |l_\epsilon(\omega)| \leqq \frac{\nu}{2} \quad \text{for all} \quad \omega \in (W_\nu, \infty),$$

$$\text{(c)} \qquad \int_{W_\nu}^{\infty} |l_\epsilon(\omega)|^2\, d\omega < \infty,$$

$$\text{(d)} \qquad \int_{W_\nu}^{\infty} |l_\epsilon'(\omega)|^2\, d\omega < \frac{\epsilon^2}{2}.$$

(Such functions are easily constructable. In fact, $l_\epsilon(\omega)$ may be chosen to be linear from

$$\omega = W_\nu \quad \text{to} \quad \omega_0(\epsilon) = \frac{3W_\nu^2 + W_\nu}{\epsilon^2}$$

with

$$l_\epsilon(\omega) = 0 \quad \text{for} \quad \omega \geqq \omega_0(\epsilon).)$$

For any $\epsilon > 0$, define

$$\Phi_\epsilon(\omega) = \begin{cases} -\dfrac{\Phi(\omega)}{2} = -\dfrac{1}{2}\arg\left\{G(i\omega) + \dfrac{1}{k}\right\} & \text{for} \quad \omega \text{ in } [-W_\nu, W_\nu], \\ l_\epsilon(\omega) & \text{for} \quad \omega > W_\nu, \\ -l_\epsilon(-\omega) & \text{for} \quad \omega < W_\nu. \end{cases}$$

By this construction it is assured that

$$|\Phi_\epsilon(\omega)| < \frac{\pi}{2} \tag{16}$$

and

$$\left|\Phi_\epsilon(\omega) + \arg\left\{G(i\omega) + \frac{1}{k}\right\}\right| < \frac{\pi}{2}. \tag{17}$$

Also for future reference the inequality

$$\int_{-\infty}^{\infty} |\Phi_\epsilon'(\omega)|^2 \, d\omega \leqq \frac{1}{4}\int_{-W_\nu}^{+W_\nu} |\Phi'(\omega)|^2 \, d\omega + \epsilon^2 \tag{18}$$

holds.

It now follows by application of Lemma 1 that there is a $y_\epsilon(\cdot)$ in $L_1(0, \infty)$ with Laplace transform $Y(s)$, $1 + Y_\epsilon(i\omega) \neq 0$ for any ω and $\arg\{1 + Y_\epsilon(i\omega)\} = \Phi_\epsilon(\omega)$. Remark 5 combined with (16) therefore yields

$$\operatorname{Re}\{1 + Y_\epsilon(i\omega)\} \geqq \delta(\epsilon) > 0$$

for some constant $\delta(\epsilon)$. This proves Lemma 3 (i). In fact (i) holds with $y(\cdot)$ equal to any $y_\epsilon(\cdot)$ chosen by the above procedure. The actual choice of $y_\epsilon(\cdot)$ which will be made will be one corresponding to ϵ sufficiently small so as to assure that Lemma 3 (ii) also holds.

In order for (ii) to hold for $y(\cdot)$ equal to some $y_\epsilon(\cdot)$, it is sufficient, by Remark 5, to show that

$$|\arg\{1 + Y_\epsilon(i\omega + \sigma)\} + \Phi(\omega)| < \frac{\pi}{2}, \qquad -\infty < \omega < \infty. \tag{19}$$

Now (19) may be rewritten as

$$|\arg\{1 + Y_\epsilon(i\omega + \sigma)\} - \Phi_\epsilon(\omega) + \Phi_\epsilon(\omega) + \Phi(\omega)| < \frac{\pi}{2}, \qquad -\infty < \omega < \infty. \tag{20}$$

From Lemma 2 it follows that for all ω,

$$(21)\quad |\arg\{1 + Y_\epsilon(i\omega + \sigma)\} - \Phi_\epsilon(\omega)| \leqq \frac{16\sqrt{\sigma}}{3\pi}\left[\int_{-\infty}^{\infty} |\Phi_\epsilon'(\omega)|^2\, d\omega\right]^{1/2},$$

and so a sufficient condition for (20) to hold is that

$$(22)\quad \frac{16\sqrt{\sigma}}{3\pi}\left[\int_{-\infty}^{\infty} |\Phi_\epsilon'(\omega)|^2\, d\omega\right]^{1/2} + \sup_{-\infty<\omega<\infty} |\Phi_\epsilon'(\omega) + \Phi(\omega)| < \frac{\pi}{2}.$$

Now

$$\sup_{\omega,\,|\omega|\leqq W_\nu} |\Phi_\epsilon(\omega) + \Phi(\omega)| = \sup_{\omega,\,|\omega|\leqq W_\nu} \frac{|\Phi(\omega)|}{2} \leqq \frac{\pi - \alpha}{2},$$

while

$$\sup_{\omega,\,|\omega|\geqq W_\nu} |\Phi_\epsilon(\omega) - \Phi(\omega)| \leqq \sup_{\omega,\,|\omega|\geqq W_\nu} |\Phi_\epsilon(\omega)| + \sup_{\omega,\,|\omega|\geqq W_\nu} |\Phi(\omega)|$$

$$\leqq \frac{\nu}{2} + \nu = \frac{3\nu}{2} = \frac{\pi - \alpha}{2}.$$

It follows that

$$\sup_{-\infty<\omega<\infty} |\Phi_\epsilon(\omega) - \Phi(\omega)| \leqq \frac{\pi - \alpha}{2}.$$

Using this relation in conjunction with inequalities (20) and (21), one finds that a sufficient condition for (19) to hold is that

$$\frac{16\sqrt{\sigma}}{3\pi}\left[\int_{-\infty}^{\infty} |\Phi_\epsilon'(\omega)|^2\, d\omega\right]^{1/2} + \frac{\pi - \alpha}{2} < \frac{\pi}{2}$$

or

$$(23)\quad \frac{16\sqrt{\sigma}}{3\pi}\left[\int_{-\infty}^{\infty} |\Phi_\epsilon'(\omega)|^2\, d\omega\right]^{1/2} < \frac{\alpha}{2}.$$

Therefore if σ_* is chosen equal to

$$\frac{(3\pi\alpha/16)^2}{\int_{-W_\nu}^{W_\nu} |\Phi'(\omega)|^2\, d\omega}$$

as hypothesized in this lemma, then for any σ, $0 < \sigma < \sigma_*$, the inequality

$$\left[\int_{-W_\nu}^{W_\nu} |\Phi'(\omega)|^2\, d\omega\right]^{1/2} \cdot \frac{16\sqrt{\sigma}}{3\pi} < \alpha$$

holds.

But then if ϵ is chosen sufficiently small, inequality (23) will hold by virtue of (18). This means, on tracing through the above argument, that the $y_\epsilon(\cdot)$ corresponding to such a choice of ϵ will be an acceptable choice for the $y(\cdot)$ in the statement of this lemma, i.e., (i) and (ii) of the state-

ment of this lemma hold for $Y_\epsilon(s)$, the Laplace transform of $y_\epsilon(\cdot)$. This completes the proof.

5. A factorization lemma. In this section a necessary and sufficient condition for the factorization property of $n(\cdot)$ as described in Theorem 2 will be derived. For convenience of notation, a class $\mathfrak{K}$ of real-valued functions is introduced.

DEFINITION. Let $\mathfrak{K}$ be the class of absolutely continuous real-valued functions $k(\cdot)$ defined on $[0, \infty)$ with each $k(\cdot)$ satisfying $0 < \inf k(t) \leqq \sup k(t) < \infty$.

LEMMA 4. *Let $k(\cdot) \in \mathfrak{K}$ be given and let σ be a nonnegative constant. The following two statements are then equivalent:*

(i) *There is a constant $K > 0$ such that*

$$\left| \frac{1}{t} \int_0^t \left| \frac{k'(\tau)}{k(\tau)} - 2\sigma \right| d\tau - 2\sigma \right| \leqq \frac{K}{t}$$

for all $t > 0$.

(ii) *There are functions $k_+(\cdot)$ and $k_-(\cdot)$ both in K satisfying the following properties for all $t \geqq 0$:*

(a) $k(t) = k_+(t) \cdot k_-(t)$,

(b) $k_+(t) \exp(-2\sigma t)$ *is monotone nonincreasing,*

(c) $k_-(t)$ *is monotone nondecreasing.*

It is somewhat simpler to write out the proof of the following equivalent version of Lemma 4 (obtained by considering $\log k(\cdot)$).

LEMMA 4′. *Let $l(\cdot)$ be a real-valued absolutely continuous bounded (above and below) function of t for t in $[0, \infty)$. Then the following two statements are equivalent:*

(i′) *There is a constant $K > 0$ such that*

$$2\sigma t - K \leqq \int_0^t |l'(\tau) - 2\sigma| \, d\tau \leqq 2\sigma t + K$$

for all $t \geqq 0$.

(ii′) *There are two real-valued absolutely continuous bounded (above and below) functions $l_+(\cdot)$ and $l_-(\cdot)$, defined on $[0, \infty)$, with the following properties:*

(a′) $l(\cdot) = l_+(\cdot) + l_-(\cdot)$,

(b′) $dl_+(t)/dt \leqq 2\sigma$ *a.e. in t,*

(c′) $dl_-(t)/dt \geqq 0$ *a.e. in t.*

Proof of Lemma 4′. Assume at first that (i′) holds:

Define

$$l_+(t) \triangleq l(0) + \int_0^t \frac{l'(\tau) + 2\sigma - |l'(\tau) - 2\sigma|}{2} \, d\tau$$

$$\triangleq l(0) + \int_0^t \min \{l'(\tau), 2\sigma\} \, d\tau$$

and

$$l_-(t) \triangleq \int_0^t \frac{l'(\tau) - 2\sigma + |\, l'(\tau) - 2\sigma \,|}{2}\, d\tau \tag{24}$$
$$\triangleq \int_0^t \max\,(l'(\tau) - 2\sigma, 0)\, d\tau.$$

Then with these definitions it is easy to check that (a′), (b′) and (c′) hold. What is left to be shown is that $l_+(\cdot)$ and $l_-(\cdot)$ are bounded functions.

But

$$l_+(t) = l(t) + \sigma t - \frac{1}{2}\int_0^t |\, l'(\tau) - 2\sigma \,|\, d\tau$$

while

$$l_-(t) = l(t) - l(0) + \frac{1}{2}\int_0^t |\, l'(\tau) - 2\sigma \,|\, d\tau - \sigma t.$$

Assumption (i′) and the fact that $l(t)$ is bounded assure that $l_+(\cdot)$ and $l_-(\cdot)$ are bounded functions and this completes the proof of (i′) → (ii′).

In the other direction assume (ii′) holds. Then it follows that

$$\frac{dl_-(t)}{dt} \geqq \max\,(l'(t) - 2\sigma, 0) \triangleq \frac{(l'(t) - 2\sigma) + |\, l'(t) - 2\sigma \,|}{2} \tag{25}$$

for almost all $t \geqq 0$.

To see this, note that if (25) did not hold, there would exist a set of positive measure on which

$$0 \leqq \frac{dl_-(t)}{dt} < l'(t) - 2\sigma.$$

But

$$\frac{dl(t)}{dt} = \frac{dl_+(t)}{dt} + \frac{dl_-(t)}{dt} \qquad \text{a.e.}$$

so that one would have

$$0 \leqq \frac{dl(t)}{dt} - \frac{dl_+(t)}{dt} < \frac{dl(t)}{dt} - 2\sigma$$

on some set of positive measure. This, of course, would imply that $dl_+(t)/dt > 2\sigma$ on a set of positive measure which would then violate (b′). Thus (25) is verified.

Now integrating both sides of (25) from 0 to t yields

$$l_-(t) - l_-(0) \geqq \frac{l(t) - l(0)}{2} - \sigma t + \frac{1}{2}\int_0^t |\, l'(\tau) - 2\sigma \,|\, d\tau$$

valid for all $t \geqq 0$. Noting that both $l_-(\cdot)$ and $l(\cdot)$ are bounded (above and below) by assumption (ii′), the estimate

$$\int_0^t | l'(\tau) - 2\sigma | \, dt \leqq 2\sigma t + K$$

follows for some constant $K > 0$ and all $t \geqq 0$.

The inequality

$$\int_0^t | l'(\tau) - 2\sigma | \, dt \geqq 2\sigma t - K$$

follows easily from the fact that

$$\max (l'(t) - 2\sigma, 0) \triangleq \frac{(l'(t) - 2\sigma) + | l'(t) - 2\sigma |}{2}$$

is nonnegative for almost all $t \geqq 0$. With these remarks, the proof of (ii′) $\rightarrow$ (i′) is complete.

COROLLARY 3. *A necessary and sufficient condition for assumption* (ii) *of Theorem 2 to hold is that either*:

(a) *there is a* $K > 0$ *such that*

$$\left| \frac{1}{t} \int_0^t \left| \frac{\dot{n}(\tau)}{n(\tau)(1 - n(\tau)/k)} - 2\sigma \right| d\tau - 2\sigma \right| \leqq \frac{K}{t} \quad \textit{for all} \quad t > 0,$$

or else

(b) *there is a* $K > 0$ *such that*

$$\left| \frac{1}{t} \int_0^t \left| \frac{\dot{n}(\tau)}{n(\tau)(1 - n(t)/k)} + 2\sigma \right| d\tau - 2\sigma \right| \leqq \frac{K}{t} \quad \textit{for all} \quad t > 0.$$

Proof. Apply Lemma 4 to

$$n^*(t) \triangleq n(t) \left(1 - \frac{n(t)}{k}\right)^{-1}$$

to get (a) or apply Lemma 4 to

$$\frac{1}{n^*(t)} \triangleq \left(1 - \frac{n(t)}{k}\right) \frac{1}{n(t)}$$

to get (b).

Concluding remarks. The techniques developed here show promise of eliminating the multiplier dependence from many recent stability results. In particular, investigations are currently under way on problems involving monotone nonlinearities (see [2]) with a view towards obtaining geometric criteria there.

Acknowledgment. The author wishes to express his deep appreciation to

G. Zames for introducing him to the areas considered here and for numerous suggestions and discussions concerning these topics.

REFERENCES

[1] R. W. Brockett and L. J. Forys, *On the stability of systems containing a time-varying gain*, Proc. Second Allerton Conference on Circuit and Systems Theory, University of Illinois, Urbana, 1964, pp. 413–430.

[2] P. L. Falb and G. Zames, *Stability conditions for systems with monotone and odd monotone non-linearities*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 221–223.

[3a] G. Zames, *On the stability of non-linear, time-varying feedback systems*, Proc. NEC, 20 (1964), pp. 725–730.

[3b] ———, *On the input-output stability of time-varying non-linear feedback systems, I, II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, 465–476.

[4] M. I. Freedman and G. Zames, *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6 (1968), pp. 487–507.

[5] E. W. Hobson, *The Theory of Functions of a Real Variable*, vol. 1, Dover, New York, 1957.

[6] E. C. Titchmarsh, *Theory of Fourier Integrals*, 2nd. ed., Clarendon Press, Oxford, 1962.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

DIFFERENTIAL GAMES WITH DELAY*

A. HALANAY†

1. Introduction. The purpose of this paper is to consider some min-max problems for systems with time-lag. The first one is a simple linear pursuit-evasion problem with prescribed duration which was studied by Ho, Bryson and Baron [1] in the case when there is no retardation. This problem can be completely solved and shows that even if the systems for the players are of the simplest form with constant concentrated retardation, the equilibrium system will be a system with distributed retardation. Only for the most general linear systems with time-lag will the equilibrium system be of the same form.

This first problem suggests a setting for the question of necessary conditions for the equilibrium strategies. As always in such cases, one must define a suitable class of admissible strategies. The pursuit-evasion problem considered above shows that there are situations when the equilibrium strategies are differentiable with respect to the state. Since the experience of control problems shows that this is unrealistic when the strategies take values in closed sets, we suppose that these sets are open, although the min-max conditions could be derived in more general cases.

The method is the one used by Berkovitz [2] and Zelikin and Tynianski [3] and consists in considering two associated variational problems; from the above discussion it follows that we are obliged to consider optimal control problems for the most general delayed systems. This is done by using the methods of Hestenes [4] (see Appendix).

It is seen that this paper is less related to the problems of the theory of differential games as this theory is surveyed by Berkovitz in [5]; indeed all the difficult game-theoretical points are avoided. The paper should be considered as a contribution to the general theory of delayed systems and more precisely to the variational aspects of this theory.

2. A linear pursuit-evasion problem. Consider the pursuit-evasion problem

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

$$\dot{y}(t) = C(t)y(t) + D(t)y(t-\tau) + H(t)v(t),$$

$$x_{t_0} = \phi, \qquad y_{t_0} = \psi, \qquad x_t(s) \triangleq x(t+s),$$

$$y_t(s) \triangleq y(t+s), \qquad s \in [-\tau, 0],$$

* Received by the editors October 24, 1967, and in revised form April 18, 1968.

† Institutul de Matematica, str. M. Eminescu 47, Bucuresti 9, Romania.

$$J(u, v) = \frac{a^2}{2}[x^*(T) - y^*(T)]K^*K[x(T) - y(T)] + \frac{1}{2}\int_{t_0}^{T}(u^*Lu - v^*Mv)\,dt, \qquad L > 0, \qquad M > 0,$$

where matrices A, B, C, D, G, H, L, M are continuous on $[t_0, T]$ and K is constant.

The couple $(\tilde{u}, \tilde{v})$ will give equilibrium strategies if $J(\tilde{u}, v) \leqq J(\tilde{u}, \tilde{v}) \leqq J(u, \tilde{v})$ for all piecewise continuous u, v.

Let $X(T, t)$, $Y(T, t)$ be the solution matrices corresponding to the adjoint systems:

$$\frac{d}{dt}X(T, t) = -X(T, t)A(t) - X(T, t + \tau)B(t + \tau), \qquad X(T, T) = E, \qquad X(T, t) \equiv 0, \qquad t > T,$$

$$\frac{d}{dt}Y(T, t) = -Y(T, t)C(t) - Y(T, t + \tau)D(t + \tau), \qquad Y(T, T) = E, \qquad Y(T, t) \equiv 0, \qquad t > T.$$

Let

$$z(t) = K\Big\{X(T, t)x(t) - Y(T, t)y(t) + \int_t^{t+\tau}[X(T, \alpha)B(\alpha)x(\alpha - \tau) - Y(T, \alpha)D(\alpha)y(\alpha - \tau)]\,d\alpha.$$

It follows that

$$z(T) = K[x(T) - y(T)], \qquad \dot{z}(t) = \tilde{G}(t)u(t) - \tilde{H}(t)v(t),$$

$$\tilde{G}(t) = KX(T, t)G(t), \qquad \tilde{H}(t) = KY(T, t)H(t),$$

$$J(u, v) = \frac{a^2}{2}|z(T)|^2 + \frac{1}{2}\int_{t_0}^{T}(u^*Lu - v^*Mv)\,dt.$$

Suppose $\tilde{G}L^{-1}\tilde{G}^* \geqq \tilde{H}M^{-1}\tilde{H}^*$ in $[t_0, T]$. Then

$$\det\left\{\frac{1}{a^2}E + \int_t^T(\tilde{G}L^{-1}\tilde{G}^* - \tilde{H}M^{-1}\tilde{H}^*)\,dt\right\} \neq 0 \quad \text{for} \quad t \in [t_0, T]$$

and

$$P(t) = \left\{\frac{1}{a^2}E + \int_t^T(\tilde{G}L^{-1}\tilde{G}^* - \tilde{H}M^{-1}\tilde{H}^*)\,dt\right\}^{-1}$$

exists in $[t_0, T]$.

It is easy to see that P is the solution of the Riccati-type matrix equation

$$\frac{dP}{dt} = P(\tilde{G}L^{-1}\tilde{G}^* - \tilde{H}M^{-1}\tilde{H}^*)P$$

and $P(T) = a^2E$.

If we denote $V(t, z) = \frac{1}{2}(P(t)z, z)$, we shall have

$$V(T, z(T)) = \tfrac{1}{2}(P(T)z(T), z(T)) = \tfrac{1}{2}a^2 \mid z(T) \mid^2,$$

$$\begin{aligned} J(u, v) &= V(T, z(T)) + \frac{1}{2}\int_{t_0}^{T} \{(Lu, u) - (Mv, v)\}\, dt \\ &= V(t_0, z(t_0)) + \int_{t_0}^{T} \frac{dV(t, z(t))}{dt}\, dt \\ &\qquad + \frac{1}{2}\int_{t_0}^{T} \{(Lu, u) - (Mv, v)\}\, dt, \end{aligned}$$

and by a direct calculation we get

$$\begin{aligned} J(u, v) = V(t_0, z(t_0)) + \frac{1}{2}\int_{t_0}^{T} &\{(L(u + L^{-1}Q^*z), u + L^{-1}Q^*z) \\ &- (M(v + M^{-1}R^*z), v + M^{-1}R^*z)\}\, dt, \end{aligned}$$

$$Q = P\tilde{G}, \qquad R = P\tilde{H}.$$

It follows that the optimal strategies are

$$\tilde{u} = -L^{-1}Q^*z, \qquad \tilde{v} = -M^{-1}R^*z$$

since $J(\tilde{u}, \tilde{v}) = V(t_0, z(t_0))$ and

$$\begin{aligned} J(\tilde{u}, v) &= V(t_0, z(t_0)) - \frac{1}{2}\int_{t_0}^{T} (M(v + M^{-1}R^*z), v + M^{-1}R^*z)\, dt \\ &\leqq V(t_0, z(t_0)) \\ &\leqq V(t_0, z(t_0)) + \frac{1}{2}\int_{t_0}^{T} (L(u + L^{-1}Q^*z), u + L^{-1}Q^*z)\, dt \\ &= J(u, \tilde{v}). \end{aligned}$$

The optimal system will be

$$\begin{aligned} \dot{x}(t) = [A(t) &- GL^{-1}Q^*KX(T, t)]x(t) \\ &+ GL^{-1}Q^*KY(T, t)y(t) + B(t)x(t - \tau) \\ &- GL^{-1}Q^*K\int_{t-\tau}^{t} [X(T, \alpha + \tau)B(\alpha + \tau)x(\alpha) \\ &\qquad - Y(T, \alpha + \tau)D(\alpha + \tau)y(\alpha)]\, d\alpha, \end{aligned}$$

$$\dot{y}(t) = -HM^{-1}R^*KX(T,t)x(t)$$
$$+ [C(t) + HM^{-1}R^*KY(T,t)]y(t) + D(t)y(t-\tau)$$
$$+ HM^{-1}R^*K\int_{t-\tau}^{t} [X(T,\alpha+\tau)B(\alpha+\tau)x(\alpha)$$
$$- Y(T,\alpha+\tau)D(\alpha+\tau)y(\alpha)]\,d\alpha.$$

The results show that although we started with a simple system with retarded argument, the synthesized system is a more general one.

If we consider systems of the form

$$\dot{x}(t) = A(t)x(t) + \sum_k B_k(t)x(t-\tau_k)$$
$$+ \int_{-\tau}^{0} A(t,s)x(t+s)\,ds + G(t)u(t),$$
$$\dot{y}(t) = C(t)y(t) + \sum_k D_k(t)y(t-\tau_k)$$
$$+ \int_{-\tau}^{0} C(t,s)y(t+s)\,ds + H(t)v(t),$$

the matrices $X(T,t)$, $Y(T,t)$ will satisfy

$$\frac{d}{dt}X(T,t) = -X(T,t)A(t) - \sum_k X(T,t+\tau_k)B_k(t+\tau_k)$$
$$- \int_{-\tau}^{0} X(T,t-\sigma)A(t-\sigma,\sigma)\,d\sigma,$$

$$\frac{d}{dt}Y(T,t) = -Y(T,t)C(t) - \sum_k Y(T,t+\tau_k)D_k(t+\tau_k)$$
$$- \int_{-\tau}^{0} Y(T,t-\sigma)C(t-\sigma,\sigma)\,d\sigma,$$

and we shall define $z(t)$ by

$$z(t) = K\{X(T,t)x(t) - Y(T,t)y(t)$$
$$+ \sum_k \int_{-\tau_k}^{0} [X(T,t+\alpha+\tau_k)B(t+\alpha+\tau_k)x(t+\alpha)$$
$$- Y(T,t+\alpha+\tau_k)D(t+\alpha+\tau_k)y(t+\alpha)]\,d\alpha$$
$$+ \int_{-\tau}^{0} \left\{\left[\int_t^{t+\alpha+\tau} X(T,s)A(s,t+\alpha-s)\,ds\right]x(t+\alpha)\right.$$
$$\left. - \left[\int_t^{t+\alpha+\tau} Y(T,s)C(s,t+\alpha-s)\,ds\right]y(t+\alpha)\right\}d\alpha\}\,.$$

Then we get the same result as above, i.e., $\tilde{u} = -L^{-1}Q^*z$, $\tilde{v} = M^{-1}R^*z$,

and the synthesized system will be of the same type as the initial one; in fact, we shall have more terms with distributed retardation than in the initial system.

For the general systems,

$$\dot{x}(t) = \int_{-\tau}^{0} [d_s \eta(t, s)]x(t + s) + G(t)u(t),$$

$$\dot{y}(t) = \int_{-\tau}^{0} [d_s \zeta(t, s)]y(t + s) + H(t)v(t),$$

and $X(T, t)$ and $Y(T, t)$ are defined by

$$X(T, t) + \int_t^T X(T, \gamma)\eta(\gamma, t - \gamma)\, d\gamma = E,$$

$$X(T, t) \equiv 0 \quad \text{for} \quad t > T, \quad X(T, T) = E,$$

$$Y(T, t) + \int_t^T Y(T, \gamma)\zeta(\gamma, t - \gamma)\, d\gamma = E,$$

$$Y(T, t) \equiv 0 \quad \text{for} \quad t > T, \quad Y(T, T) = E,$$

and

$$z(t) = K\Big\{X(T, t)x(t) - Y(T, t)y(t) + \int_{-\tau}^{0}\Big[d_s \int_t^{t+\tau} X(T, \gamma)\eta(\gamma, t + s - \gamma)\, d\gamma\Big]x(t + s) - \int_{-\tau}^{0}\Big[d_s \int_t^{t+\tau} Y(T, \gamma)\zeta(\gamma, t + s - \gamma)\, d\gamma\Big]y(t + s)\Big\}.$$

We have then

$$z(t) = K[x(T) - y(T)] - \int_t^T \{KX(T, \alpha)G(\alpha)u(\alpha) - KY(T, \alpha)H(\alpha)v(\alpha)\}\, d\alpha,$$

and the same as above holds; the synthesized system will now be of exactly the same type as the initial one. For all the computations concerning the general system with time-lag, see [6].

3. Necessary conditions for equilibrium strategies. The problem we solved gives a setting for the general question of necessary conditions of optimality.

Consider a differential game with delay of the form

$$\dot{x}(t) = f(t, x_t, u, v), \quad x_{t_0} = \phi, \quad x_t(s) \triangleq x(t + s), \quad s \in [-\tau, 0],$$

$$I(u, v) = g(x(T)) + \int_{t_0}^{T} L(t, x_t, u, v)\, dt,$$

where T is fixed, f and L are continuously differentiable with respect to x, u, v, and continuous with respect to t; g is C^1.

We shall consider the admissible strategies $u(t, z)$ piecewise continuous with respect to t in $[t_0, T]$, differentiable with respect to z, and taking their values in a set U, supposed to be *open*; the admissible strategies $v(t, z)$ will be considered of the same type.

To the couple (u, v) of admissible strategies will correspond the solution defined by the system

$$\dot{x}(t) = f(t, x_t, u(t, x_t), v(t, x_t)), \qquad x_{t_0} = \phi,$$

and

$$I(u, v) = g(x(T)) - \int_{t_0}^{T} L(t, x_t, u(t, x_t), v(t, x_t))\, dt.$$

For the admissible strategies we shall require that the corresponding solution be defined for $t_0 \leqq t \leqq T$. The couple $(\tilde{u}, \tilde{v})$ will be an equilibrium one if for all admissible strategies (u, v), $I(\tilde{u}, v) \leqq I(\tilde{u}, \tilde{v}) \leqq I(u, \tilde{v})$.

We are now interested in necessary conditions for $(\tilde{u}, \tilde{v})$ to be equilibrium strategies. The method will be essentially the same as in [2], [3]. Suppose that $(\tilde{u}, \tilde{v})$ is optimal and let $\tilde{x}$ be the corresponding solution; let $\bar{u}(t) = \tilde{u}(t, \tilde{x}_t)$, $\bar{v}(t) = \tilde{v}(t, \tilde{x}_t)$. Consider the control system

$$\dot{x}(t) = f(t, x_t, u, \tilde{v}(t, x_t)), \qquad x_{t_0} = \phi,$$

$$J_1(u) = g(x(T)) + \int_{t_0}^{T} L(t, x_t, u, \tilde{v}(t, x_t))\, dt.$$

For this system $\tilde{x}$ will be the solution which corresponds to the control $\bar{u}$, and $J_1(\bar{u}) = I(\tilde{u}, \tilde{v})$, $J_1(u) = I(u, \tilde{v})$; since the piecewise continuous controls $u(t)$ are a particular case of admissible strategies, we deduce that $J_1(\bar{u}) \leqq J_1(u)$ for all piecewise continuous u with values in U.

For the control problem considered, the maximum principle is true in the following form: If $\bar{u}$ is optimal, then for all $t \in [t_0, T]$ where $\bar{u}$ is continuous and $u \in U$, we have

$$H_1(t, \tilde{x}_t, u, \psi) \leqq H_1(t, \tilde{x}_t, \bar{u}(t), \psi),$$

where

$$H_1(t, \tilde{x}_t, u, \psi) = \psi f(t, \tilde{x}_t, u, \tilde{v}(t, \tilde{x}_t)) = L(t, \tilde{x}_t, u, \tilde{v}(t, \tilde{x}_t))$$

and ψ is defined by

$$\psi(t) + \int_t^T [\psi(\sigma)\eta_1(\sigma, t - \sigma) - \zeta_1(\sigma, t - \sigma)]\, d\sigma = -\frac{\partial g}{\partial x}(\tilde{x}(T)),$$

$$\psi(t) \equiv 0 \quad \text{for} \quad t > T$$

and

$$\psi(T) = -\frac{\partial g}{\partial x}(\tilde{x}(T)).$$

The functions with bounded variation with respect to the second argument η_1, ζ_1, correspond, according to the well-known theorem of Riesz, to the operators

$$\frac{\partial f}{\partial x_t}(t, \tilde{x}_t, \bar{u}, \bar{v}) + \frac{\partial f}{\partial v}(t, \tilde{x}_t, \bar{u}, \bar{v}) \frac{\partial \bar{v}}{\partial x_t}(t, \tilde{x}_t)$$

and

$$\frac{\partial L}{\partial x_t}(t, \tilde{x}_t, \bar{u}, \bar{v}) + \frac{\partial L}{\partial v}(t, \tilde{x}_t, \bar{u}, \bar{v}) \frac{\partial \bar{v}}{\partial x_t}(t, \tilde{x}_t),$$

respectively.

Consider now the control system

$$\dot{x}(t) = f(t, x_t, \tilde{u}(t, x_t), v), \qquad x_{t_0} = \phi,$$

$$J_2(v) = g(x(T)) + \int_{t_0}^{T} L(t, x_t, \tilde{u}(t, x_t), v)\, dt.$$

Then $\tilde{x}$ is the solution corresponding to the control $\bar{v}$:

$$J_2(\bar{v}) = I(\tilde{u}, \bar{v}), \qquad J_2(v) = I(\tilde{u}, v);$$

hence $J_2(\bar{v}) \geqq J_2(v)$ for all piecewise continuous v with values in V.

The maximum principle gives

$$xf(t, \tilde{x}_t, \bar{u}(t), v) - L(t, \tilde{x}_t, \bar{u}(t), v) \geqq \chi f(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)) - L(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t))$$

for all $v \in V$ and $t \in [t_0, T]$ where $\bar{v}$ is continuous, χ being the solution of the equation

$$\chi(t) + \int_t^T [\chi(\sigma)\eta_2(\sigma, t - \sigma) - \zeta_2(\sigma, t - \sigma)]\, d\sigma = -\frac{\partial g}{\partial x}(\tilde{x}(T)),$$

$$\chi(t) \equiv 0 \quad \text{for} \quad t > T, \quad \chi(T) = -\frac{\partial g}{\partial x}(\tilde{x}(T)).$$

The functions with bounded variation with respect to the second argument η_2 and ζ_2 correspond to the operators

$$\frac{\partial f}{\partial x_t}(t, \tilde{x}_t, \bar{u}, \bar{v}) + \frac{\partial f}{\partial u}(t, \tilde{x}_t, \bar{u}, \bar{v}) \frac{\partial \tilde{u}}{\partial x_t}(t, \tilde{x}_t)$$

and

$$\frac{\partial L}{\partial x_t}(t, \tilde{x}_t, \bar{u}, \bar{v}) + \frac{\partial L}{\partial u}(t, \tilde{x}_t, \bar{u}, \bar{v}) \frac{\partial \tilde{u}}{\partial x_t}(t, \tilde{x}_t).$$

Using the fact that U and V are supposed to be *open*, we have from the first maximum condition

$$\psi \frac{\partial f}{\partial u}(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)) = \frac{\partial L}{\partial u}(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)), \tag{1}$$

and from the second,

$$\chi \frac{\partial f}{\partial v}(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)) = \frac{\partial L}{\partial v}(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)). \tag{2}$$

Let η_3, η_1', η_2', ζ_3 be the functions with bounded variation with respect to the second argument corresponding to the operators

$$\frac{\partial f}{\partial x_t}(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)), \quad \frac{\partial \tilde{v}}{\partial x_t}(t, \tilde{x}_t), \quad \frac{\partial \tilde{u}}{\partial x_t}(t, \tilde{x}_t), \quad \frac{\partial L}{\partial x_t}(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)),$$

respectively. Then we have

$$\eta_1 = \eta_3 + \frac{\partial f}{\partial v}(t, \tilde{x}_t, \bar{u}, \bar{v})\eta_1', \qquad \eta_2 = \eta_3 + \frac{\partial f}{\partial u}(t, \tilde{x}_t, \bar{u}, \bar{v})\eta_2',$$

$$\zeta_1 = \zeta_3 + \frac{\partial L}{\partial v}(t, \tilde{x}_t, \bar{u}, \bar{v})\eta_1', \qquad \zeta_2 = \zeta_3 + \frac{\partial L}{\partial u}(t, \tilde{x}_t, \bar{u}, \bar{v})\eta_2'.$$

The equations for ψ and χ can be written

$$\psi(t) + \int_t^T \Big[\psi(\sigma)\eta_3(\sigma, t-\sigma) + \psi(\sigma)\frac{\partial f}{\partial v}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_1'(\sigma, t-\sigma) - \zeta_3(\sigma, t-\sigma) - \frac{\partial L}{\partial v}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_1'(\sigma, t-\sigma) \Big] d\sigma = -\frac{\partial g}{\partial x}(\tilde{x}(T)),$$

$$\chi(t) + \int_t^T \Big[\chi(\sigma)\eta_3(\sigma, t-\sigma) + \chi(\sigma)\frac{\partial f}{\partial u}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_2'(\sigma, t-\sigma) - \zeta_3(\sigma, t-\sigma) - \frac{\partial L}{\partial u}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_2'(\sigma, t-\sigma) \Big] d\sigma = -\frac{\partial g}{\partial x}(\tilde{x}(T)),$$

and taking into account (1) and (2),

$$\psi(t) + \int_t^T \Big[\psi(\sigma)\eta_3(\sigma, t-\sigma)$$

$$+ \psi(\sigma) \frac{\partial f}{\partial v}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_1'(\sigma, t-\sigma) - \zeta_3(\sigma, t-\sigma)$$

$$- \chi(\sigma) \frac{\partial f}{\partial v}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_1'(\sigma, t-\sigma)\Big] d\sigma = -\frac{\partial g}{\partial x}(\tilde{x}(T)),$$

$$\chi(t) + \int_t^T \Big[\chi(\sigma)\eta_3(\sigma, t-\sigma)$$

$$+ \chi(\sigma) \frac{\partial f}{\partial u}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_2'(\sigma, t-\sigma) - \zeta_3(\sigma, t-\sigma)$$

$$- \psi(\sigma) \frac{\partial f}{\partial u}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_2'(\sigma, t-\sigma)\Big] d\sigma = -\frac{\partial g}{\partial x}(\tilde{x}(T)).$$

It follows that the difference $\psi - \chi$ satisfies the equation

$$\psi(t) - \chi(t) + \int_t^T [\psi(\sigma) - \chi(\sigma)]\Big[\eta_3(\sigma, t-\sigma)$$

$$+ \frac{\partial f}{\partial v}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_1'(\sigma, t-\sigma)$$

$$+ \frac{\partial f}{\partial u}(\sigma, \tilde{x}_\sigma, \bar{u}(\sigma), \bar{v}(\sigma))\eta_2'(\sigma, t-\sigma)\Big] d\sigma = 0,$$

$$\psi(t) - \chi(t) \equiv 0 \quad \text{for} \quad t \geqq T,$$

and hence $\psi(t) - \chi(t) \equiv 0$ for all t.

The equation for ψ is now

$$\psi(t) + \int_t^T [\psi(\sigma)\eta_3(\sigma, t-\sigma) - \zeta_3(\sigma, t-\sigma)]\, d\sigma = -\frac{\partial g}{\partial x}(\tilde{x}(T)),$$

$$\psi(t) \equiv 0 \quad \text{for} \quad t > T,$$

and we obtain the following result. The optimal strategies $\tilde{u}(t, x_t)$, $\tilde{v}(t, x_t)$ and the corresponding optimal solution $\tilde{x}_t$ satisfy the minimax principle:

$$\psi(t)f(t, \tilde{x}_t, u, \bar{v}(t)) - L(t, \tilde{x}_t, u, \bar{v}(t))$$

$$\leqq \psi(t)f(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t)) - L(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t))$$

$$\leqq \psi(t)f(t, \tilde{x}_t, \bar{u}(t), v) - L(t, \tilde{x}_t, \bar{u}(t), v)$$

for all $u \in U$, $v \in V$, ψ being defined above.

Recall that η_3 and ζ_3 are the functions corresponding to the operators $(\partial f/\partial x_t)(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t))$, $(\partial L/\partial x_t)(t, \tilde{x}_t, \bar{u}(t), \bar{v}(t))$, respectively.

Apply this general result to the case of linear systems with quadratic cost functional considered in §2.

We have, denoting $\begin{pmatrix} x \\ y \end{pmatrix} = z$,

$$f(t, z_t, \bar{u}(t), \bar{v}(t)) = \begin{bmatrix} \int_{-\tau}^{0} [d_s \eta(t, s)]x(t + s) + G(t)\bar{u}(t) \\ \int_{-\tau}^{0} [d_s \zeta(t, s)]y(t + s) + H(t)\bar{v}(t) \end{bmatrix}$$

and

$$\frac{\partial f}{\partial z_t}(t, z_t, \bar{u}(t), \bar{v}(t)) = \begin{bmatrix} \int_{-\tau}^{0} [d_s \eta(t, s)]x(t + s) \\ \int_{-\tau}^{0} [d_s \zeta(t, s)]y(t + s) \end{bmatrix}.$$

The corresponding function is

$$\begin{pmatrix} \eta(t, s) & 0 \\ 0 & \zeta(t, s) \end{pmatrix};$$

$L(t, z_t, \bar{u}, \bar{v})$ does not depend on z_t and hence $\partial L/\partial z_t \equiv 0$.

The equation for ψ will now be

$$(\psi_1(t)\ \psi_2(t)) + \int_t^T (\psi_1(\sigma)\ \psi_2(\sigma)) \begin{pmatrix} \eta(\sigma, t - \sigma) & 0 \\ 0 & \zeta(\sigma, t - \sigma) \end{pmatrix} d\sigma = \text{const.},$$

$\psi(t) \equiv 0$ for $t > T$,

$$\psi(T) = -a^2([\tilde{x}^*(T) - \tilde{y}^*(T)]K^*K, -[\tilde{x}^*(T) - \tilde{y}^*(T)]K^*K);$$

hence

$$\psi_1(t) + \int_t^T \psi_1(\sigma)\eta(\sigma, t - \sigma)\, d\sigma = -a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^*K,$$

$$\psi_2(t) + \int_t^T \psi_2(\sigma)\zeta(\sigma, t - \sigma)\, d\sigma = a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^*K.$$

It follows that, with the notations used in §2,

$$\psi_1(t) = -a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^*KX(T, t),$$

$$\psi_2(t) = a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^*KY(T, t).$$

The minimax conditions (1) and (2) now give

$$\psi_1 G = \bar{u}^*(t)L, \qquad \psi_2 H = -\bar{v}^*(t)M;$$

hence

$$\bar{u}^* = \psi_1 G L^{-1}, \qquad \bar{v}^* = -\psi_2 H M^{-1},$$
$$\bar{u}^*(t) = -a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^* K X(T,t)G(t)L^{-1}(t),$$
$$\bar{v}^*(t) = -a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^* K Y(T,t)H(t)M^{-1}(t),$$

and with the notations of §2,

$$\bar{u}^* = -a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^* \tilde{G} L^{-1},$$
$$\bar{v}^* = -a^2[\tilde{x}^*(T) - \tilde{y}^*(T)]K^* \tilde{H} M^{-1}.$$

Thus

$$\bar{u} = -a^2 L^{-1}\tilde{G}^* K[\tilde{x}(T) - \tilde{y}(T)],$$
$$\bar{v} = -a^2 M^{-1}\tilde{H}^* K[\tilde{x}(T) - \tilde{y}(T)].$$

To see that these formulas agree with the ones obtained in §2, recall that

$$\bar{u}(t) = -L^{-1}\tilde{G}^* P\tilde{z}(t),$$
$$\bar{v}(t) = -M^{-1}\tilde{H}^* P\tilde{z}(t),$$

and since

$$\dot{z} = \tilde{G}u - \tilde{H}v,$$

we have for $\tilde{z}$ the system

$$\dot{z} = -(\tilde{G}L^{-1}\tilde{G}^* - \tilde{H}M^{-1}\tilde{H}^*)Pz$$

and, using the equation for P, we obtain

$$\dot{z} = -P^{-1}\dot{P}z.$$

Hence $P\dot{z} + \dot{P}z = 0$ and

$$P(t)\tilde{z}(t) = P(T)\tilde{z}(T) = a^2 K[\tilde{x}(T) - \tilde{y}(T)].$$

It follows that

$$\bar{u} = -a^2 L^{-1}\tilde{G}^* K[\tilde{x}(T) - \tilde{y}(T)],$$
$$\bar{v} = -a^2 M^{-1}\tilde{H}^* K[\tilde{x}(T) - \tilde{y}(T)],$$

which is the same result that we obtained from the minimax principle.

Appendix. Optimal control systems with time-lag. Consider a control system with time-lag

$$\dot{x}(t) = f(t, x_t, u_t, b), \quad x_{t_0} = \alpha, \quad u_{t_0} = \eta, \quad b \in B \subset R^s, \quad x(T) = \omega(b),$$

and the functionals

$$I_\gamma = g_\gamma(b) + \int_{t_0}^{T} L_\gamma(t, x_t, u_t, b)\, dt.$$

The couple (u, b) is admissible if u is a piecewise continuous function on $[t_0, T]$, $u(t) \in U$, where U is a given set, $u_{t_0} = \eta$ and if, for the corresponding solution of the system with $x_{t_0} = \alpha$, we have $x(T) = \omega(b)$, $I_\gamma \leqq 0$, $1 \leqq \gamma \leqq p'$, $I_\gamma = 0$, $p' < \gamma \leqq p$; the couple is optimal if it is admissible and minimizes I_0 in the class of admissible couples.

Let $(\tilde{u}, \tilde{b})$ be an optimal couple, $\tilde{x}$ the corresponding optimal solution. Let

$$A(t) = \frac{\partial f}{\partial x_t}(t, \tilde{x}_t, \tilde{u}_t, \tilde{b}), \quad c_\gamma(t) = \frac{\partial L_\gamma}{\partial x_t}(t, \tilde{x}_t, \tilde{u}_t, \tilde{b}).$$

By the Riesz theorem we can write

$$A(t)\phi = \int_{-\tau}^{0} [d_s\, \eta(t, s)]\phi(s), \qquad c_\gamma(t)\phi = \int_{-\tau}^{0} [d_s\, \zeta_\gamma(t, s)]\phi(s).$$

Let $q_\gamma(t)$ be defined by

$$q_\gamma(t) + \int_t^T [q_\gamma(\sigma)\eta(\sigma, t - \sigma) + \zeta_\gamma(\sigma, t - \sigma)]\, d\sigma = 0,$$

$$q_\gamma(t) \equiv 0 \quad \text{for} \quad t > T,$$

and $p_i(t)$ be defined by

$$p_i(t) + \int_t^T p_i(\sigma)\eta(\sigma, t - \sigma)\, d\sigma = \text{const.},$$

$$p_i(t) \equiv 0 \quad \text{for} \quad t > T,$$

$$p_{ij}(T) = \delta_{ij}.$$

Define

$$F_\gamma(t, \phi_1, \phi_2, b) = L_\gamma(t, \phi_1, \phi_2, b) - c_\gamma(t)\phi_1 + q_\gamma[f(t, \phi_1, \phi_2, b) - A(t)\phi_1], \qquad 0 \leqq \gamma \leqq p,$$

$$F_{p+i}(t, \phi_1, \phi_2, b) = p_i[f(t, \phi_1, \phi_2, b) - A(t)\phi_1].$$

Remark that by the definitions of A and c_γ, we shall have $\partial F_\gamma/\partial\phi_1 = 0$, $\partial F_{p+i}/\partial\phi_1 = 0$ along the optimal solution considered. Indeed

$$F_\gamma(t, \phi_1, \tilde{u}_t, \tilde{b}) - F_\gamma(t, \tilde{x}_t, \tilde{u}_t, \tilde{b}) = L_\gamma(t, \phi_1, \tilde{u}_t, \tilde{b}) - L_\gamma(t, \tilde{x}_t, \tilde{u}_t, \tilde{b}) - c_\gamma(t)\phi_1$$

$$+ c_\gamma(t)\tilde{x}_t + q_\gamma(t)[f(t, \phi_1, \tilde{u}_t, \tilde{b}) - f(t, \tilde{x}_t, \tilde{u}_t, \tilde{b}) - A(t)(\phi_1 - \tilde{x}_t)]$$
$$= c_\gamma(t)(\phi_1 - \tilde{x}_t) + o(\|\phi_1 - \tilde{x}_t\|) - c_\gamma(t)(\phi_1 - \tilde{x}_t) + q_\gamma(t)o(\|\phi_1 - \tilde{x}_t\|)$$
$$= o(\|\phi_1 - \tilde{x}_t\|);$$

for F_{p+i} the proof is the same.

We choose $G_\gamma(b)$, $G_{p+i}(b)$ such that

$$\vartheta_\gamma(u, b) = G_\gamma(b) + \int_{t_0}^{T} F_\gamma(t, x_t, u_t, b)\, dt = I_\gamma(u, b),$$

$$\vartheta_{p+i}(u, b) = G_{p+i}(b) + \int_{t_0}^{T} F_{p+i}(t, x_t, u_t, b)\, dt = -\omega^i(b) + x^i(T).$$

To do this we compute[1]

$$\int_{t_0}^{T} F_\gamma(t, x_t, u_t, b)\, dt = \int_{t_0}^{T} L_\gamma(t, x_t, u_t, b)\, dt - \int_{t_0}^{T} c_\gamma(t)x_t\, dt + \int_{t_0}^{T} q_\gamma(t)f(t, x_t, u_t, b)\, dt - \int_{t_0}^{T} q_\gamma(t)A(t)x_t\, dt.$$

We have

$$\int_{t_0}^{T} [q_\gamma(t)A(t) + c_\gamma(t)]x_t\, dt$$
$$= \int_{t_0}^{T} q_\gamma(t)\left[\int_{-\tau}^{0} d_s\, \eta(t, s)x(t + s)\right] dt + \int_{t_0}^{T}\left[\int_{-\tau}^{0} d_s\, \zeta_\gamma(t, s)x(t + s)\right] dt$$
$$= \int_{t_0}^{T}\left[\int_{-\tau}^{0} d_s[q_\gamma(t)\eta(t, s) + \zeta_\gamma(t, s)]x(t + s)\right] dt.$$

After some transformations we obtain

$$\int_{t_0}^{T} [q_\gamma(t)A(t) + c_\gamma(t)]x_t\, dt$$
$$= \int_{t_0}^{T} d_t \int_{t_0}^{T} [q_\gamma(\sigma)\eta(\sigma, t - \sigma) + \zeta_\gamma(\sigma, t - \sigma)]\, d\sigma\, x(t) + \int_{t_0 - \tau}^{t_0} d_\sigma \int_{t_0}^{T} [q_\gamma(t)\eta(t, \sigma - t) + \zeta_\gamma(t, \sigma - t)]\, dt\, x(\sigma)$$

[1] In all the calculations we suppose that $\zeta_\gamma(t, s) \equiv 0$ for $t > T$; in fact, ζ_γ is defined by c_γ only for $t_0 \leqq t \leqq T$.

and, using the equation for q_γ,

$$\int_{t_0}^{T} [dq_\gamma(t)]x(t) + \int_{t_0}^{T} [q_\gamma(t)A(t) + c_\gamma(t)]x_t\, dt$$
$$= \int_{t_0-\tau}^{t_0} d_\sigma \int_{t_0}^{T} [q_\gamma(t)\eta(t, \sigma - t) + \zeta_\gamma(t, \sigma - t)]\, dt\, x(\sigma).$$

Hence we have

$$\int_{t_0}^{T} F_\gamma(t, x_t, u_t, b)\, dt = \int_{t_0}^{T} L_\gamma(t, x_t, u_t, b)\, dt - q_\gamma(t_0)x(t_0)$$
$$- \int_{t_0-\tau}^{t_0} d_\sigma \int_{t_0}^{T} [q_\gamma(t)\eta(t, \sigma - t) + \zeta_\gamma(t, \sigma - t)]\, dt\, x(\sigma),$$

and we can choose

$$G_\gamma(b) = g_\gamma(b) + g_\gamma(t_0)x(t_0)$$
$$+ \int_{t_0-\tau}^{t_0} d_\sigma \int_{t_0}^{T} [q_\gamma(t)\eta(t, \sigma - t) + \zeta_\gamma(t, \sigma - t)]\, dt\, x(\sigma).$$

In the same way,

$$\int_{t_0}^{T} F_{p+i}(t, x_t, u_t, b)\, dt = - p_i(t_0)x(t_0) + x^i(T)$$
$$- \int_{t_0-\tau}^{t_0} d_\sigma \int_{t_0}^{T} p_i(t)\eta(t, \sigma - t)\, dt\, x(\sigma)$$

and

$$G_{p+i}(b) = -\omega^i(b) + p_i(t_0)x(t_0) + \int_{t_0-\tau}^{t_0} d_\sigma \int_{t_0}^{T} p_i(t)\eta(t, \sigma - t)\, dt\, x(\sigma).$$

In order to obtain the maximum principle in the form we used here we shall consider that the control u is not delayed and shall use the abstract theorem of Hestenes [4]. If $\psi = -\sum \lambda_\gamma q_\gamma - \sum \lambda_{p+i} p_i$, we obtain for ψ the equation

$$\psi(t) + \int_t^T [\psi(\sigma)\eta(\sigma, t - \sigma) - \sum \lambda_\gamma\, \zeta_\gamma(\sigma, t - \sigma)]\, d\sigma = \text{const.},$$

$$\psi(t) \equiv 0 \quad \text{for} \quad t > T, \quad \psi_i(T) = -\lambda_{p+i}.$$

The Hamiltonian will be

$$H = \psi f - \sum \lambda_\gamma L_\gamma.$$

In the case considered for the theory of differential games we had no isoperimetric conditions so that

$$H = \psi f - L.$$

Since we can take here $x(T) = b$, hence $\omega(b) = b$, the transversality condition gives

$$0 = \frac{\partial g}{\partial b^\sigma} + \sum \lambda_i \frac{\partial G_i}{\partial b^\sigma}(\tilde{b}) = \frac{\partial g}{\partial x}(\tilde{x}(T)) - \lambda_i;$$

hence $\lambda_i = (\partial g/\partial x)(\tilde{x}(T))$ and $\psi(T) = -(\partial g/\partial x)(\tilde{x}(T))$.

Acknowledgment. The author is indebted to the referees for helpful suggestions. He would also like to express his appreciation to H. T. Banks for pointing out and correcting some errors in his book [7, Chap. IV, §3] concerning general systems with time-lag.

REFERENCES

[1] Yu-Chi Ho, A. E. Bryson and M. L. Baron, *Differential games and optimal pursuit-evasion strategies*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 385–389.

[2] L. Berkovitz, *A variational approach to differential games*, Ann. of Math. Studies, no. 52, Princeton University Press, Princeton, 1964, pp. 127–174.

[3] M. I. Zelikin and N. T. Tynianski, *Determined differential games*, Uspehi Mat. Nauk, 20 (1965), 4 (124), pp. 151–157.

[4] M. Hestenes, *On variational theory and optimal control theory*, this Journal, 3 (1965), pp. 23–48.

[5] L. Berkovitz, *A survey of differential games*, Mathematical Theory of Optimal Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 342–372.

[6] H. T. Banks, *Representations for solutions of linear functional differential equations*, Tech. Rep. 68-2, Center for Dynamical Systems, Brown University, Providence, 1968.

[7] A. Halanay, *Differential Equations: Stability, Oscillations, Time Lags*, Academic Press, New York, 1966.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

ERRATA: ON THE CLOSURE AND CONVEXITY OF ATTAINABLE SETS IN FINITE AND INFINITE DIMENSIONS*

H. HERMES

It has been pointed out to me by Taketomo Mitsui, University of Tokyo' that the argument given as a proof of property V, p. 414, is not valid' Specifically, the argument, as given, would be valid if we were to show that $\{x(\cdot, u^n)\}$ converges, rather than to select a convergent subsequence. This will now be done.

We had, for every nonnegative integer n, $z(s) = f(s, x(s, u^n), u^{n+1}(s))$ almost everywhere in $[0, t]$; hence,

$$
\begin{aligned}
&|\dot{x}(s, u^n) - \dot{x}(s, u^{n+1})| \\
&\qquad = |f(s, x(s, u^n), u^n(s)) - f(s, x(s, u^{n+1}), u^{n+1}(s))| \\
(*) \qquad &\qquad = |f(s, x(s, u^n), u^n(s)) - f(s, x(s, u^{n-1}), u^n(s)) \\
&\qquad\quad + f(s, x(s, u^n), u^{n+1}(s)) - f(s, x(s, u^{n+1}), u^{n+1}(s))| \\
&\qquad \leqq K\{|x(s, u^n) - x(s, u^{n-1})| + |x(s, u^n) - x(s, u^{n+1})|\}.
\end{aligned}
$$

As commented in Remark 2, p. 411, the assumptions 1, 2 of p. 410 imply $\mathfrak{F}$ is a bounded subset of $\mathfrak{L}_\infty[0, T]$; let ρ be the bound. Then for any k, $|\dot{x}(s, u^k)| \leqq \rho$ and $|\dot{x}(s, u^0) - \dot{x}(s, u^1)| \leqq 2\rho$ almost everywhere in $[0, t]$. Integrating from 0 to t implies $|x(t, u^0) - x(t, u^1)| \leqq 2\rho t$. Using this estimate in (*) for $n = 1$ gives

$$|\dot{x}(s, u^1) - \dot{x}(s, u^2)| \leqq K\{|x(s, u^1) - x(s, u^2)| + 2\rho s\}.$$

Integrating this from 0 to t gives

$$|x(t, u^1) - x(t, u^2)| \leqq K\left\{\frac{2\rho t^2}{2} + \int_0^t |x(s, u^1) - x(s, u^2)|\, ds\right\}.$$

This may be used in a similar fashion in (*) for $n = 2$ to obtain

$$|x(x, u^2) - x(t, u^3)| \leqq \frac{2K^2\rho t^3}{3!} + K^2 \int_0^t \int_0^s |x(\sigma, u^1) - x(\sigma, u^2)|\, d\sigma\, ds + K \int_0^t |x(s, u^2) - x(s, u^3)|\, ds.$$

* This Journal, 5 (1967), pp. 409–417. Received by the editors June 24, 1968.

Proceeding inductively, one gets

$$| x(t, u^n) - x(t, u^{n+1}) | \leqq \left[\frac{2K^n \rho t^{n+1}}{(n+1)!} + K^n \int_0^t \int_0^{t_1} \cdots \int_0^{t_{n-1}} | x(t_n, u^1) - x(t_n, u^2) | \, dt_n \cdots dt_1\right] + K \int_0^t | x(s, u^n) - x(s, u^{n+1}) | \, ds.$$

The standard equality

$$\int_0^t \int_0^{t_1} \cdots \int_0^{t_{n-1}} f(t_n) \, dt_n \cdots dt_1 = \frac{1}{n!}\left[\int_0^t f(\tau) \, d\tau\right]^n$$

yields

$$| x(t, u^n) - x(t, u^{n+1}) | \leqq \left\{\frac{2K^n \rho t^{n+1}}{(n+1)!} + \frac{K^n}{n!}\left[\int_0^t | x(\tau, u^1) - x(\tau, u^2) | \, d\tau\right]^n\right\} + K \int_0^t | x(\tau, u^n) - x(\tau, u^{n+1}) | \, d\tau.$$

Now apply the Gronwall inequality, take a limit as $n \to \infty$, and we obtain $| x(t, u^n) - x(t, u^{n+1}) | \to 0$ uniformly in $[0, T]$. This shows the sequence $\{x(\cdot, u^n)\}$ converges uniformly and the remainder of the argument as given on p. 414 completes the proof.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

ON THE OPTIMAL CONTROL OF A SYSTEM GOVERNED BY A LINEAR PARABOLIC EQUATION WITH WHITE NOISE INPUTS*

H. J. KUSHNER†

1. Introduction. Let $\mathcal{L}$ be a smooth elliptic operator whose coefficients are defined on $\bar{D} \times [0, T]$, where $\bar{D} \in E^n$ (Euclidean n-space).

We consider the problem of the optimal control of systems with the formal representation

$$(*) \qquad W_t = \mathcal{L}W + bu + \sum_i \sigma_i \xi_i$$

satisfying $W(x, t) \to 0$ as $x \to \partial D$, $t \geqq 0$, and with control

$$u(x, t) = \int k(v, x, t) W(v, t)\, dv$$

and cost criterion

$$C^u(\phi, t) \equiv E_\phi{}^u \int_t^T \int W(x, s) W(y, s) S(x, y, s)\, dx\, dy\, ds$$

$$+ E_\phi{}^u \int_t^T \int P(x, s) u^2(x, s)\, dx\, ds,$$

where ξ_i is the formal derivative of the Wiener process $z_i(t)$ and $E_\phi{}^u$ is the expectation given the control u, and initial condition $\phi(\cdot)$. A precise meaning is given to all terms in the sequel. An equation of the form $(*)$ seems like a useful model of a variety of noise disturbed objects, but it also arises in the following way. Suppose that an object is governed by $H_t = \mathcal{L}H + bu$ and the noise corrupted observations having the Itô differential (for each $x \in \bar{D}$) $dy_i(x, t) = dt \int m_i(v, x, t) H(v, t)\, dv + dw_i$ are taken, where the w_i are Wiener processes. Then, the conditional expectation[1] $W(x, t) \equiv E\{H(x, t) \mid y(v, s), s \leqq t, v \in \bar{D}\}$ has a representation in the form $(*)$. In fact, (26) is the relevant Riccati equation (with a reversed time parameter). The results herein concern the first boundary value problem for a

* Received by the editors September 25, 1967, and in revised form March 22, 1968.

† Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant AF-AFOSR-693-67, in part by the National Aeronautics and Space Administration under Grant 40-002-015, and in part by the National Science Foundation under Grant GK-967.

[1] The filtering problem will be treated in a subsequent work. This paper is devoted strictly to the control problem.

single parabolic equation. However, it is clear that the method is applicable (and results easily extendible) to the second boundary value problem, or to a family of parabolic equations. The latter model is quite versatile. For example, we can use a vector parameter process $W(x, t)$ generated by (*) as the input to another system (e.g., $(\partial/\partial t - \mathfrak{L})Y = W + b'u'$). We can thus generate a distributed system analogue of the linear Gauss-Markov processes, and treat the corresponding average quadratic cost control problem.

The results are based on the results of Lemma 1 which provide criteria which guarantee that there is a version of a vector parameter process which with probability one (w.p.1) is continuous or differentiable in some particular parameters. Without these latter results, (1) would lose its intuitive meaning (as would stochastic differential equations if the paths were not known to be continuous w.p.1). Previous works dealing with "random" partial differential equations [2, Chap. 14], [3] were concerned with the nature of the random solution corresponding to a random, but smooth, boundary condition. Once the necessary smoothness properties of the process $W(x, t)$ are established, much of the analysis is similar to the analysis of the corresponding deterministic problem. However, to our knowledge the few treatments of the deterministic problem (e.g., see the interesting reference [4]) are essentially formal in nature. Most of the proofs are slightly abbreviated. We have chosen to omit the details of several arguments dealing mainly with the smoothness properties of potentials and related integrals. The arguments are tedious and standard. Some are based on existence theorems (Lemma 2.3) and most others use the arguments of [5, Chap. 1, §3.5].

In §2, some needed results on processes with a vector parameter set are given. The proofs of the statements of Lemmas 2.1 and 2.2 are found in [1]. Theorems 3.1 and 3.2 define the solution of (*) and its basic properties; continuity w.p.1, existence of Hölder continuous w.p.1 second derivatives (with respect to the x_i), etc. The optimality and approximation in policy space results appear in §4. Although we deal with a single white noise input, the results are obviously valid for the no more general infinite-dimensional white noise input of Lemma 2.2.

2. Mathematical preliminaries.

DEFINITION. Following the usual usage, a version of the vector parameter scalar-valued process $f(y)$ is any scalar-valued process $\tilde{f}(y)$ such that $P\{\tilde{f}(y) = f(y)\} = 1$ for all vector parameter values y. Write D_i, D_iD_j, $D_iD_jD_k$ and D_t for the differential operators $\partial/\partial x_i$, $\partial^2/\partial x_i\partial x_j$, $\partial^3/\partial x_i\partial x_j\partial x_k$ and $\partial/\partial t$, respectively. D is a bounded open set, $\bar{D}$ its closure and $\bar{R}$ denotes $\bar{D} \times [0, T]$.

LEMMA 2.1. *Let the boundary ∂D of D have the property that any line intersects it only finitely often. Let the functions*

$$(**) \qquad \begin{aligned} &\alpha(x, t, s), \quad \{D_i\alpha(x, t, s)\}, \quad \{D_iD_j\alpha(x, t, s)\}, \\ &\{D_iD_jD_k\alpha(x, t, s)\}, \quad \{D_iD_jD_kD_l\alpha(x, t, s)\} \end{aligned}$$

be defined on $\bar{D} \times [0, T] \times [0, T] = \bar{R} \times [0, T]$, continuous in (x, t) for each s, and bounded (in absolute value) by a square integrable function of s. Let f be any function in the set (**), and let $z(t)$ be a Wiener process. Then $\int_0^T f^2(x, t, s)\, ds \leqq M < \infty$ for some real number M, and $\int_0^t f(x, t, s)\, dz_s$ can be defined to be a separable and measurable process with parameter (x, t). There is a null set N and a separable and measurable version of $\int_0^t \alpha(x, t, s)\, dz_s = \psi(x, t)$ which, for $\omega \notin N$, is continuous in (x, t) and has three continuous (in (x, t)) derivatives with respect to the components of x. These derivatives are equal to continuous (for $\omega \notin N$), separable and measurable versions of $\int_0^t D_i\alpha(x, t, s)\, dz_s$, $\int_0^t D_iD_j\alpha(x, t, s)\, dz_s$, $\int_0^t D_iD_jD_k\alpha(x, t, s)\, dz_s$, respectively.

Let in addition, for some real numbers $K < \infty$, $\beta > 0$,

$$(***) \qquad \begin{aligned} &E\left\{\int_0^{t+\Delta} f(x, t + \Delta, s)\, dz_s - \int_0^t f(x, t, s)\, dz_s\right\}^2 \\ &= \int_0^t [f(x, t + \Delta, s) - f(x, t, s)]^2\, ds + \int_t^{t+\Delta} f^2(x, t + \Delta, s)\, ds \\ &\leqq K\Delta^\beta, \end{aligned}$$

where f is any member of (**). Let g be any member of the first three sets of (**). Then the continuous version (for $\omega \notin N$) of $\int_0^t g(x, t, s)\, dz_s = \phi(x, t)$ is Hölder continuous on $\bar{R}$, i.e., there is some $K(\omega) < \infty$ w.p.1 and a real $\gamma > 0$ so that

$$|\phi(x + \delta, t + \Delta) - \phi(x, t)| \leqq K(\omega)[|\Delta|^\gamma + |\delta|^\gamma],$$

where $|\cdot|$ refers to the Euclidean norm.

Lemma 2.2, although not used in the paper, can be used to generalize slightly its results.

LEMMA 2.2. *Let $z_i(t)$, $i = 1, \cdots$, be a family of independent Wiener processes. Let ∂D satisfy the condition of Lemma 2.1. Let $\alpha_i(x, t, s)$ be a family of functions, each of which satisfies the conditions on the $\alpha(x, t, s)$ of Lemma 2.1,*

except that with M_i and K_i replacing the M and K of Lemma 2.1, *when f corresponds to α_i, we assume that $\sum_i M_i < \infty$, $\sum_i K_i < \infty$. Then the functions $\psi(x, t) = \sum_i \int_0^t \alpha_i(x, t, s)\, dz_{is}$ and $\phi(x, t) = \sum_i \int_0^t g_i(x, t, s)\, dz_{is}$ have the properties of ψ and ϕ of Lemma* 2.1. *g_i is either α_i, $D_j\alpha_i$ or $D_jD_k\alpha_i$.*

Let us collect the following assumptions here. Note that Hölder continuity in (x, t) on the compact set $\bar{R}$ is equivalent to uniform Hölder continuity in (x, t) on $\bar{R}$.

(E1) Let $R = D \times [0, T]$, where D is a bounded and open Borel measurable domain. To each point in ∂D, let there exist a neighborhood V and a function $h(\cdot)$ so that $\partial D \cap V$ has the representation $x_i = h(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$ for some component x_i, where $h(x)$ has Hölder continuous fourth partial derivatives.[2]

(E2) On $\bar{R}$, the coefficients of $\mathfrak{L}$ are bounded and Hölder continuous, together with their first two derivatives with respect to the x_i.

(E3) There exists a real number $K > 0$ such that

$$\sum_{i,j} a_{ij}(x, t)\xi_i\xi_j \geqq K \sum_i \xi_i^2$$

for any vector ξ.

(E4) Let $\sigma(x, t)$ and its first four derivatives in the components of x be Hölder continuous in $\bar{R}$.

(E5) Let $\sigma(x, t)$ and $\mathfrak{L}\sigma(x, t)$ tend to zero as $x \to \partial D$ in $\bar{R}$.

(E6) Let $b(x, t)$ be Hölder continuous on $\bar{R}$ and $k(y, x, t)$ be bounded, measurable and Hölder continuous in x, t on $\bar{R}$, uniformly in y.

It will be helpful to collect the following results here. They will be used without reference in the sequel.

LEMMA 2.3a [5, Chap. 3, Theorem 7 and §5]. *Suppose* (E1)–(E3) *hold. Let $\phi(x, 0)$ have Hölder continuous second derivatives in $\bar{D}$ and satisfy $\phi(x, 0) \to 0$ and $\mathfrak{L}\phi(x, 0) \to 0$ as $x \to \partial D$. Let $\phi(\partial D, t) \equiv 0$. Suppose that $f(x, t)$ is Hölder continuous in $\bar{R}$ and tends to zero as $x \to \partial D$.Then there is a continuous unique solution to $U_t = \mathfrak{L}U + f$ on $\bar{R}$ which satisfies the boundary condition $\phi(x, t)$. $U_t(x, t)$, $D_iU(x, t)$, $D_iD_jU(x, t)$, $D_iD_jD_kU(x, t)$ and $D_iD_jD_kD_lU(x, t)$ are Hölder continuous on $\bar{R}$ and $\mathfrak{L}U(x, t) \to 0$ as $x \to \partial D, t \geqq 0$.*

Define the Banach space $\bar{C}_\alpha^{\,0}$ of functions on $\bar{R}$ which satisfy $f(x, t) \to 0$ as $x \to \partial D$ and with the norm

$$\|f\|_\alpha = \sup_{x,t \in \bar{R}} |f(x, t)| + \sup_{\substack{x,t \in \bar{R} \\ y,s \in \bar{R}}} \frac{|f(x, t) - f(y, s)|}{|x - y|^\alpha + |t - s|^{\alpha/2}}.$$

Let $\bar{C}^0_{2+\alpha}$ be the sub- (Banach) space of $\bar{C}_\alpha^{\,0}$ of functions which satisfy

[2] (E1) implies the condition on $\bar{R}$ of Lemma 2.1.

$\mathfrak{L}f(x, t) \to 0$ *as* $x \to \partial D$, *and have the norm*

$$\| f \|_{2+\alpha} = \| f \|_\alpha + \sum_i \| D_i f \|_\alpha + \sum_{i,j} \| D_i D_j f \|_\alpha .$$

Then, with homogeneous boundary conditions, the equation $U_t - \mathfrak{L}U = f$ *represents a continuous linear map of* $\bar{C}_\alpha^0$ *into* $\bar{C}_{2+\alpha}^0$. (This follows from [5, Chap. 3, (2.21)] and the first part of the lemma.)

LEMMA 2.3b [5, Chap. 3, Theorem 16]. *Suppose* (E1)–(E3) *hold. Then there is a Green's function* $G(x, x'; t, t')$ *for* $U_t = \mathfrak{L}U$. $G(x, x'; t, t') \to 0$ *as* $x \to \partial D$ *for* $t > t'$. $D_i G(x, x'; t, t')$, $D_i D_j G(x, x'; t, t')$ *and* $G_t(x, x'; t, t')$ *are continuous in* (x, t) *on* $D \times (t', T]$. *If* $\sigma(x, t)$ *is continuous in* x *for each* t, *then* $\alpha(x, t, t') \equiv \int G(x, x'; t, t')\sigma(x', t')\, dx'$ *satisfies* $(\partial/\partial t - \mathfrak{L}) \cdot \alpha(x, t, t') = \sigma(x, t)$ *on* $D \cap (t', T]$ *and tends to zero as* $x \to \partial D$ *for* $t > t'$. *If* $\sigma(x, t)$ *and* $\mathfrak{L}\sigma(x, t)$ *tend to zero as* $x \to \partial D$ *and are Hölder continuous on* $\bar{R}$, *then* $\alpha(x, t, t')$ *satisfies the same conditions for* $(x, t) \in \bar{D} \cap [t', T]$. *If* $\sigma(x, t)$ *is bounded and measurable, then* $\int_0^t dt' \alpha(x, t, t')$ *tends to zero as* $x \to \partial D$ *and is continuous on* $\bar{R}$. (The last statement follows from the arguments concerning potentials in [5, Chap. 1, §3–§5].)

LEMMA 2.3c [5. Chap. 3, Theorem 17]. *Suppose* (E1)–(E3) *hold. Then the Green's function for the adjoint operator* $\partial/\partial t + \mathfrak{L}^*$ *is*

$$G^*(x, x'; t, t') = G(x', x; t', t).$$

Note that $G^*(x, x'; t, t')$ is defined for $t < t'$ and that $(\partial/\partial t + \mathfrak{L}^*)G^* \cdot (x, x'; t, t') = 0$ on $D \times [0, t')$.

3. The stochastic partial differential equation.

THEOREM 3.1. *Suppose* (E1)–(E5) *hold. Let* $\phi(x)$ *have Hölder continuous second derivatives on* $\bar{R}$ *and tend to zero as* $x \to \partial D$. *Let* $\mathfrak{L}\phi(x) \to 0$ *as* $x \to \partial D$. *Then there is a process* $W(x, t)$ *with parameter set* $\bar{R}$ *satisfying* (z' *denotes* $z(t')$)

$$W(x, t) = \int G(x, x'; t, 0)\phi(x')\, dx' + \int_0^t dz' \left\{ \int G(x, x'; t, t')\sigma(x', t')\, dx' \right\}. \tag{1}$$

There is a version of $W(x, t)$ *which is continuous on* $\bar{R}$ *w.p.*1 *and satisfies* $W(x, 0) = \phi(x)$, $W(\partial D, t) = 0$. $D_i W(x, t)$ *and* $D_i D_j W(x, t)$ *are also Hölder continuous on* $\bar{R}$ *w.p.*1 *and* $\mathfrak{L}W(x, t) \to 0$ *as* $x \to \partial D$. *Also, w.p.*1,

$$W(x, t) = \int G(x, x'; t, s)W(x', s)\, dx' + \int_s^t dz' \int G(x, x'; t, t')\sigma(x', t')\, dx'. \tag{2}$$

For each fixed x in $\bar{D}$, $W(x, t)$ has the Itô differential

$$dW = \mathcal{L}W(x, t)\, dt + \sigma(x, t)\, dz. \tag{3}$$

Remark. $W(x, t)$ is not (w.p.1) differentiable in t. The smoothness of $\sigma(x, t)$ determines the smoothness of $W(x, t)$. Lemma 2.1 plays a crucial role here. It is not a priori obvious that the last term in (1) has a version which is sufficiently smooth as a function of (x, t) w.p.1. The order of integration in the stochastic integral in (1) must be preserved. Also, the theorem implies that the $\mathcal{L}W$ term in (3) has a version which is continuous w.p.1 on $\bar{R}$.

It can be shown that the process $W(\cdot, t)$, *with parameter* t, is a continuous Markov process with values in a Banach space of functions which satisfy the appropriate boundary conditions ($W(x, t) \to 0$ and $\mathcal{L}W(x, t) \to 0$ as $x \to \partial D$) and have Hölder continuous second derivatives (for some fixed nonrandom Hölder exponent).

Proof. The proof is a consequence of Lemmas 2.1 and 2.3. Let $\beta(x, t)$ be the first term on the right of (1). Then by Lemma 2.3, $\mathcal{L}\beta = \beta_t$; $\beta(x, t)$ satisfies the boundary conditions and $\mathcal{L}\beta(x, t) \to 0$ as $x \to \partial D$ for $t \geqq 0$. $\beta(x, t)$ and $\mathcal{L}\beta(x, t)$ tend to $\phi(x)$ and $\mathcal{L}\phi(x)$, respectively, as $t \to 0$. Write

$$\alpha(x, t, t') = \int G(x, x'; t, t')\sigma(x', t')\, dx'.$$

Then for $t \geqq t'$ (t' fixed), $\alpha(x, t, t')$, $D_i\alpha(x, t, t')$, $D_iD_j\alpha(x, t, t')$, $D_iD_jD_k\alpha(x, t, t')$ and $D_iD_jD_kD_l\alpha(x, t, t')$ are Hölder continuous on $\bar{R}$ and each satisfies the conditions on $f(x, t, t')$ of Lemma 2.1. Hence, by Lemma 2.1, there is a version of

$$\psi(x, t) = \int_0^t dz'\alpha(x, t, t')$$

which is Hölder continuous w.p.1 on $\bar{R}$, and which has Hölder continuous second derivatives with respect to the x_i w.p.1 on $\bar{R}$. $D_i\psi(x, t)$ and $D_iD_j\psi(x, t)$ can be identified with the Hölder continuous versions (which exist w.p.1 in $\bar{R}$) of

$$D_i\psi(x, t) = \int_0^t dz' D_i\alpha(x, t, t') = \int_0^t dz' \int D_i G(x, x'; t, t')\sigma(x', t')\, dx',$$

$$D_iD_j\psi(x, t) = \int_0^t dz' D_iD_j\alpha(x, t, t')$$

$$= \int_0^t dz' \int D_iD_j G(x, x'; t, t')\sigma(x', t')\, dx',$$

respectively. Thus $\mathcal{L}\psi(x, t)$ has a version which is Hölder continuous w.p.1,

and which is clearly a Hölder continuous version of $\int_0^t dz' \mathfrak{L}\alpha(x, t, t')$. Using the continuity of $\mathfrak{L}\psi(x, t)$ and the fact that $\mathfrak{L}\psi(x, t) \to 0$ in probability (since $\mathfrak{L}\alpha(x, t, t') \to 0$) as $x \to \partial D$, we have $\mathfrak{L}\psi(x, t) \to 0$ as $x \to \partial D$. Thus (1) satisfies the required boundary conditions. Equation (3) follows from the definition of the Itô differential of (1) for each fixed x, and the observation that $d \int_0^t dz' \alpha(x, t, t') = dz\alpha(x, t, t) + dt \int_0^t dz' \alpha_t(x, t, t')$, where $\alpha(x, t, t) = \sigma(x, t)$ and $\alpha_t(x, t, t') = \mathfrak{L}\alpha(x, t, t')$.

Equation (2) is obviously true w.p.1 for each fixed x, t, s. To show that it is true w.p.1 on $\bar{D} \times [0, T] \times [0, T]$, note first that

$$\psi(x, t, s) \equiv \int_s^t dz' \left\{ \int G(x, x'; t, s)\sigma(x', s)\, dx' \right\}$$

can be defined (using previous arguments) to be continuous (as a function of x, t, s) on $\bar{D} \times [0, T] \times [0, T]$, except for ω in some null set N. Define $\psi(x, t)$ as the (unique) $\lim_{s \to 0} \psi(x, t, s)$. The limit exists for $\omega \notin N$, and is a version of the continuous (for $\omega \notin N$) function $\psi(x, t)$ defined previously. Then, for $\omega \notin N$,

$$\psi(x, t) = \psi(x, s) + \psi(x, t, s).$$

Now writing (1) in the equivalent form

$$W(x, t) = \int G(x, x'; t, 0)\phi(x')\, dx' + \int G(x, x'; t, s)\psi(x', s, 0)\, dx' + \psi(x, t, s)$$

and using the semigroup property

$$\int G(x, x'; t, 0)\rho(x')\, dx' = \int G(x, x''; t, s)\, dx'' \int G(x'', x'; s, 0)\rho(x')\, dx'$$

and the continuity w.p.1 of $\psi(x, t, s)$, we get (2), and the proof is complete.

The concern of this paper is restricted to systems with controls which are linear in $W(x, t)$ and which appear linearly in the differential equation. Thus adding a control term to (3) we have, formally,

$$dW = \mathfrak{L}W(x, t) \cdot dt + \sigma(x, t)\, dz + b(x, t)u(x, t)\, dt, \tag{3a}$$

where the control is

$$u(x, t) = \int W(y, t)k(y, x, t)\, dy. \tag{4}$$

Theorem 3.2 gives meaning to (3a).

THEOREM 3.2. *Suppose* (E6) *and all the assumptions of Theorem* 3.1 *hold, and let* $u(x, t)$ *be given by* (4). *Then there is a Hölder continuous* (*w.p.*1) *version of the following*:

$$W(x,t) = \int G(x,x';t,0)\phi(x')\,dx' + \int_0^t dz' \int G(x,x';t,t')\sigma(x',t')\,dx' + \int_0^t dt' \int G(x,x';t,t')b(x',t')u(x',t')\,dx'. \tag{5}$$

Furthermore (*w.p.*1) $D_iW(x, t)$ *and* $D_iD_jW(x, t)$ *are Hölder continuous on* $\bar{R}$, *and both* $W(x, t)$ *and* $\mathcal{L}W(x, t)$ *tend to zero as* $x \to \partial D$.

There is a kernel $B(x, x'; t, t')$ *so that*

$$W(x,t) = \int_0^t dt' \int B(x,x';t,t')q(x',t')\,dx', \tag{6}$$

where $q(x, t)$ *is the sum of the first two terms on the right side of* (5). $B(x, x'; t, t')$ *maps*[3] $\bar{C}^0_{2+\alpha}(\bar{R})$ *into* $\bar{C}^0_{2+\alpha}(\bar{R})$. $W(x, t)$ *has the Itô differential*

$$dW(x,t) = \mathcal{L}W(x,t)\,dt + \sigma(x,t)\,dz + b(x,t)\int k(y,x,t)W(y,t)\,dy\,dt. \tag{7}$$

Proof. Let $W(x, t)$ be Hölder continuous on $\bar{R}$ (exponent α) and tend to zero as $x \to \partial D$. Let the Hölder exponent in (E6) be $\gamma \geqq \alpha$. Then

$$f(x',t') = b(x',t')\int W(y,t')k(y,x',t')\,dy \tag{8}$$

is Hölder continuous on $\bar{R}$ (exponent α) and tends to zero as $x' \to \partial D$. Thus, by Lemma 2.3a, the last term in (5) maps $W(x, t) \in \bar{C}_\alpha{}^0$ continuously into $\bar{C}^0_{2+\alpha}$. Hence (5) also is a continuous linear map of $\bar{C}^0_{2+\alpha}$ into $\bar{C}^0_{2+\alpha}$.

Write (5) as

$$W(x,t) = q(x,t) + \int_0^t dt \int M(x,x';t,t')W(x',t')\,dx', \tag{9}$$

where

$$M(x,x';t,t') = \int G(x,y;t,t')b(y,t')k(x',y,t)\,dy.$$

The kernel $M(x, x'; t, t')$ must also correspond to a continuous linear map of

[3] See Lemma 2.3a for the definition of $\bar{C}^0_{2+\alpha}$.

$\bar{C}^0_{2+\alpha}$ into $\bar{C}^0_{2+\alpha}$. By Theorem 3.1 and Lemma 2.3a, $q(x, t)$ is in $\bar{C}^0_{2+\alpha}$ for some $\alpha > 0$ w.p.1. Then, the theory of Volterra integral equations asserts the existence of a $W(x, t) \in \bar{C}^0_{2+\alpha}$ satisfying (9) (and, hence, (5)) w.p.1. It also yields the representation (6).

The assertion concerning the Itô differential follows exactly as in the proof of Theorem 3.1.

4. The solution to the optimum control problem. The solution is divided into four theorems. Theorem 4.1 establishes some required properties of a partial differential integral equation (the analogue of the Ricatti equation). Theorem 4.2 establishes a formula for the cost corresponding to a fixed control. Then, in Theorem 4.3 the usual dynamic programming technique of quasi-linearization (or approximation in policy space) is applied to obtain a sequence of costs (and improved controls) which, in Theorem 4.4, converges to the minimum cost (and optimal control).

The *adjoint of* $\mathcal{L}$, *operating on functions of* x, is written as

$$\mathcal{L}_x^*\xi(x) = \sum_{i,j} D_iD_j[a_{ij}(x, t)\xi(x)] - \sum D_i[d_i(x, t)\xi(x)] + c(x, t)\xi(x).$$

Define the Banach space $\hat{C}^0_{2+\alpha}$ of functions on $\hat{R} = \bar{D} \times \bar{D} \times [0, T]$ satisfying the condition that $f(x, y, t)$, $\mathcal{L}_x^* f(x, y, t)$ and $\mathcal{L}_y^* f(x, y, t) \to 0$ as $x \to \partial D$ or $y \to \partial D$ or $t \to T$, and with norm

$$\| f \|_{2+\alpha} = \| f \|_\alpha + \sum_i \| D_{x_i} f \|_\alpha + \sum_i \| D_{y_i} f \|_\alpha + \sum_{i,j} \| D_{y_i} D_{y_j} f \|_\alpha + \sum_{i,j} \| D_{x_i} D_{x_j} f \|_\alpha + \| f_t \|_\alpha,$$

where

$$\| f \|_\alpha = \sup_{x,y,t \in \hat{R}} | f(x, y, t) | + \sup_{\substack{x,y,t \in \hat{R} \\ x',y',t' \in \hat{R}}} \frac{| f(x, y, t) - f(x', y', t') |}{| x - x' |^\alpha + | y - y' |^\alpha + | t - t' |^{\alpha/2}}.$$

THEOREM 4.1. *Assume the conditions of Theorem* 3.2. *Let* $Q(x, y, t)$ *be symmetric and Hölder continuous on* $\bar{D} \times \bar{D} \times [0, T]$. *Let* $Q(x, y, t) \to 0$ *as* $x \to \partial D$ *or* $y \to \partial D$. *Write*

$$R_t(x, y, t) + (\mathcal{L}_x^* + \mathcal{L}_y^*)R(x, y, t) + \int b(v, t)[k(x, v, t)R(v, y, t) + k(y, v, t)R(x, v, t)]\, dv = -Q(x, y, t). \tag{10}$$

There is a unique symmetric (*in* x, y) *and continuous solution to* (10) *which, in addition, is in* $\hat{C}^0_{2+\alpha}$.

Proof. By (E3), the adjoint operator and its Green's function are de-

fined. The proof involves some standard calculations, similar to those of [5, Chap. 1, §§3–5], and most of the details are left to the reader. Consider first the adjoint equation (11), defined in $\bar{D} \times \bar{D} \times [0, T]$,

$$\tilde{R}_t(x, y, t) + (\mathfrak{L}_x^* + \mathfrak{L}_y^*)\tilde{R}(x, y, t) = -Q(x, y, t) \tag{11}$$

with boundary conditions $\tilde{R}(x, y, t) \to 0$ as $x \to \partial D$ or $y \to \partial D$ or $t \to T$. The unique solution to (11) can be verified to be the symmetric function[4, 5]

$$\begin{aligned}\tilde{R}(x, y, t) &= \int_t^T ds \iint dx'\, dy' G(x', x; s, t)G(y', y; s, t)Q(x', y', s) \\ &= \int_t^T ds \iint dx'\, dy' G^*(x, x'; t, s)G^*(y, y'; t, s)Q(x', y', s).\end{aligned} \tag{12}$$

Write (12) as

$$R(x, y, t) = \int_t^T \int ds\, dx' G^*(x, x'; t, s)h(y, x'; t, s),$$

where

$$h(y, x'; t, s) = \int G^*(y, x'; t, s)Q(x', y', s)\, dy'.$$

$h(y, x'; t, s)$ is uniformly Hölder continuous in y, x', t, s for $s \geqq t$, and $h(y, x'; t, s) \to 0$ as $x \to \partial D$ or $y \to \partial D$. Let $t_0 \leqq t$ and consider

$$\tilde{R}(x, y, t \mid t_0) = \int_t^T \int ds\, dx' G^*(x, x'; t, s)h(y, x'; t_0, s).$$

$\tilde{R}(x, y, t \mid t_0)$ is the solution to the adjoint equation $U_t + \mathfrak{L}^* U = h$ (with $t = t_0$ in h) to which Lemma 2.3a is applicable.[6] Thus, for each fixed y, $\tilde{R}(x, y, t \mid t_0)$ and $\mathfrak{L}_x^* \tilde{R}(x, y, t \mid t_0)$ tend to zero as $x \to \partial D$ or $t \to T$, by virtue of the properties of $Q(x, y, t)$; also $\tilde{R}(x, y, t \mid t_0)$ is symmetric in x, y. Since $\tilde{R}(x, y, t_0 \mid t_0)$ and $\mathfrak{L}_x^* \tilde{R}(x, y, t_0 \mid t_0)$ tend to zero as $x \to \partial D$ or $y \to \partial D$, and t_0 is arbitrary, we conclude that the terms $\tilde{R}(x, y, t \mid t) = \tilde{R}(x, y, t)$ and $\mathfrak{L}_x^* \tilde{R}(x, y, t)$ and $\mathfrak{L}_y^* \tilde{R}(x, y, t)$ are Hölder continuous and tend to zero as $x \to \partial D$ or $y \to \partial D$ or $t \to T$. Now, to complete the proof that $\tilde{R}(x, y, t) \in \hat{C}^0_{2+\alpha}$, we need only show that $\tilde{R}_t(x, y, t)$ is Hölder continuous.

[4] Consider the differential equation (*) $\dot{P} = AP + PA' - Q$, with boundary condition $P(T) = 0$. Let $\Phi(t, s)$ be the fundamental matrix of $\dot{x} = Ax$. Then (*) has the solution (**) $P(t) = \int_t^T \Phi(t, s)Q(s)\Phi'(t, s)\, ds$. Note the similarity in form between (**) and (12). (Here $'$ denotes transpose.)

[5] Recall that $G_t^*(x, x'; t, t') + \mathfrak{L}_x^* G(x, x'; t, t') = 0$ for $t < t'$.

[6] The boundary conditions are $U(x, t) = 0$ on $\bar{D} \times \{T\} + \partial D \times [0, T]$. If the time parameter is reversed (changing the terminal manifold $\bar{D} \times \{T\}$ to an initial manifold $\bar{D} \times \{0\}$) then, since $\mathfrak{L}^*$ satisfies (E1)–(E3), Lemma 2.3a is applicable.

But this is true since $\tilde{R}(x, y, t \mid t_0)$ has Hölder continuous derivatives with respect to t and t_0 in $t \geqq t_0 \geqq 0$ (uniformly in x, y in $\bar{D} \times \bar{D}$).

For the rest of the proof, write (10) as the Volterra integral equation

$$R(x, y, t) = \tilde{R}(x, y, t) + \int_t^T ds \iint dx'\, dy' G^*(x, x'; t, s) G^*(y, y'; t, s) M(x', y', s), \tag{13}$$

where $M(x', y', s)$ is the integral term in the middle of (10) (with x', y', s substituted for x, y, t, respectively). Obtain (14) by changing the order of integration in (13):

$$\begin{aligned} R(x, y, t) &= \tilde{R}(x, y, t) \\ &\quad + \int_t^T ds \iint R(v, w, s) K(x, y; v, w; t, s)\, dv\, dw \\ &\quad + \int_t^T ds \iint R(v, w, s) K(y, x; w, v; t, s)\, dv\, dw, \end{aligned} \tag{14}$$

where K is defined by:

$$K(x, y; v, w; t, s) = b(v, s) G^*(y, w; t, s) \int G^*(x, x'; t, s) k(x', v, s)\, dx'. \tag{14a}$$

If there is a solution of the desired form to (14), then there is a solution of the desired form to (13). It can be verified that under the imposed conditions, the kernel K represents[7] a continuous linear map of $\hat{C}^0_{2+\alpha}$ into $\hat{C}_{2+\alpha}$.

Finally, it can be verified, via the theory of Volterra equations, that (14) does have a unique solution of the desired form. The proof is complete.

Let the control u be given by

$$u(x, s) = \int k^u(v, x, s) W(v, s)\, dv.$$

Then the cost becomes

$$C^u(\phi, t) = E_\phi{}^u \iint_t^T dx\, dy\, ds W(x, s) W(y, s) Q^u(x, y, s),$$

where

$$Q^u(x, y, s) = S(x, y, s) + \int k^u(x, v, s) k^u(y, v, s) P(v, s)\, dv.$$

Recall that the system that we are concerned with is defined by Theorem 3.2 and has the Itô differential

$$dW(x, t) = \mathfrak{L} W(x, t)\, dt + \sigma(x, t)\, dz + b(x, t) u(x, t)\, dt$$

[7] The map is given by the sum of the two integrals of (14).

and boundary conditions $W(x, t) \to 0$ as $x \to \partial D$.

THEOREM 4.2. *Assume the conditions of Theorem* 4.1. *Let* $W(x, t)$ *satisfy* (5), *with the control* $u(x, t)$ *given by* (4), *and let the initial condition* (*given at time* $t \in [0, T]$) $W(x, t) = \phi(x)$ *satisfy the conditions on* $\phi(x)$ *of Theorem* 3.1.

Suppose that $S(x, y, t)$, $P(v, t)$, $k^u(x, v, t)$ *and* $b(v, t)$ *are Hölder continuous in their arguments and* $S(x, y, t)$ *and* $k^u(x, v, t)$ *tend to zero as* $x \to \partial D$ *or* $y \to \partial D$. *Let* $P(x, t) > 0$ *on* $\bar{D} \times [0, T]$ *and let* $S(x, y, t)$ *be symmetric and nonnegative definite on* $\bar{D} \times \bar{D}$ *for each* t *in* $[0, T]$. *Then*

$$
\begin{aligned}
C^u(\phi, t) &\equiv E_\phi{}^u \iint_t^T dx\, dy\, ds Q^u(x, y, s) W(x, s) W(y, s) \\
&= p(t) + \int dx\, dy R^u(x, y, t) \phi(x) \phi(y),
\end{aligned} \tag{15}
$$

where

$$
p(t) = \frac{1}{2} \iint_t^T dx\, dy\, ds R^u(x, y, s) \sigma(x, s) \sigma(y, s),
$$

and $R^u(x, y, s)$ *is the function introduced in Theorem* 4.1 *corresponding to* $Q(x, y, s) = Q^u(x, y, s)$.

Proof. The assumptions on $S(x, y, t)$, $P(v, t)$, $k^u(x, v, t)$ and $b(v, t)$ guarantee that $Q^u(x, y, t)$ satisfies the conditions on $Q(x, y, t)$ in Theorem 4.1. For fixed x, y, the function

$$
R^u(x, y, r) W(x, r) W(y, r) \equiv F(x, y, r)
$$

has the Itô differential (Theorem 3.2) in $[0, T]$ (we use a version of $W(x, t)$ for which $\mathfrak{L}W(x, t)$ is continuous w.p.1—see Theorem 3.2):

$$
\begin{aligned}
dF(x, y, r) &= R_r{}^u(x, y, r) W(x, r) W(y, r)\, dr + R^u(x, y, r)[dW(x, r) W(y, r) \\
&\quad + W(x, r)\, dW(y, r) + dW(x, r)\, dW(y, r)] \\
&= R_r{}^u(x, y, r) W(x, r) W(y, r)\, dr \\
&\quad + R^u(x, y, r) W(y, r) \Big[\mathfrak{L}W(x, r)\, dr + \sigma(x, r)\, dz \\
&\qquad + b(x, r) \Big(\int W(v, r) k(v, x, r)\, dv \Big) dr \Big] \\
&\quad + R^u(x, y, r) W(x, r) \Big[\mathfrak{L}W(y, r)\, dr + \sigma(y, r)\, dz \\
&\qquad + b(y, r) \Big(\int W(v, r) k(v, y, r)\, dv \Big) dr \Big] \\
&\quad + \tfrac{1}{2} R(x, y, r) \sigma(x, r) \sigma(y, r)\, dr.
\end{aligned}
$$

This, together with $R^u(x, y, T) = 0$, implies (w.p.1 for each x, y, t) that

$$-F(x, y, t) = \int_t^T dF(x, y, r).$$

Furthermore,

$$\begin{aligned}
&-E_\phi{}^u F(x, y, t) \\
&= -E_\phi{}^u R^u(x, y, t)W(x, t)W(y, t) \\
&= E_\phi{}^u \int_t^T \Big[R_r{}^u(x, y, r)W(x, r)W(y, r) \\
&\quad + R^u(x, y, r)W(y, r)\Big(\mathfrak{L}W(x, r)\, dr + b(x, r)\int W(v, r)k(v, x, r)\, dv\Big) \\
&\quad + R^u(x, y, r)W(x, r)\Big(\mathfrak{L}W(y, r)\, dr + b(y, r)\int W(v, r)k(v, y, r)\, dv\Big)\Big] \\
&\quad + \frac{1}{2}\int_t^T R^u(x, y, r)\sigma(x, r)\sigma(y, r)\, dr.
\end{aligned} \tag{16}$$

Each of the integrands on the right side of (16) can be defined to be a measurable function of ω, x, y, r and absolutely integrable over $\Omega \times \bar{D} \times \bar{D} \times [t, T]$. Thus (16), together with Fubini's theorem, implies that

$$\begin{aligned}
&-E_\phi{}^u \int dx\, dy F(x, y, t) \\
&= -\int R^u(x, y, t)\phi(x)\phi(y)\, dx\, dy \\
&= E_\phi{}^u \iint_t^T dx\, dy\, dr \Big[R_r{}^u(x, y, r)W(x, r)W(y, r) \\
&\quad + R^u(x, y, r)W(y, r)\Big(\mathfrak{L}W(x, r) + b(x, r)\int W(v, r)k(v, x, r)\, dv\Big) \\
&\quad + R^u(x, y, r)W(x, r)\Big(\mathfrak{L}W(y, r) + b(y, r)\int W(v, r)k)v, y, r)\, dv\Big)\Big] \\
&\quad + p(t).
\end{aligned} \tag{17}$$

Now $W(x, t)$ (w.p.1) and $R^u(x, y, t)$ (for each fixed $y \in \bar{D}$) are continuous and have uniformly continuous first and second derivatives, with respect to the x_i, in the domain $\bar{R}$. Also $W(x, t)$ (w.p.1) and $R^u(x, y, t)$ tend to zero as $x \to \partial D$. Thus, upon partially integrating and using Green's identity to eliminate the boundary integrals which are obtained (which are zero, owing to the first two sentences of this paragraph), we get (for

ω not in some null set)

$$\iint_t^T dx\, dy\, dr R^u(x, y, r)W(y, r)\mathfrak{L}W(x, r)$$
$$= \iint_t^T dx\, dy\, dr \mathfrak{L}_x{}^* R^u(x, y, r)W(y, r)W(x, r).$$

Substituting this in (17), and using the symmetry of $R(x, y, r)$, yields, after another change in the order of integration,

$$-\int R^u(x, y, t)\phi(x)\phi(y)\, dx\, dy$$
$$= E_\phi{}^u \iint_t^T dx\, dy\, dr W(x, r)W(y, r)$$
$$\cdot \Big[R_r{}^u(x, y, r) + (\mathfrak{L}_x{}^* + \mathfrak{L}_y{}^*)R^u(x, y, r) \tag{18}$$
$$+ \int b(x', r)k(x, x', r)R^u(x', y, r)\, dx'$$
$$+ \int b(x', r)k(y, x', r)R^u(x, x', r)\, dx' \Big] + p(t).$$

Finally, using the relation (10) in (18) yields (15), and the proof is complete.

Write $C^u(W(x, t), t)$ for the function $C^u(\phi, t)$ with $W(x, t)$ substituted for $\phi(t)$. Write $d^vC^u(W(x, t), t)$ for the Itô differential of the cost (corresponding to control u) but where, in the expression for $dW(x, t)$, a control $v(x, t)$ replaces the control $u(x, t)$.

THEOREM 4.3. *Let $u(x, t)$ be a given control (then $k^u(x, y, t)$ is given) and assume the other conditions of Theorem* 4.2. *Let $\tilde{u}(x, t)$ be the function $v(x, t)$ which minimizes*

$$E_\phi{}^v \int_t^T d^vC^u(W(x, s), s)$$
$$+ E_\phi{}^v \int_t^T \int dx\, dy\, ds S(x, y, s)W(x, s)W(y, s) \tag{19}$$
$$+ E_\phi{}^v \int_t^T \int dx\, ds P(x, s)v^2(x, s)$$

or, equivalently, which minimizes

$$E_\phi{}^v \int dx\, dy R^u(x, y, z)[W(y, s)b(x, s)v(x, s) + W(x, s)b(y, s)v(y, s)] \tag{20}$$
$$+ E_\phi{}^v \int dx P(x, s)v^2(x, s).$$

Then

$$C^{\tilde{u}}(\phi, t) \leqq C^{u}(\phi, t). \tag{21}$$

$\tilde{u}(x, s)$ *is given by*

$$\tilde{u}(x, s) = -b(x, s) \int \frac{R^{u}(x, y, s)}{P(x, s)} W(y, s)\, dy \tag{22}$$

and

$$k^{\tilde{u}}(g, x, s) = -b(x, s)R^{u}(x, g, s)/P(x, s), \tag{23}$$

and the corresponding $R^{\tilde{u}}(x, y, t)$ *satisfies the conditions on the* $R^{u}(x, y, t)$ *of Theorem* 4.2. *Also*

$$\int R^{\tilde{u}}(x, y, t)\phi(x)\phi(y)\, dx\, dy \leqq \int R^{u}(x, y, t)\phi(x)\phi(y)\, dx\, dy \tag{24}$$

for any bounded and measurable function $\phi(x)$.

Proof. By Theorem 4.2, (19) is nonnegative and equals zero when $v(x, t) = u(x, t)$. Then any minimizing $v(x, t)$ (provided that the corresponding integrals of (19) exist) must leave (19) nonpositive. This, together with the facts that $C^{v}(\phi, t)$ is the sum of the last two integrals in (19) and that the first integral of (19) equals $-C^{u}(\phi, t)$, implies (21). The $v(x, t)$ minimizing (19) is the $v(x, t)$ which minimizes

$$\begin{aligned} E_{\phi}^{v} \int_{t}^{T} \int dx\, dy\, ds[&R_{s}^{u}(x, y, s)W(x, s)W(y, s) \\ &+ R^{u}(x, y, s)W(y, s)(\mathfrak{L}W(x, s) + b(x, s)v(x, s)) \\ &+ R^{u}(x, y, s)W(x, s)(\mathfrak{L}W(y, s) + b(y, s)v(y, s))] \\ &+ E_{\phi}^{v} \int_{t}^{T} \int dx\, dy\, dsS(x, y, s)W(x, s)W(y, s) \\ &+ E_{\phi}^{v} \int_{t}^{T} \int dx\, dsP(x, s)v^{2}(x, s) \end{aligned}$$

or, equivalently, which minimizes (20). The $v(x, t)$ minimizing (20) is given by (22). The statement below (23) is easily established (via Theorem 4.2) since the $k^{\tilde{u}}(y, x, t)$ of (23) satisfies the conditions on $k^{v}(y, x, t)$ in the hypothesis of Theorem 4.2. Equation (24) is valid for all doubly differential functions $\phi(x)$ which are zero on ∂D since

$$\begin{aligned} C^{v}(\phi, t) &= \int dx\, dyR^{v}(x, y, t)\phi(x)\phi(y) \\ &\qquad + \frac{1}{2} \int\int_{t}^{T} dx\, dy\, drR^{v}(x, y, s)\sigma(x, s)\sigma(y, s) \\ &\leqq \int dx\, dyR^{u}(x, y, t)\phi(x)\phi(y) \\ &\qquad + \frac{1}{2} \int\int_{t}^{T} dx\, dy\, dsR^{u}(x, y, s)\sigma(x, s)\sigma(y, s) \end{aligned}$$

for all such $\phi(x)$. Hence (24) is valid for all functions which are (almost everywhere) pointwise limits of a bounded sequence of such $\phi(x)$.

Theorem 4.4 is the optimality theorem. Let k^n, Q^n and R^n correspond to k^{u_n}, Q^{u_n} and R^{u_n}, respectively.

THEOREM 4.4. *Let $u_0(x, t)$ be given and let the corresponding $k^0(y, x, t)$ satisfy the conditions on $k^u(y, x, t)$ in Theorem 4.2. Suppose the other conditions of Theorem 4.2 hold. Define $u_n(x, t)$ from $u_{n-1}(x, t)$, $n = 1, \cdots$, via the procedure of Theorem 4.3. Then $R^n(x, y, t)$ converges pointwise (almost everywhere) to an $R(x, y, t)$ which satisfies the conditions of Theorem 4.1. The control $u(x, t)$ corresponding to $R(x, y, t)$ via (25) (see (22)),*

$$\begin{aligned} u(x, s) &= -b(x, s) \int \frac{R(x, y, s)W(y, s)}{P(x, s)}\, dy \\ &\equiv \int k(y, x, s)W(y, s)\, dy, \end{aligned} \tag{25}$$

is optimal in that $C^u(\phi, t) \leqq C^{\bar{v}}(\phi, t)$ for any other control

$$\bar{v}(x, s) = \int k^{\bar{v}}(y, x, s)W(y, s)\, dy, \tag{25a}$$

where $k^v(y, x, t)$ satisfies the condition on the $k^u(y, x, t)$ in Theorem 4.2. $R(x, y, t)$ also satisfies the boundary conditions on the $R^u(x, y, t)$ of Theorem 4.2, and the Ricatti equation

$$\begin{aligned} &R_t(x, y, t) + (\mathfrak{L}_x^* + \mathfrak{L}_y^*)R(x, y, t) \\ &+ \int b(v, t)[k(x, v, t)R(v, y, t) + k(y, v, t)R(x, v, t)]\, dv = -Q(x, y, t), \end{aligned} \tag{26}$$

where $k(y, x, t)$ is given by (25), $k(x, v, t) = -b(v, t)R(v, x, t)/P(v, t)$ and also

$$Q(x, y, t) = S(x, y, t) + \int k(x, v, t)k(y, v, t)P(v, s)\, dv. \tag{27}$$

Proof. The proof is divided into three steps. First, we show that $R^n(x, y, t)$ converges (almost everywhere) to some function $R(x, y, t)$; second that $R(x, y, t)$ is smooth and satisfies (26), and third, that $R(x, y, t)$ corresponds to the optimal control. By (24) (where u and $\tilde{u}$ are replaced by u_n and u_{n+1}) and the nonnegative definiteness of the $R^i(x, y, t)$,

$$\int dx\, dy[R^n(x, y, t) - R^{n+1}(x, y, t)]\phi(x)\phi(y) \geqq 0 \tag{28}$$

for any bounded measurable $\phi(x)$. Also the $R^n(x, y, t)$ are continuous in $\bar{D} \times \bar{D} \times [0, T]$. Hence, the diagonal values $R^n(x, x, t)$ are nonnegative and nonincreasing as n increases, and $R^n(x, x, t) \downarrow R(x, x, t)$ (al-

most everywhere) for some function $R(x, x, t)$. This, together with $\max_{x,y} | R^n(x, y, t) | \leqq \max_x | R^n(x, x, t) |$, implies that the $R^n(x, y, t)$ are uniformly bounded. In fact, the pointwise convergence implies that the diagonal values converge "almost uniformly" in the sense that for any fixed $\epsilon > 0$, there is a function $\alpha(N, \epsilon)$ tending to zero as $N \to \infty$ and a set $S_\alpha \subset \bar{D}$ with Lebesgue measure $\alpha(N, \epsilon)$ so that, for any $m, n > N$,

$$0 \leqq R^n(x, x, t) - R^m(x, x, t) < \epsilon \tag{29}$$

on $\bar{D} - S_\alpha$. Next, suppose that, for some $m > n > N$,

$$R^m(x', x'', t) - R^n(x', x'', t) > 2\epsilon \tag{29a}$$

on some $(x', x'') \in (\bar{D} - S_\alpha) \times (\bar{D} - S_\alpha)$. Then, by continuity and symmetry, there are neighborhoods A', A'' of x', x'', respectively (A', A'' are assumed to be in $\bar{D} - S_\alpha$) so that (29a) and (29) hold on $A' \times A'' \cup A'' \times A'$. *Let $I(A)(x)$ be the characteristic function of the set A.* Set $\phi(x) = I(A' \cup A'')(x)$. Then using this and the diagonal (29) bound in (28) gives

$$\int [R^n(x, y, t) - R^m(x, y, t)] I(A' \cup A'')(x) \cdot I(A' \cup A'')(y)\, dx\, dy \leqq (2\epsilon - 4\epsilon)\mu(A')\mu(A'') < 0$$

($u(\cdot)$ is Lebesgue measure on $\bar{D}$), a contradiction. Since α can be made arbitrarily small by increasing N, we conclude that $R^n(x, y, t)$ converges almost everywhere to a function $R(x, y, t)$. Furthermore, it is clear that $R(x, y, t)$ is symmetric, measurable and bounded almost everywhere by $r(x, y, t)$, where $r(x, y, t)$ is some function which tends to zero as $x \to \partial D$ or $y \to \partial D$.

To continue, we use the representation (see Theorems 4.1 and 4.3 for terminology)

$$\begin{aligned} R^{n+1}(x, y, t) &= \tilde{R}^{n+1}(x, y, t) \\ &\quad + \int_t^T \int ds\, dv\, dw K^{n+1}(x, y; v, w; t, s) R^{n+1}(v, w, s) \\ &\quad + \int_t^T \int ds\, dv\, dw K^{n+1}(y, x; w, v; t, s) R^{n+1}(v, w, s), \end{aligned} \tag{30}$$

$$\tilde{R}^{n+1}(x, y, t) = \int_t^T \int ds\, dx'\, dy' G^*(x, x'; t, s) G^* \cdot (y, y'; t, s) Q^{n+1}(x', y', s), \tag{31}$$

$$\begin{aligned} Q^{n+1}(x', y', s) &= S(x', y', s) \\ &\quad + \int k^{n+1}(x', v, s) k^{n+1}(y', v, s) P(v, s)\, dv, \end{aligned} \tag{32}$$

$$k^{n+1}(x', v, s) = -b(v, s)R^n(v, x', s)/P(v, s), \tag{33}$$

$$K^{n+1}(x, y; v, w; t, s) = \frac{-b^2(v, s)}{P(v, s)} G^*(y, w; t, s) \cdot \int G^*(x, x'; t, s)R^n(v, x', s)\, dx'. \tag{34}$$

The left side of (30) tends (almost everywhere) to $R(x, y, t)$. Similarly (and we omit the uninteresting details), the limit of each sequence of integrals can be replaced by the integral of the (almost everywhere) limit of the integrands.

Thus, almost everywhere,

$$R(x, y, t) = \tilde{R}(x, y, t) + \int_t^T \int ds\, dv\, dw K(x, y; v, w; t, s)R(v, w, s) + \int_t^T \int ds\, dv\, dw K(y, x; w, v; t, s)R(v, w, s), \tag{35}$$

where the kernel K is given by (34) with $R(v, x', s)$ replacing $R^n(v, x', s)$.

Consider the $\tilde{R}(x, y, t)$ term in (35). Since $Q(x, y, t)$ is symmetric, bounded, measurable and tends to zero (almost everywhere) as $x \to \partial D$ or $y \to \partial D$, $\tilde{R}(x, y, t)$ is Hölder continuous and tends to zero as $x \to \partial D$ or $y \to \partial D$ or $t \to T$. If a member of this latter class is substituted for the $Q(x, y, t)$, then $\tilde{R}(x, y, t) \in \hat{C}^0_{2+\alpha}$. Similarly, the map K (the sum of the integrals in (35)) takes $R(x, y, s)$ into a Hölder continuous function which tends to zero as $x \to \partial D$, $y \to \partial D$ or $t \to T$. If a member of this class is substituted for the $R(x, y, s)$ in the kernel K, then the sum of the integrals in (35) is in $\hat{C}^0_{2+\alpha}$. (Recall the identical assertion in the proof of Theorem 4.1.) These considerations imply that $R(x, y, t)$ is indeed in $\hat{C}^0_{2+\alpha}$.

Upon differentiating (35) and using $G_t(x, x'; t, s) = -\mathfrak{L}_x{}^* G(x, x'; t, s)$, we obtain (26).

Now, note that the $u(x, t)$ in (25) is the $v(x, t)$ which minimizes (19) and (20). Thus, letting $\bar{v}(x, t)$ be a control of the form (25a), we see that (19) yields

$$\begin{aligned} E_\phi{}^u \int_t^T d^u C^u(W(x, s), s) + E_\phi{}^u \int_t^T \int ds\, dx\, dy S(x, y, s)W(x, s)W(y, s) \\ + E_\phi{}^u \int_t^T \int ds\, dx P(x, s)u^2(x, s) \\ \leqq E_\phi{}^{\bar{v}} \int_t^T d^{\bar{v}} C^u(W(x, s), s) + E_\phi{}^{\bar{v}} \int_t^T \int ds\, dx\, dy S(x, y, s)W(x, s)W(y, s) \\ + E_\phi{}^{\bar{v}} \int_t^T \int ds\, dx P(x, s)\bar{v}^2(x, s). \end{aligned} \tag{36}$$

Since the first terms on the left and right of (36) are equal, (36) implies $C^u(\phi, t) \leqq C^{\bar{v}}(\phi, t)$.

REFERENCES

[1] H. J. KUSHNER, *An application of Sobolev's imbedding theorem to criteria for the continuity of vector parameter processes*, Ann. Math. Statist., to appear.

[2] A. BLANC-LAPIERRE AND R. FORTET, *Théorie des fonctions aléatoires*, Masson et Cie, Paris, 1953.

[3] J. KAMPÉ DE FERIET, *Random solutions of partial differential equations*, Proc. Third Berkeley Symposium on Probability and Statistics, vol. 3, University of California Press, Berkeley, 1960, pp. 199–208.

[4] P. K. C. WANG, *Control of distributed parameter systems*, Advances in Control Systems, vol. 1, C. T. Leondes, ed., Academic Press, New York, 1964, pp. 75–172.

[5] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

ALGEBRAIC STRUCTURE OF GENERALIZED POSITIVE REAL MATRICES*

B. D. O. ANDERSON AND J. B. MOORE†

Abstract. Square matrices $Z(\cdot)$ of real rational functions of a complex variable are considered with two properties: (1) $Z(\infty)$ has finite elements; (2) $Z(j\omega) + Z'(-j\omega)$ is nonnegative definite Hermitian for all real ω, other than those for which $j\omega$ is a pole of an element of $Z(\cdot)$. Necessary and sufficient conditions for the nonnegativity property are derived which involve the existence of constant matrices satisfying several algebraic equations. The work thereby extends earlier results on the structure of rational positive real matrices.

1. Introduction. This paper investigates the structure of a class of matrices occurring in various systems theory problems. We shall work with $n \times n$ matrices $Z(\cdot)$ of real rational functions of a complex variable s; the matrices of particular interest are those for which $Z(\infty) < \infty$ (that is, no element of $Z(\cdot)$ has a pole at infinity) and for which $Z(j\omega) + Z'(-j\omega)$ is a nonnegative definite Hermitian matrix for all real ω with $j\omega$ not a pole of any element of $Z(\cdot)$.

The so-called positive real matrices [1] possess the aforementioned properties, but also possess additional properties restricting the nature of poles of the matrix elements. The structure of such matrices has been investigated from the systems theory point of view [2], [3], and applications of the structure properties have been discussed [4]. There are, however, system theoretic problems involving matrices $Z(\cdot)$ with the finite-at-infinity constraint and the $j\omega$-axis nonnegativity constraint, but without the additional constraints imposed by $Z(\cdot)$ being positive real.

We shall term such matrices *generalized positive real.* Examples of problems involving generalized positive real matrices as distinct from positive real matrices may be found in [5], which discusses system instability, and in [6] and [7], which discuss inverse optimal control problems.

In [5], a single-input, single-output, time-invariant, finite-dimensional system is considered, with a time-varying feedback gain coupling the output to the input. The Nyquist plot of the open-loop system is supposed not to intersect a certain disk in the complex plane; as a consequence of this, a certain scalar function $z(\cdot)$ of a complex variable s is generalized positive real. The number of encirclements of the disk by the Nyquist plot determines whether $z(\cdot)$ is or is not positive real. When it is not positive real, an instability criterion is deduced.

To deal more effectively with systems theory problems involving rational

* Received by the editors October 31, 1967, and in revised form April 19, 1968.

† Department of Electrical Engineering, University of Newcastle, New South Wales 2308, Australia.

generalized positive real matrices, we shall give a structural description parallel to that which is known for positive real matrices [3].

Section 2 presents the main result of the paper (see Theorem 1) while §3 discusses several implications of the results.

2. Main results. In this section we consider square matrices $Z(\cdot)$ of real rational functions of a complex variable s subject to the constraint that $Z(\infty) < \infty$. A nonnegativity constraint will also be used extensively:

$$Z(j\omega) + Z'(-j\omega) \geqq 0 \quad \text{for almost all}^1 \text{ real } \omega \tag{1}$$

(the notation $A \geqq B$ [$A > B$] for Hermitian A and B means $A - B$ is nonnegative [positive] definite).

Before stating results for matrices $Z(\cdot)$ satisfying the above constraints, two results on rational positive real matrices will be reviewed.

LEMMA 1. *Suppose $Z(\cdot)$ is an $n \times n$ matrix of real rational functions of a complex variable s, with $Z(\infty) < \infty$. Suppose it is decomposed in the form* [3]

$$Z(s) = J + H'(sI - F)^{-1}G, \tag{2}$$

where F, G, H, J are real constant matrices. Then $Z(\cdot)$ is positive real if (1) *holds and all eigenvalues of F have negative real parts.*

Proof. This lemma is but a restatement of the definition of a positive real property.

LEMMA 2. *Let $Z(\cdot)$ be an $n \times n$ matrix of real rational functions of a complex variable s such that $Z(\infty) < \infty$. Let $\{F, G, H, J\}$ be a realization for $Z(\cdot)$, that is, a quadruple for which* (2) *holds. Suppose further that $[F, G]$ is completely controllable* [8] *and $[F, H]$ completely observable; then a necessary and sufficient condition for $Z(\cdot)$ to be positive real is that there exist real matrices $P = P' > 0$, L, W_0 such that*

$$PF + F'P = -LL', \tag{3a}$$

$$PG = H - LW_0, \tag{3b}$$

$$W_0'W_0 = J + J'. \tag{3c}$$

This lemma is proved in [3].

The following extended form of Lemma 2, relaxing simultaneously the complete observability requirement and the nonsingularity of P requirement, will be required in the sequel.

LEMMA 3. *With the same hypothesis as in Lemma 2, save for requiring the complete observability of $[F, H]$, $Z(\cdot)$ is positive real if and only if there exist real matrices $P = P' \geqq 0$, L, W_0 such that* (3a, b, c) *hold.*

[1] We exclude those ω for which $j\omega$ is a pole of some element of $Z(\cdot)$.

Proof. Necessity follows as in [3]. Sufficiency only will be established here. Select a coordinate transformation T (see [8]) such that

$$TFT^{-1} = \begin{bmatrix} F_{11} & 0 \\ F_{21} & F_{22} \end{bmatrix}, \tag{4a}$$

$$H'T^{-1} = [H_1' \quad 0], \tag{4b}$$

$$TG = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}, \tag{4c}$$

with $[F_{11}, H_1]$ completely observable. Let $\hat{P}$, $\hat{L}$, W_0 be such that

$$\hat{P}F_{11} + F_{11}'\hat{P} = -\hat{L}\hat{L}', \tag{5a}$$

$$\hat{P}G_1 = H_1 - \hat{L}W_0, \tag{5b}$$

$$W_0'W_0 = J + J', \tag{5c}$$

with $\hat{P} = \hat{P}' > 0$. Existence of $\hat{P}$, $\hat{L}$, W_0 is guaranteed by Lemma 2 and the fact that $\{F_{11}, G_1, H_1, J\}$ is a completely controllable, completely observable realization of $Z(\cdot)$. Then it is straightforward to check that

$$P = T'\begin{bmatrix} \hat{P} & 0 \\ 0 & 0 \end{bmatrix} T, \tag{6a}$$

$$L = T'\begin{bmatrix} \hat{L} \\ 0 \end{bmatrix}, \tag{6b}$$

and W_0 satisfy (3a, b, c), where the zero blocks in (6a) augment $\hat{P}$ to be of the same dimension as F, and the zero block in (6b) augments the number of rows of $\hat{L}$ to equal the dimension of F.

The extension of Lemmas 2 and 3 to generalized positive real matrices, covered in Theorem 1 below, relies on associating with a generalized positive real $Z(s)$ a positive real $Y(s)$, applying Lemmas 2 and 3 to $Y(s)$ to conclude the existence of certain matrices, and then defining a set of matrices to be associated with $Z(s)$ using those associated with $Y(s)$.

THEOREM 1. *Let $Z(\cdot)$ be an $n \times n$ matrix of real rational functions of a complex variable s such that $Z(\infty) < \infty$. Let $\{F, G, H, J\}$ be a realization of $Z(s)$ with $[F, G]$ completely controllable and $[F, H]$ completely observable. Then a necessary and sufficient condition for*

$$Z(j\omega) + Z'(-j\omega) \geqq 0 \tag{1}$$

to hold for all real ω for which $j\omega$ is not a pole of any element of $Z(\cdot)$ is that there exist real matrices $P = P'$, $\det P \neq 0$, L and W_0 such that

$$PF + F'P = -LL', \tag{3a}$$

$$PG = H - LW_0, \tag{3b}$$

$$W_0'W_0 = J + J'. \tag{3c}$$

Moreover, P is not positive definite if $Z(\cdot)$ is not positive real.

The proof of sufficiency is relatively simple; the proof of necessity harder. We turn to the former first.

Proof of sufficiency. Explicit calculation yields

$$
\begin{aligned}
& Z(j\omega) + Z'(-j\omega) \\
&= J + J' + H'(j\omega I - F)^{-1}G + G'(-j\omega I - F')^{-1}H \\
&= J + J' + G'[P(j\omega I - F)^{-1} + (-j\omega I - F')^{-1}P]G \\
&\quad + W_0'L'(j\omega I - F)^{-1}G + G'(-j\omega I - F')^{-1}LW_0 \qquad \text{(using (3b))} \\
&= J + J' + G'(-j\omega I - F)^{-1}[-PF - F'P](j\omega I - F)^{-1}G \\
&\quad + W_0'L'(j\omega I - F)^{-1}G + G'(-j\omega I - F')^{-1}LW_0 \qquad (7) \\
&= W_0'W_0 + G'(-j\omega I - F')^{-1}LL'(j\omega I - F)^{-1}G \\
&\quad + W_0'L'(j\omega I - F)^{-1}G + G'(-j\omega I - F')^{-1}LW_0 \\
&\qquad\qquad \text{(using (3a), (3c))} \\
&= [W_0' + G'(-j\omega I - F')^{-1}L][W_0 + L'(j\omega I - F)^{-1}G].
\end{aligned}
$$

Because of the form of the right-hand side, (1) is established.

Proof of necessity. We begin by defining the matrix

$$S(s) = [Z(s) - I][Z(s) + I]^{-1}. \tag{8}$$

It is not hard to verify that if

$$S(s) = J_s + H_s'(sI - F_s)^{-1}G_s, \tag{9}$$

the matrices F, G, H, J and F_s, G_s, H_s, J_s are related through the invertible equations

$$F = F_s + G_s(I - J_s)^{-1}H_s', \tag{10a}$$

$$G = 2G_s(I - J_s)^{-1}, \tag{10b}$$

$$H' = (I - J_s)^{-1}H_s', \tag{10c}$$

$$J = -I + 2(I - J_s)^{-1}. \tag{10d}$$

Note. Because (1) holds and because $Z(\infty) < \infty$, the matrix $I + J$ or $I + Z(\infty)$ is nonsingular. This means that J_s is well-defined as $I - 2[I + J]^{-1}$, and precisely because of the way J_s is defined, $I - J_s = 2[I + J]^{-1}$ is nonsingular. These two facts guarantee that all quantities in (9) are well-defined or, equivalently, that (10a, b, c, d) are invertible.

The following sequence of implications should be noted:

$[F, G]$ is completely controllable implies $[F, \frac{1}{2}G(I - J_s)]$ or $[F, G_s]$ is completely controllable.

implies $[F - G_S K_S', G_S]$ is completely controllable for any K_S of appropriate dimension.

implies $[F_S, G_S]$ is completely controllable, taking $K_S' = (I - J_S)^{-1} H_S'$.

It is also not hard to verify the following formula:

$$Z(s) + Z'(-s) = \tfrac{1}{2}[I + Z'(-s)][I - S'(-s)S(s)][I + Z(s)]. \tag{11}$$

Equations (1) and (11) together imply that

$$I = S'(-j\omega)S(j\omega) \geqq 0 \tag{12}$$

for all real ω.

Now let a matrix K be chosen so that the eigenvalues of the matrix $F_S - G_S K'$ all possess negative real part. Such a K always exists when $[F_S, G_S]$ is completely controllable (see [9]). (Note that it may be possible to choose $K = 0$.)

Define

$$Q(s) = S(s)R(s), \tag{13}$$

where

$$R(s) = I - K'(sI - F_S + G_S K')^{-1} G_S. \tag{14}$$

Then simple manipulation yields

$$Q(s) = J_S + (H_S' - J_S K')(sI - F_S + G_S K')^{-1} G_S. \tag{15}$$

Equations (12) and (13) also imply

$$R'(-j\omega)R(j\omega) - Q'(-j\omega)Q(j\omega) \geqq 0 \quad \text{for all real} \quad \omega, \tag{16}$$

which, in full, is

$$\begin{aligned} &(I - J_S' J_S) - [K' + J_S'(H_S' - J_S K')](j\omega I - F_S + G_S K')^{-1} G_S \\ &- G_S'(-j\omega I - F_S' + K G_S')^{-1}[K + (H_S - K J_S')J_S] \\ &+ G_S'(-j\omega I - F_S' + K G_S')^{-1}[KK' - (H_S - K J_S')(H_S' - J_S K')] \\ &\cdot (j\omega I - F_S + G_S K')^{-1} G_S \geqq 0 \quad \text{for all real} \quad \omega. \end{aligned} \tag{17}$$

Define now the matrix P_Q as the unique symmetric solution of

$$\begin{aligned} P_Q(F_S - G_S K') + (F_S' - K G_S')P_Q \\ = KK' - (H_S - K J_S')(H_S' - J_S K'). \end{aligned} \tag{18}$$

The eigenvalue restriction on $F_S - G_S K'$ guarantees the existence of a unique and symmetric P_Q satisfying (18) (see [10]). Then (17) becomes,

using manipulations like those used in deducing (7),

$$Y(j\omega) + Y'(-j\omega) \geqq 0 \quad \text{for all real} \quad \omega, \tag{19}$$

where, if $Y(s) = J_Y + H_Y'(sI - F_Y)^{-1}G_Y$,

$$G_Y = G_S, \tag{20a}$$

$$F_Y = F_S - G_S K', \tag{20b}$$

$$H_Y = -K - H_S J_S + K J_S' J_S - P_Q G_S, \tag{20c}$$

$$J_Y = \tfrac{1}{2}(I - J_S' J_S). \tag{20d}$$

Because $F_S - G_S K' = F_Y$ has all eigenvalues with negative real parts, (19) implies by Lemma 1 that $Y(\cdot)$ is positive real. Lemma 3 may therefore be applied to yield the existence of matrices $P_Y = P_Y' \geqq 0$, L_Y and W_{0Y} for which

$$P_Y F_Y + F_Y' P_Y = -L_Y L_Y', \tag{21a}$$

$$P_Y G_Y = H_Y - L_Y W_{0Y}, \tag{21b}$$

$$W_{0Y}' W_{0Y} = J_Y + J_Y'. \tag{21c}$$

For convenience, these may be rewritten, using (20a, b, c, d), as

$$P_Y F_S + F_S' P_Y = -L_Y L_Y' + P_Y G_S K' + K G_S' P_Y, \tag{22a}$$

$$P_Y G_S = -K - H_S J_S + K J_S' J_S - P_Q G_S - L_Y W_{0Y}, \tag{22b}$$

$$W_{0Y}' W_{0Y} = I - J_S' J_S. \tag{22c}$$

Recapitulating, we have passed from $Z(s)$ to $S(s)$, thence to $R(s)$ and $Q(s)$, and finally to $Y(s)$. The quantities of further interest are, in order of their definition, F, G, H and J, then F_S, G_S, H_S, J_S (related to F, G, H and J via (10a, b, c, d), K and P_Q (here (18) is relevant), and finally P_Y, L_Y and W_{0Y} (see (22a, b, c)).

We now claim that matrices P, L and W_0 satisfying (3a, b, c) are given by

$$P = \tfrac{1}{2}(P_Q + P_Y), \tag{23a}$$

$$L = (1/\sqrt{2})[L_Y + K W_{0Y}' + H_S(I - J_S')^{-1} W_{0Y}'], \tag{23b}$$

$$W_0 = \sqrt{2} W_{0Y}(I - J_S)^{-1}. \tag{23c}$$

Equation (3c) is easy to prove using (10d), (22c) and (23c); to prove (3a) and (3b) requires some manipulation, an outline of which will now be given.

Using (10a, b, c, d) and (23a, b, c), we have

$$
\begin{aligned}
PG - H + LW_0 = {} & (P_Q + P_Y)G_S(I - J_S)^{-1} - H_S(I - J_S')^{-1} \\
& + L_Y W_{0Y}(I - J_S)^{-1} + KW_{0Y}'W_{0Y}(I - J_S)^{-1} \\
& + H_S(I - J_S)^{-1}W_{0Y}'W_{0Y}(I - J_S)^{-1}.
\end{aligned} \tag{24}
$$

Now if (22b) and (22c) are used to substitute for $P_Y G_S$ and $W_{0Y}'W_{0Y}$, and all possible cancellations made, the right-hand side of (24) becomes zero. This proves (3b). At the same time, (10a, b, c, d) and (23a, b, c) give

$$
\begin{aligned}
PF + F'P + LL' = {} & \tfrac{1}{2}(P_Q + P_Y)[F_S + G_S(I - J_S)^{-1}H_S'] \\
& + \tfrac{1}{2}[F_S + H_S(I - J_S')^{-1}G_S'](P_Q + P_Y) \\
& + \tfrac{1}{2}[L_Y + KW_{0Y}' + H_S(I - J_S')^{-1}W_{0Y}][L_Y' + W_{0Y}K' \\
& + W_{0Y}'(I - J_S)^{-1}H_S'].
\end{aligned} \tag{25}
$$

The first and second terms on the right side of (25) may be manipulated to yield

$$
\begin{aligned}
& 2(PF + F'P + LL') \\
& = P_Q(F_S - G_S K') + (F_S' - KG_S')P_Q + P_Q G_S K' + KG_S' P_Q \\
& \quad + P_Q G_S(I - J_S)^{-1}H_S' + H_S(I - J_S')^{-1}G_S' P_Q \\
& \quad + P_Y F_S + F_S' P_Y + P_Y G_S(I - J_S)^{-1}H_S' + H_S(I - J_S')^{-1}G_S' P_Y \\
& \quad + [L_Y + KW_{0Y}' + H_S(I - J_S')^{-1}W_{0Y}][L_Y' + W_{0Y}K' \\
& \quad + W_{0Y}'(I - J_S)^{-1}H_S'].
\end{aligned} \tag{26}
$$

Equation (18) eliminates P_Q from the first two terms. Equations (22a) and (22b) eliminate P_Q and P_Y from the next eight terms. What is left is then an expression involving K, H_S, J_S, G_S, L_Y and W_{0Y}. Using (22c) causes the right-hand side then to equal zero.

Next, the nonsingularity of P will be demonstrated. The symmetry of P follows from (23a) and the symmetry of P_Q and P_Y. Suppose P is singular, so that there exists a nonsingular T for which

$$
\tilde{P} = T'PT = \begin{bmatrix} I_r & 0 & 0 \\ 0 & -I_s & 0 \\ 0 & 0 & 0_t \end{bmatrix}, \tag{27}
$$

where I_r is the $r \times r$ unit matrix and 0_t a zero $t \times t$ matrix ($t > 0$). Set $\tilde{F} = T^{-1}FT$, $\tilde{G} = T^{-1}G$, $\tilde{H} = H'T$, $\tilde{L}' = L'T$. Then (3a) and (3b) become

$$
\tilde{P}\tilde{F} + \tilde{F}'\tilde{P} = -\tilde{L}\tilde{L}', \tag{28a}
$$

$$
\tilde{P}\tilde{G} = \tilde{H} - \tilde{L}W_0. \tag{28b}
$$

Partition $\tilde{F}$, $\tilde{G}$, $\tilde{H}$ and $\tilde{L}$ as

$$\tilde{F} = \begin{bmatrix} F_{rr} & F_{rs} & F_{rt} \\ F_{sr} & F_{ss} & F_{st} \\ F_{tr} & F_{ts} & F_{tt} \end{bmatrix}, \tag{29a}$$

$$\tilde{G} = \begin{bmatrix} G_r \\ G_s \\ G_t \end{bmatrix}, \tag{29b}$$

$$H = \begin{bmatrix} H_r \\ H_s \\ H_t \end{bmatrix}, \tag{29c}$$

$$L = \begin{bmatrix} L_r \\ L_s \\ L_t \end{bmatrix}. \tag{29d}$$

Then it is easily checked that (27) and (28a) force F_{rt}, F_{st}, F_{tr}, F_{ts} and L_t to equal zero. Equation (28b) then forces H_t to equal zero, which is incompatible with the complete observability of the pair

$$\begin{bmatrix} F_{rr} & F_{rs} & 0 \\ F_{st} & F_{ss} & 0 \\ 0 & 0 & F_{tt} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} H_r \\ H_s \\ H_t \end{bmatrix}.$$

Finally, note that if P is positive definite, this fact, together with (3a, b, c), implies the positive real nature of $Z(\cdot)$ (see [3]). Thus if $Z(\cdot)$ is not positive real, P is not positive definite.

Just as Lemma 3 extends the result of Lemma 2, so the following extension of Theorem 1 is possible.

COROLLARY. *With the same hypothesis as Theorem* 1, *save for the requirement that* $[F, H]$ *be completely observable, a necessary and sufficient condition for*

$$Z(j\omega) + Z'(-j\omega) \geqq 0 \tag{1}$$

for all real ω, *where* $j\omega$ *is not a pole of any element of* $Z(\cdot)$, *is that there exist real matrices* $P = P'$, L, W_0 *such that* (3a, b, c) *hold. Moreover,* $Z(\cdot)$ *is positive real if and only if* P *is nonnegative or positive definite.*

3. Concluding remarks. It is possible to give a simple frequency domain interpretation of the basic equations (3a, b, c). Defining

$$W(s) = W_0 + L'(sI - F)^{-1}G, \tag{30}$$

we have, using arguments appearing in the proof of sufficiency for Theorem

1, that

$$Z(j\omega) + Z'(-j\omega) = W'(-j\omega)W(j\omega). \tag{31}$$

The determination for a prescribed $Z(\cdot)$ of a $W(\cdot)$ satisfying (31) is termed spectral factorization. As is discussed in, for example, [11], for a prescribed $Z(\cdot)$, there are many possible $W(\cdot)$ satisfying (31); here we have elected to find a spectral factor $W(\cdot)$ which not only has the same poles as $Z(\cdot)$, but which can have two matrices of a realizing quadruple identical with those of $Z(\cdot)$.

It is of interest to observe how P, L and W_0 in (3a, b, c) may be calculated, given F, G, H and J. Section 2 shows how the determination of P for a generalized positive real $Z(\cdot)$ can be made to depend on the determination of P for a positive real $Z(\cdot)$, which is discussed in [12] and [13]; the former reference shows how to determine P by solving a quadratic matrix equation, while the latter determines P as the limiting solution of a matrix Riccati differential equation.

When P in (3a, b, c) has been found, the determination of L and W_0 proves straightforward.

For stability and instability studies, the positive definiteness or lack of positive definiteness of the P matrix becomes important, since Lyapunov functions for systems with which a generalized positive real matrix is associated may well have a term $x'Px$ appearing in them. For examples, [5] and [14] can be consulted.

In inverse optimal control problems (see [6] and [7]) typically an equation such as (31) has to be solved, with the constraint that with $Z(s)$ of the form $J + H'(sI - F)^{-1}G$, then $W(s)$ should have the form $W_0 + L'(sI - F)^{-1}G$; usually L has to be found, and the preceding two sections exhibit procedures for this.

REFERENCES

[1] R. W. Newcomb, *Linear Multiport Synthesis*, McGraw-Hill, New York, 1966.

[2] R. E. Kalman, *On a new characterization of linear passive systems*, Proc. First Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1963, pp. 456–470.

[3] B. D. O. Anderson, *A system theory criterion for positive real matrices*, this Journal, 5 (1967), pp. 171–182.

[4] ———, *Development and applications of a system theory criterion for rational positive real matrices*, Proc. Fourth Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1966, pp. 400–407.

[5] R. W. Brockett and H. B. Lee, *Frequency domain instability criteria for time-varying and nonlinear systems*, Proc. IEEE, 55 (1967), pp. 604–619.

[6] R. E. Kalman, *When is a linear control system optimal?* Trans. ASME Ser. D. J. Basic Engrg., 86 (1964), pp. 1–10.

[7] B. D. O. Anderson, *The inverse problem of optimal control*, Rep. SEL-66-038 (TR no. 6560-3), Stanford Electronics Laboratories, Stanford, California, 1966.

[8] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.
[9] B. D. O. ANDERSON AND D. G. LUENBERGER, *Design of multivariable control systems*, Proc. Inst. Elec. Engrs., 114 (1967), pp. 395–399.
[10] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1969.
[11] D. C. YOULA, *On the factorization of rational matrices*, IRE Trans. Information Theory, IT-7 (1961), pp. 172–189.
[12] B. D. O. ANDERSON, *An algebraic solution to the spectral factorization problem*, Trans. IEEE Automatic Control, AC-12 (1967), pp. 410–414.
[13] ———, *Quadratic minimization, positive real matrices and spectral factorization*, submitted for publication.
[14] ———, *Stability of control systems with multiple nonlinearities*, J. Franklin Inst., 281 (1966), pp. 155–160.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

STABILITY ANALYSIS OF NONLINEAR AND TIME-VARYING DISCRETE SYSTEMS*

KUMPATI S. NARENDRA† AND YO-SUNG CHO‡

1. Introduction. Ever since Popov [1] derived frequency domain conditions for the stability of continuous feedback systems, considerable work has been done to derive similar criteria for the determination of the stability of nonlinear sampled-data systems. The class of systems generally treated contains a linear time-invariant discrete system and a single no-memory nonlinear or time-varying function. Starting with the work of Tsypkin [2a], [2b], one can derive many interesting and practically applicable frequency domain stability criteria for both autonomous and nonautonomous discrete feedback systems. Significant contributions were made by Szegö, Jury, Lee, Pearson and Gibson [3]–[6] who derived less conservative conditions on the linear part of the system than did Tsypkin [2] by placing constraints on the slope of the nonlinearity. The most general stability conditions known to date as a consequence of the work of these authors can be expressed in terms of the linear part $G(z)$ and the nonlinear function $F(\cdot)$ in Fig. 1 as follows: If

$$F(0) = 0, \quad 0 < \frac{F(y)}{y} < K_2 \quad \text{and} \quad \frac{dF(y)}{dy} < K', \tag{1.1}$$

then

$$\frac{1}{K_2} + \operatorname{Re}[1 + q(z-1)]G(z) - \frac{K'}{2}|q|\,|(z-1)G(z)|^2 \geqq 0$$

for all $|z| = 1$ is a sufficient condition for the stability of the feedback system. In 1963, Tsypkin [6], after pointing out the difficulty involved in determining whether or not the inequality (1.1) can be satisfied, suggested a simpler condition (later extended by Jury and Lee [4b], [4c]) for the case of slope restricted monotonic functions in the feedback path. For this case if $0 < F(y)/y < K_2$ and $0 < dF(y)/dy < \infty$, Tsypkin's condition may be stated as

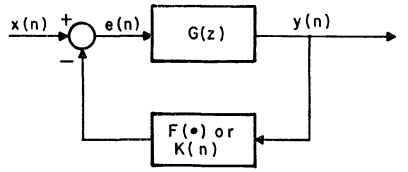
$$\operatorname{Re} G(z)\left[1 + q\left(\frac{z-1}{z}\right)\right] + \frac{1}{K_2} \geqq 0 \quad \text{for all} \quad |z| = 1. \tag{1.2}$$

In this paper several new results are presented for the stability of discrete

* Received by the editors September 22, 1967, and in revised form March 11, 1968.

† Department of Engineering and Applied Science, Dunham Laboratory, Yale University, New Haven, Connecticut 06520.

‡ Honeywell Incorporated, Waltham, Massachusetts 02154.



G(z): LINEAR TIME-INVARIANT DISCRETE OPERATOR
(z IS A z - TRANSFORM VARIABLE)

F(•) OR K(n): NONLINEAR FUNCTION OR TIME - VARYING GAIN FUNCTION

FIG. 1. *Nonlinear time-varying discrete feedback system*

feedback systems with monotonic, odd monotonic or time-varying gain functions in the feedback path. In all cases a "passive operator" technique is used to generate classes of multipliers $M(z)$ such that the frequency domain stability criteria derived have, in general, the form

$$\text{(1.3)} \quad \text{Re}\left[G(z) + \frac{1}{b}\right] M(z) + \frac{1}{K_2} - \frac{1}{b} \geqq 0 \quad \text{for all} \quad |z| = 1,$$

where

$$0 < \frac{F(y)}{y} < K_2 \quad \text{and} \quad 0 < \frac{F(y_1) - F(y_2)}{y_1 - y_2} < b.$$

Since, in general, the frequency domain criteria are difficult to verify, equivalent geometrical criteria which by-pass the need for determining the multiplier $M(z)$ are derived. Examples are also provided to demonstrate the applicability of the results.

The scope of the paper may be divided into three parts:

(i) The determination of frequency domain conditions for the stability of discrete feedback systems with monotonic and odd monotonic nonlinear functions. (For the monotonic case, Tsypkin's result in [6] is found to be a special case of the results derived here.)

(ii) The determination of frequency domain conditions for the stability of linear discrete feedback systems with time-varying gains. (The bound on the slope of the time-varying gain is found to depend on the location of the singularities of $M(z)$ which makes $\text{Re}\, G(z)M(z) + 1/K_2 \geqq 0$ for all $|z| = 1$.)

(iii) The derivation of geometrical criteria to simplify the application of criteria (i) and (ii) so that stability of the feedback system can be assured from either a plot of $G(z)$ for all $|z| = 1$, or the root locus plot of $G(z)$ for constant gains K in the feedback path in the range $0 < K < K_2$.

The criteria derived in the paper are also applied to examples treated by other authors for comparison purposes.

2. Definitions and mathematical preliminaries. In this section, the definitions required to prove the stability criteria in §3 and §4 are stated briefly.

2.1. Function spaces. The input and output functions in this paper are assumed to belong to one of the two spaces (a) and (b) of real-valued sequences defined for all nonnegative integers:

(a) l_2-space: Let X be the set of sequences x of real numbers $x(n)$ defined for all nonnegative integers n. Then $l_2 = \{x \mid x \in X, \sum_{n=0}^{\infty} x(n)^2 < \infty\}$, and the norm is defined as $\| x \|_2 = \{\sum_{n=0}^{\infty} x(n)^2\}^{1/2}$.

(b) l_{2e}-space: Let x_ν denote the truncated sequence which assumes the values $x_\nu(n) = x(n)$ $(n \leqq \nu$ and $\nu \in N$, where N is the subinterval of nonnegative integers $[0, \infty))$ and $x_\nu(n) = 0$ elsewhere. Then $l_{2e} = \{x \mid x \in X, x_\nu \in l_2$, where $\nu \in N\}$, and $\| x \|_{2e} = \{\sum_{n=0}^{\infty} x(n)^2\}^{1/2}$.

Hence it should be noted that

$$\| x \|_{2e} = \begin{cases} \| x \|_2 & \text{if} \quad x \in l_2, \\ \infty & \text{if} \quad x \notin l_2. \end{cases}$$

l_{2e} may be considered to be an extension of l_2 which consists of sequences with bounded and unbounded norms.

2.2. Passive operators. The main tool used in this paper for determining the stability of nonlinear or time-varying discrete systems is the concept of a passive operator.

(a) Inner product: We define an inner product in l_2 by

$$\langle x, y \rangle = \sum_{n=0}^{\infty} x(n)y(n). \tag{2.1}$$

Since every truncated sequence x_ν, $y_\nu \in l_2$ if $x, y \in l_{2e}$, we can define

$$\langle x, y \rangle_\nu = \langle x_\nu, y_\nu \rangle = \sum_{n=0}^{\infty} x_\nu(n)y_\nu(n) = \sum_{n=0}^{\nu} x(n)y(n), \tag{2.2}$$

where $\nu \in N$.

(b) Passive operator: Let a discrete operator W map l_{2e} into itself. If $\langle x, W(x) \rangle_\nu \geqq 0$ for all $\nu \in N$ and for all $x \in l_{2e}$, then the operator W is said to be passive.

(c) δ-passive operator: If $y = W(x)$ and there exists a positive real number δ such that $\langle x, y \rangle_\nu \geqq \delta \langle x, x \rangle_\nu$, then the operator W is δ-passive.

(d) δ-M-passive operator: If there exists a positive real number δ such that $\langle x, y \rangle_\nu \geqq \delta \langle y, y \rangle_\nu$, then the operator is δ-M-passive.

2.3. Classes of nonlinear and time-varying functions. The nonlinear functions $F(\cdot)$ considered here are limited only to monotonic and odd monotonic real-valued functions with known slope bounds.

(a) If a real-valued function F satisfies:

(i) $F(y) = 0$ if and only if $y = 0$,

(ii) $K_1 < F(y)/y < K_2$, where $0 \leqq K_1 < K_2 < \infty$,

(iii) $a < [F(y_1) - F(y_2)]/(y_1 - y_2) < b$ (bounded slope), where $0 \leqq a < b < \infty$,

then it is denoted by $F \in F_m \triangleq F_m[(K_1, K_2), (a, b)]$. By inspection for all $F \in F_m$, F^{-1} exists and $F^{-1} \in F_m$. It is sometimes useful to consider the special case $F \in F_m{}^* \triangleq F_m[(0, K_2), (0, K_2)]$.

(b) If $F \in F_m$ is odd, i.e., $F(-y) = -F(y)$, then $F \in F_{0m} \triangleq F_{0m}[(K_1, K_2), (a, b)]$. Obviously $F \in F_{0m}$ implies $F^{-1} \in F_{0m}$ exists. Again a special case is $F \in F_{0m}^* \triangleq F_{0m}[(0, K_2), (0, K_2)]$.

(c) The linear time-varying gain function $K(n)$ satisfies the condition $0 < K(n) < \infty$.

2.4. Class of linear time-invariant discrete operators G. G is assumed to satisfy the condition

$$y(n) = G(x) = \sum_{m=0}^{n} g(n - m)x(m),$$

where an input sequence x of G is in l_{2e} and g is an impulse response function of G satisfying $\sum_{n=0}^{\infty} | g(n) | < \infty$ and $\sum_{n=0}^{\infty} g(n)^2 < \infty$.

2.5. Lemmas.

LEMMA 2.1. *If $F \in F_{0m}$, then for all real numbers y_1 and y_2,*

$$(y_1 + y_2)F(y_1) - (y_1 - y_2)F(y_2) \geqq 0 \tag{2.3}$$

(see [14]).

LEMMA 2.2. *If an operator W maps l_{2e} into itself and the z-transform of W is $W(z)$, then W is passive if the Fourier transform of W exists and*

$$\operatorname*{Re}_{-\pi<\omega<\pi} W(e^{i\bar{\omega}}) \geqq 0,$$

where $\bar{\omega} = \omega T$.

Proof. Assuming $T = 1$, let $x \in l_{2e}$ and $y = W(x)$.

$$\langle x, y\rangle_\nu = \sum_{n=0}^{\nu} x(n)y(n) = \sum_{n=0}^{\nu} x(n) \sum_{m=0}^{n} x(m)w(n - m) \quad \text{for all} \quad \nu \in N,$$

where w is the impulse response of W. Hence,

$$\langle x, y\rangle_\nu = \operatorname{Re} \frac{1}{2\pi} \int_{-\pi}^{\pi} \overline{X_\nu(e^{i\omega})} X_\nu(e^{i\omega})\ W(e^{i\omega})\ d\omega \geqq 0$$

by the Lyapunov-Parseval theorem [2c], where $X_\nu(e^{i\omega}) = \sum_{n=0}^{\nu} x(n)e^{-i\omega}$. Hence W is a passive operator.

LEMMA 2.3. *If* $F \in F_m$ *and* $u \in l_{2e}$, *then*

$$\sum_{n=0}^{\nu} F[u(n+1)]\{u(n+1) - \eta u(n)\} \geqq 0$$

for all $\nu \in N$, $0 \leqq \eta \leqq 1$ *and* $u(0) = 0$.

Proof. It will be shown by induction that

$$\sum_{n=0}^{\nu} F[u(n+1)]\{u(n+1) - \eta u(n)\} \geqq \int_0^{u(\nu+1)} F(\sigma)\, d\sigma \geqq 0 \tag{2.4}$$

for all $\nu \in N$.

If the inequality (2.4) is satisfied for $\nu = r$, i.e.,

$$\sum_{n=0}^{r} F[u(n+1)]\{u(n+1) - \eta u(n)\} \geqq \int_0^{u(r+1)} F(\sigma)\, d\sigma,$$

then for $\nu = r + 1$,

$$\begin{aligned}\sum_{n=0}^{r+1} F[u(n+1)]\{u(n+1) - \eta u(n)\} &\geqq \int_0^{u(r+1)} F(\sigma)\, d\sigma + F[u(r+2)]\{u(r+2) - \eta u(r+1)\} \\ &\geqq \int_0^{u(r+2)} F(\sigma)\, d\sigma.\end{aligned}$$

For $\nu = 1$, the result follows trivially; hence by induction, Lemma 2.3 is proved.

LEMMA 2.4. (a) *If* $F \in F_m$ *and* $M(z) = (z - \eta)/z$, $0 \leqq \eta \leqq 1$, *then* $M[F(\cdot)]$ *is a passive operator.*

(b) *If* $0 \leqq \eta < 1$, $M[F(\cdot)]$ *is* δ*-passive.*

Proof. (a) Let $u, v \in l_{2e}$ such that $v = M[F(u)]$. Let $w(n)$ be defined by the relationship $v(n) = w(n+1) - \eta w(n)$; then $u(n) = F^{-1}[w(n+1)]$. By Lemma 2.3,

$$\langle v, u\rangle_\nu = \sum_{n=0}^{\nu} \{w(n+1) - \eta w(n)\}\{F^{-1}w(n+1)\} \geqq 0 \tag{2.5}$$

for all $\nu \in N$ and $0 \leqq \eta \leqq 1$.

(b) If $0 \leqq \eta < 1$,

$$\begin{aligned}\langle v, u\rangle_\nu &= \eta\left(\frac{1}{\eta} - 1\right)\sum_{n=0}^{\nu} w(n+1)F^{-1}[w(n+1)] \\ &\quad + \eta \sum_{n=0}^{\nu} \{w(n+1) - w(n)\}F^{-1}[w(n+1)] \\ &\geqq \frac{\eta}{K_2}\left(\frac{1}{\eta} - 1\right)\sum_{n=0}^{\nu} w(n+1)\cdot w(n+1)\end{aligned}$$

for all $\nu \in N$.

Hence $M[F(\cdot)]$ is δ-passive.

LEMMA 2.5. *If* $F \in F_m$ *and* $M(z) = (z-1)/(z-\eta)$, $0 \leqq \eta \leqq 1$, *then* $M[F(\cdot)]$ *is a passive operator.*

Proof. Let $v = M[F(u)]$ and $v(n) = w(n+1) - w(n)$. Then

$$\langle u, M[F(u)]\rangle_\nu = \sum_{n=0}^{\nu} \{F^{-1}[w(n+1) - \eta w(n)]\}\{w(n+1) - w(n)\}. \tag{2.6}$$

Subtracting

$$\sum_{n=0}^{\nu} \{F^{-1}[(1-\eta)w(n+1)]\}\{w(n+1) - w(n)\} \tag{2.7}$$

from (2.6), we have, by inequality (2.2),

$$\sum_{n=0}^{\nu} \{w(n+1) - \eta w(n) - (1-\eta)w(n+1)\} \cdot \{F^{-1}[w(n+1) - \eta w(n)] - F^{-1}[(1-\eta)w(n+1)]\} \geqq 0.$$

However, (2.7) is nonnegative by Lemma 2.3, and hence (2.6) is also nonnegative.

LEMMA 2.6. *If* $F \in F_m$ *and*

$$M(z) = \prod_{i=1}^{m} a_i \frac{z-\gamma_i}{z-\eta_i}, \tag{2.8}$$

where $a_i \geqq 0$, $1 \geqq \gamma_i > \eta_i > \gamma_{i+1} \geqq 0$, *then* $M[F(\cdot)]$ *is passive.*

Proof. It is known that $M(z)$ can be expanded as

$$M(z) = \alpha_0 + \beta_0 \frac{z-1}{z} + \sum_{i=1}^{m} \alpha_i \frac{z-1}{z-\eta_i} \triangleq M_{(1)}(z) + M_{(2)}(z) + \sum_{i=1}^{m} M_i(z),$$

where all coefficients α_0, β_0, α_i are nonnegative.

Hence,

$$M[F(\cdot)] = M_{(1)}[F(\cdot)] + M_{(2)}[F(\cdot)] + \sum_{i=1}^{m} M_i[F(\cdot)]$$

is passive by Lemmas 2.4 and 2.5.

LEMMA 2.7. *If* $F \in F_{0m}$ *and*

$$M(z) = M_{(1)}(z) + M_{(2)}(z) + M_{(3)}(z) + M_{(4)}(z), \tag{2.9}$$

where

$$M_{(1)}(z) = \sum_{i=1}^{m} a_i \frac{z-\gamma_i'}{z-\eta_i'}, \quad a_i \geqq 0, \quad 1 \geqq \gamma_i' > \eta_i' \geqq 0 \quad \textit{for all } i,$$

$$M_{(2)}(z) = \sum_{j=1}^{n} b_j \frac{z-\gamma_j}{z-\eta_j}, \quad b_j \geqq 0, \quad 1 \geqq \eta_j > \gamma_j \geqq 0, \gamma_j > \frac{\eta_j}{2-\eta_j},$$

$$M_{(3)}(z) = \sum_{p=1}^{q} c_p \frac{z+\mu_p'}{z+\nu_p'}, \quad c_p \geqq 0, \quad 1 \geqq \nu_p' > \mu_p' \geqq 0,$$

$$M_{(4)}(z) = \sum_{r=1}^{s} d_r \frac{z+\mu_r}{z+\nu_r}, \quad d_r \geqq 0, \quad 1 \geqq \mu_r > \nu_r \geqq 0,$$

then $M[F(\cdot)]$ *is a passive operator.*

Proof. Since

$$M[F(\cdot)] = M_{(1)}[F(\cdot)] + M_{(2)}[F(\cdot)] + M_{(3)}[F(\cdot)] + M_{(4)}[F(\cdot)], \tag{2.10}$$

$M[F(\cdot)]$ is passive if each term of the right-hand side of (2.10)] is passive. By Lemma 2.6, $M_{(1)}[F(\cdot)]$ is passive. The passivity of $M_{(2)}[F(\cdot)]$ is proved by showing that

$$A \equiv \langle M_{2j}[F(u)], u\rangle_\nu \geqq 0 \quad \text{for all} \quad j \text{ and } \nu \in N, \tag{2.11}$$

where $M_{(2)}(z) = \sum_{j=1}^{n} M_{2j}(z)$. Defining $w(n)$ by the relation

$$F[u(n)] = w(n+1)/\eta_j - w(n),$$

we see that (2.11) becomes

$$A = \frac{b_j \gamma_j}{\eta_j} \sum_{n=0}^{\nu} \left\{ F^{-1}\left[\frac{w(n+1)}{\eta_j} - w(n)\right]\right\} \left\{\frac{w(n+1)}{\gamma_j} - w(n)\right\}. \tag{2.12}$$

If the hypotheses concerning the coefficients are satisfied, the following is a nonnegative quantity by Lemma 2.3:

$$\frac{b_j \gamma_j}{\eta_j} \sum_{n=0}^{\nu} \left\{ F^{-1}\left[\left(\frac{1}{\gamma_j} - \frac{1}{\eta_j}\right) w(n+1)\right]\right\} \cdot \left\{\left(\frac{2}{\eta_j} - \frac{1}{\gamma_j}\right) w(n+1) - w(n)\right\}. \tag{2.13}$$

Subtracting this nonnegative quantity (2.13) from (2.12), one sees that

$$A \geqq \sum_{n=0}^{\nu} \{(y_1 + y_2)F^{-1}(y_1) - (y_1 - y_2)F^{-1}(y_2)\} \quad \text{for all} \quad \nu \in N, \tag{2.14}$$

where

$$y_1(n) = \frac{w(n+1)}{\eta_j} - w(n) \quad \text{and} \quad y_2(n) = \left(\frac{1}{\gamma_j} - \frac{1}{\eta_j}\right) w(n+1).$$

However, since $F^{-1} \in F_{0m}$, (2.14) is nonnegative by Lemma 2.1, and hence (2.12) is also nonnegative. Next consider

$$\langle M_{(3)}[F(u)], u\rangle_\nu = \sum_{p=1}^{q} \langle M_{3p}[F(u)], u\rangle_\nu,$$

where $M_{3p}(z) = c_p(z + \mu_p')/(z + \nu_p')$. Defining $w(n)$ by the relation $F[u(n)] = w(n+1) + \nu_p' w(n)$, we have

$$\begin{aligned} &\langle M_{3p}[F(u)], u\rangle_\nu \\ &= c_p \frac{\mu_p'}{\nu_p'} \sum_{n=0}^{\nu} \{w(n+1) + \nu_p' w(n)\}\{F^{-1}[w(n+1) + \nu_p' w(n)]\} \\ &\quad + c_p\left(1 - \frac{\mu_p'}{\nu_p'}\right) \sum_{n=0}^{\nu} w(n+1)\{F^{-1}[w(n+1) + \nu_p' w(n)]\}. \end{aligned} \tag{2.15}$$

The first term on the right-hand side of (2.15) is greater than zero by inspection. The second term is greater than or equal to zero if $\mu_p'/\nu_p' < 1$ and $\nu_p' \leqq 1$; hence $M_{(3)}[F(\cdot)]$ is a passive operator.

The passivity of $M_{(4)}[F(\cdot)]$ can be proved in a similar manner.

Comments on the multipliers $M(z)$ in Lemmas 2.5 *and* 2.6. $M(z)$ in Lemma 2.5 has alternating poles and zeros on the real axis of the z-plane in the interval [0, 1] with the singularity nearest to the point (1, 0) being a zero. Hence $\arg M(z) = \phi$ (where $z = e^{i\bar{\omega}}$, $\bar{\omega} = \omega T$ and T is the sampling period) lies in the interval $0 \leqq \phi \leqq -\frac{1}{2}\bar{\omega} + \frac{1}{2}\pi$, where $0 \leqq \bar{\omega} \leqq \pi$ (see [8, p. 440]).

The multiplier $M(z)$ in Lemma 2.6 is more general than that of Lemma 2.5. $M_{(2)}(z)$ also has alternating poles and zeros, but the singularity nearest to the point (1, 0) is a pole, so $M_{(1)}(z) + M_{(2)}(z)$ may have complex zeros.

$M_{(3)}(z)$ has alternating real poles and real zeros in the interval $[-1, 0]$ with the zero nearest to the origin. Finally, $M_{(4)}(z)$ has the form of the inverse of $M_{(3)}(z)$. Hence $M_{(4)}(z)$ and $M_{(3)}(z)$ together can have complex zeros. If one uses techniques similar to those used in Lemmas 2.4 and 2.5 of [7a], it is possible to derive multipliers having complex zeros and complex poles but with relatively complicated relationships between the coefficients of the multiplier.

Since the phase characteristic of $M_{(1)}(z)$ has the same form as $M(z)$ in Lemma 2.5, only $M_{(2)}(z)$, $M_{(3)}(z)$ and $M_{(4)}(z)$ will be considered. Simple calculations show that the phase angle ϕ_2 of $M_{(2)}(e^{i\bar{\omega}})$ should lie in the interval

$$-\frac{\pi}{2} + \tan^{-1}\frac{\sin\bar{\omega}}{\cos\bar{\omega} - \dfrac{\cos\bar{\omega}}{2 - \cos\bar{\omega}}} \leqq \phi_2 \leqq 0,$$

where $0 \leqq \bar{\omega} \leqq \pi$.

Similarly $\phi_3 = \arg M_{(3)}(e^{i\bar{\omega}})$ and $\phi_4 = \arg M_{(4)}(e^{i\bar{\omega}})$ lie in the intervals $0 \leqq \phi_3 \leqq \frac{1}{2}\bar{\omega}$ and $-\frac{1}{2}\bar{\omega} \leqq \phi_4 \leqq 0$, respectively, for all $0 \leqq \bar{\omega} \leqq \pi$.

The phase curves of $M_{(i)}$, $i = 1, \cdots, 4$, lie in the regions indicated in Fig. 2.

LEMMA 2.8. *If $M(z)$ is a linear time-invariant discrete operator and $K(nT)$ is a time-varying function, and there exists a positive real constant β such that:*

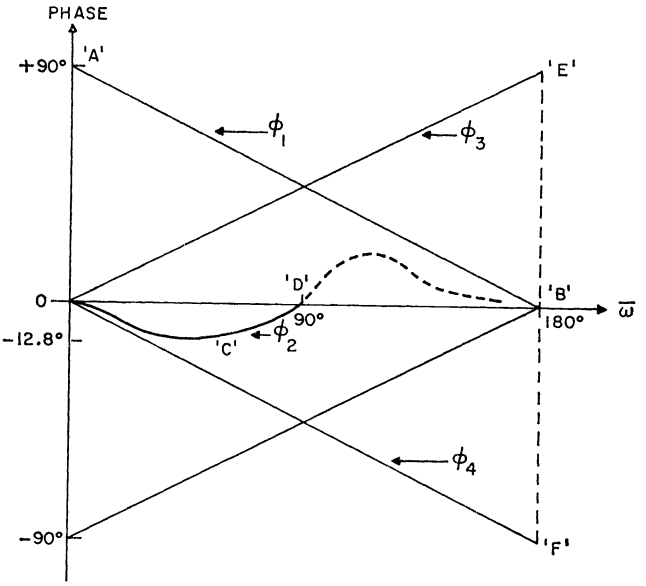
(i) $M_1(z) = M(z/\beta)$ *is passive,*

(ii)

$$0 < K(nT) < \infty, \quad \frac{K(nT + T)}{K(nT)} < \beta^2, \tag{2.16}$$

then KM is also a passive operator.

Proof. Without any loss of generality, the proof will be given for $T = 1$.



ϕ_1 LIES IN THE TRIANGLE OAB

ϕ_2 LIES IN THE REGION OABDC

ϕ_3 LIES IN THE TRIANGLE OEB

ϕ_4 LIES IN THE TRIANGLE OFB

FIG. 2. *Phase angles of multipliers*

Since $M_1(z)$ is a passive operator, its input and output sequences x_1, y_1 satisfy the inequality

$$\sum_{n=0}^{\nu} x_1(n)y_1(n) \geqq 0 \quad \text{for all} \quad \nu \in N.$$

If $x(n) = x_1(n)e^{-\alpha n}$ and $\beta = e^{\alpha}$, $\alpha > 0$, then $x \in l_{2e}$ if $x_1 \in l_{2e}$. If the input to $M(z)$ is $x(n)$, the output of $M(z)$ is $y_1(n)e^{-\alpha n}$ (see [8, p. 162]). Then

$$\begin{aligned} \langle x, Ky \rangle_\nu &= \sum_{n=0}^{\nu} x_1(n)e^{-\alpha n}y_1(n)K(n)e^{-\alpha n} \\ &= \sum_{n=0}^{\nu} \beta^{-2n}K(n)x_1(n)y_1(n). \end{aligned} \tag{2.17}$$

Since $\sum_{n=0}^{\nu} x_1(n)y_1(n) \geqq 0$, passivity of (2.17) is established if $K(n+1)/K(n) < e^{2\alpha} = \beta^2$. Note that when T tends to zero in (2.16), corresponding to the continuous case, the condition (2.16) becomes

$\dot{K}/K < 2\alpha$ (see [7c]), where $\dot{K} = dK(t)/dt$. In this case the equivalent condition of (i) of Lemma 2.8 is that the multiplier $M(s - \alpha)$ is positive real (s is the Laplace transform variable).

2.6. Stability theorems. If a feedback system has the form shown in Fig. 1 and the subsystems G and F are assumed to be operators mapping l_{2e} into itself, the operator equation may be formulated as

$$FG(e) + e(n) = x(n),$$

$$G(e) = y(n).$$

The objective is to find sufficient conditions under which $x \in l_2$ and $e, y \in l_{2e}$ imply $e, y \in l_2$. In such a case the system is defined to be l_2-stable. Moreover, if $x(n) \to 0$ as $n \to \infty$ implies $e(n), y(n) \to 0$ as $n \to \infty$, the system is said to be asymptotically stable. The l_2-stability condition is stated without proof in Theorem 2.1. The proof can be found in [7b] and is established by restricting both F and G to be passive operators. Similar results may also be found in [11], [12]. If F and G do not satisfy the conditions stated in Theorem 2.1, stability can still be established by finding a multiplier M such that FM^{-1} and MG are passive operators. The asymptotic stability condition is derived in Theorem 2.2.

THEOREM 2.1. *In the feedback system in Fig.* 1, *if the operator F is δ-M-passive or δ-passive, and G is $-(\epsilon$-passive$)$ or $-(\epsilon$-M-passive$)$ respectively, where δ and ϵ are positive real numbers and $\delta > \epsilon \geqq 0$, then the system is l_2-stable. The converse statement resulting from interchanging operators F and G is also true.*

THEOREM 2.2. *If the system is l_2-stable, and F and G belong to the classes stated in* §2.3 *and* §2.4 *respectively, then $|y(n)| < \infty$ for all positive integers n and $x(n \to \infty) \to 0$ implies $e(n \to \infty) \to 0$ and $y(n \to \infty) \to 0$.*

Proof. (i) $|y(n)| < \infty$ for all n:

$$(2.18)\quad e(n) = x(n) - F[y(n)] \quad \text{and} \quad y(n) = \sum_{m=0}^{n} g(n-m)e(m).$$

By Schwarz's inequality,

$$|y(n)|^2 \leqq \sum_{m=0}^{n} |g(m)|^2 \sum_{m=0}^{n} |e(m)|^2 < \infty.$$

Hence, $|y(n)| < \infty$ for all n.

(ii) $e(n \to \infty) \to 0$ and $y(n \to \infty) \to 0$ as $x(n \to \infty) \to 0$: Let

$$G(s) = \sum_{n=0}^{\infty} g(n)e^{-ns} \quad \text{and} \quad E(s) = \sum_{n=0}^{\infty} e(n)e^{-ns}.$$

Then since $g, e \in l_2$,

$$(2.19)\qquad y(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} G(i\omega)E(i\omega)\exp(in\omega)\,d\omega.$$

Since $G(i\omega)E(i\omega)$ is an absolutely integrable function in the range $-\pi \leqq \omega \leqq \pi$, $y(n) \to 0$ as $n \to \infty$ (see [10, p. 174, Corollary]). Hence $e(n \to \infty) \to 0$ as $x(n \to \infty) \to 0$ by (2.18).

3. Nonlinear time-invariant systems.

3.1. Main theorems. The theorems stated in this section represent the main results for nonlinear time-invariant systems.

THEOREM 3.1.[1] *In the given discrete system of Fig.* 1, *if* $F(\cdot) \in F_m[(0, K_2), (0, b)]$ *and a multiplier* $M(z)$ (*defined in* (2.8)) *exists such that*

$$\text{(3.1)} \qquad \text{Re}\left[G(z) + \frac{1}{b}\right] M(z) + \alpha_0\left(\frac{1}{K_2} - \frac{1}{b}\right) \geqq 0 \quad \textit{for all} \quad |z| = 1,$$

where $\alpha_0 = M(z = 1)$, *then the system is asymptotically stable.*

Proof. Transforming the feedback system in Fig. 1, one obtains an equivalent system containing new operators $G_1(z) \triangleq G(z) + 1/b$ and $F_1 \triangleq F/(b - F)$, where $F_1(\cdot) \in F_m[(0, bK_2/(b - K_2)), (0, \infty)]$. But $M^{-1}(z)$ followed by $F_1(\cdot)$ is δ-M-passive since

$$\text{(3.2)} \qquad \langle F_1[M^{-1}(y_1)], y_1\rangle_\nu \geqq \alpha_0\left(\frac{1}{K_2} - \frac{1}{b}\right) \langle F_1[M^{-1}(y_1)], F_1[M^{-1}(y_1)]\rangle_\nu$$

as shown below.

Hence by Theorems 2.1 and 2.2, if $G_1(z)M(z) + \alpha_0(1/K_2 - 1/b)$ is a passive operator, the system is asymptotically stable.

Proof of inequality (3.2). Let $w(n)$ be defined by the relationship $w(n) = M^{-1}[y_1(n)]$. Then the left-hand side of inequality (3.2) may be expressed as (see Lemma 2.6)

$$\begin{aligned} \langle F_1(w), M(w)\rangle_\nu &= \langle F_1(w), M_{(1)}(w) + M_{(2)}(w) + \sum_{i=1}^{m} M_i(w)\rangle_\nu \\ &\geqq \langle F_1(w), M_{(1)}(w)\rangle_\nu \qquad \text{(by Lemma 2.6)} \\ &\geqq \langle F_1(w), \alpha_0 w\rangle_\nu \quad \text{for all} \quad \nu \in N. \end{aligned} \tag{3.3}$$

Since $F_1(w) < bK_2w/(b - K_2)$ or $w > (1/K_2 - 1/b)F_1(w)$, (3.3) becomes

$$\langle F_1(w), M(w)\rangle_\nu \geqq \alpha_0\left(\frac{1}{K_2} - \frac{1}{b}\right) \langle F_1(w), F_1(w)\rangle_\nu,$$

which completes the proof.

---

[1] After the completion of this work, it was brought to the authors' attention that in a recent report [15] other conditions have been derived for monotonic and odd monotonic nonlinearities by employing a frequency-time domain method. The multiplier for the case of monotonic nonlinearities derived in that report appears more general than those derived here. However, it is noted that the multiplier used in Example 3.2 of their report to prove the stability of the system containing an odd monotonic function is a special case of the multiplier derived in this paper.

THEOREM 3.2. *Under the same conditions as in Theorem* 3.1, *if* $F(\cdot) \in F_{0m}[(0, K_2), (0, b)]$, *then the system is asymptotically stable if*

$$\text{(3.4)} \quad \text{Re}\left[G(z) + \frac{1}{b}\right] M(z) + \alpha\left(\frac{1}{K_2} - \frac{1}{b}\right) \geqq 0 \quad \textit{for all} \quad |z| = 1,$$

where $M(z)$ *is defined in* (2.9) *and* $\alpha = M_{(1)}(1) + M_{(3)}(0)$.

Proof. Transforming the system as before, one obtains $G_1(z) = G(z) + 1/b$ and an odd monotonic function

$$F_1(\cdot) \in F_{0m}\left[\left(0, \frac{bK_2}{b - K_2}\right), (0, \infty)\right].$$

Since it will be shown that

$$\text{(3.5)} \quad \langle F_1[M^{-1}(y)], y\rangle_n \geqq \alpha\left(\frac{1}{K_2} - \frac{1}{b}\right) \langle F_1[M^{-1}(y)], F_1[M^{-1}(y)]\rangle_n,$$

the system is asymptotically stable by Theorems 2.1 and 2.2 if condition (3.4) is satisfied.

Proof of inequality (3.5). If $w(n)$ is defined by the relationship $w(n) = M^{-1}[y(n)]$, the left-hand side of (3.5) may be rewritten as (see Lemma 2.7)

$$\begin{aligned}\langle F_1(w), M(w)\rangle_n &= \langle F_1(w), M_{(1)}(w) + M_{(2)}(w) + M_{(3)}(w) + M_{(4)}(w)\rangle_n \\ &\geqq \langle F_1(w), M_{(1)}(w)\rangle_n + \langle F_1(w), M_{(3)}(w)\rangle_n \\ &\geqq \alpha_0\langle F_1(w), w\rangle_n + \alpha_1\langle F_1(w), w\rangle_n\end{aligned}$$

(where $\alpha_0 = M_{(1)}(1)$ and $\alpha_1 = M_{(3)}(0)$)

$$\geqq \alpha\left(\frac{1}{K_2} - \frac{1}{b}\right) \langle F_1(w), F_1(w)\rangle_n.$$

Comments on Theorems 3.1 *and* 3.2. (i) Since the form of the multipliers $M(z)$ in (3.1) and (3.4) are known, the frequency domain conditions derived may be used to determine the stability of discrete feedback systems with monotonic and odd monotonic nonlinearities. It should be noted that Tsypkin's stability criterion (1.2) for systems with a monotonic nonlinear function is only a special case of (3.1) and may be obtained by setting $M(z) = (z - \eta)/z$, where $0 \leqq \eta \leqq 1$ and $b \to \infty$.

(ii) If the monotonic functions F in either Theorems 3.1 or 3.2 have the same sector and slope bounds ($b = K_2$), the stability conditions have the form

$$\text{Re}\left[G(z) + \frac{1}{K_2}\right] M(z) \geqq 0 \quad \text{for all} \quad |z| = 1.$$

(iii) Using a specific form of the multipliers derived, we can derive sim-

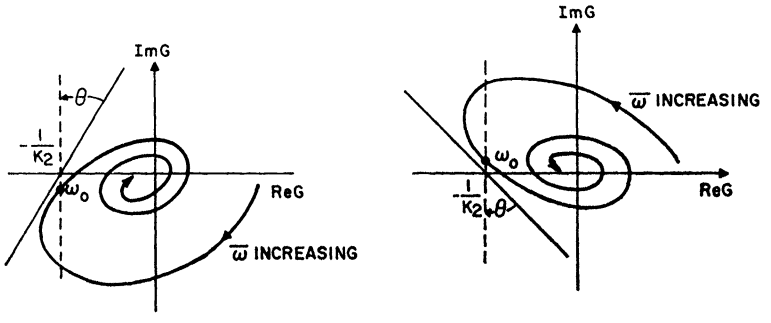


FIG. 3. *Proof of Theorem 3.3*

ple geometrical criteria for determining the stability of systems with monotonic and odd monotonic nonlinearities directly from a frequency response plot of $G(z)$ as shown in the next section.

3.2. Geometrical stability criteria. Tsypkin's circle criterion [2b] for nonlinear time-varying discrete systems is well known and has considerable practical value since the criterion only requires the frequency response of the linear time-invariant part of the system and by-passes the need to determine a multiplier $M(z)$ which usually requires a large amount of computation.

In this section, similar geometrical stability criteria are obtained for the class of discrete systems, in Fig. 1, containing a linear time-invariant plant $G(z)$, satisfying the conditions stated in §2.4, and a monotonic function whose sector and slope bounds are equal. It appears at the present time that the criteria derived in this section can be easily generalized since the forms chosen here for $M(z)$ are only special cases of those allowed by Theorems 3.1 and 3.2. Similar geometrical criteria for continuous systems are given in [7b].

Two examples are discussed at the end of the section to show the applicability of the criteria developed in this section.

THEOREM 3.3. *Let the Nyquist plot of* $G(e^{i\bar{\omega}})$ *for all* $0 \leqq \bar{\omega} \leqq \pi$ *lie entirely to the right of a straight line, whose slope* γ *is nonnegative passing through* $(-1/K_2, 0)$. *Let* ω_0 *be such that* $\operatorname{Re} G(e^{i\omega_0}) = -1/K_2$ (see Fig. 3a) *and* $\operatorname{Re} G(e^{i\bar{\omega}}) \geqq -1/K_2$ *for* $\bar{\omega} \geqq \omega_0$ *and* $\operatorname{Im} G(e^{i\bar{\omega}}) \leqq 0$ *for* $\omega_0 \geqq \bar{\omega} \geqq 0$. *Then the system is asymptotically stable for all monotonic functions* $F(\cdot) \in F_m[(0, K_2), (0, K_2)]$ *in the feedback path if*

$$\theta \leqq -\frac{1}{2}\omega_0 + \frac{\pi}{2}, \tag{3.6}$$

where θ *is the angle made by the straight line and the imaginary axis, i.e.,*

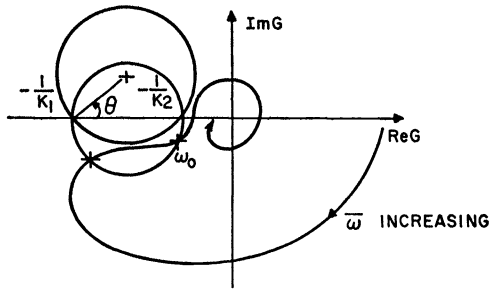


FIG. 4. *Theorem* 3.4

$\theta = \cot^{-1}\gamma$. *If* Im $G(e^{i\bar{\omega}}) \geqq 0$ *for* $\omega_0 \geqq \bar{\omega} \geqq 0$, *the same argument can be used to prove the asymptotic stability of the system with nonpositive* γ *and*

$$\theta \geqq \frac{1}{2}\omega_0 - \frac{\pi}{2}$$

(see Fig. 3b).

Proof. Case 1. $\gamma \geqq 0$. If the hypotheses on $G(z)$ are satisfied, then

$$-\left(\frac{\pi}{2} + \theta\right) \leqq \arg\left(G(e^{i\bar{\omega}}) + \frac{1}{K_2}\right) \leqq 0 \qquad \text{for} \quad 0 \leqq \bar{\omega} \leqq \omega_0,$$

$$-\frac{\pi}{2} \leqq \arg\left(G(e^{i\bar{\omega}}) + \frac{1}{K_2}\right) \leqq \frac{\pi}{2} - \theta \quad \text{for} \quad \omega_0 \leqq \bar{\omega} \leqq \pi.$$

Hence choosing $M(z) = (z-1)/(z-\eta)$, where $0 \leqq \eta \leqq 1$, we can always satisfy the following condition by the proper choice of η:

$$\text{(3.7)} \qquad \text{Re}\left(G(z) + \frac{1}{K_2}\right) M(z) \geqq 0 \quad \text{for all} \quad |z| = 1.$$

Hence the system is asymptotically stable by Lemma 2.2 and Theorem 3.1.

Case 2. $\gamma \leqq 0$. With $M(z) = (z-\eta)/(z-1)$, the inequality (3.7) can be satisfied if the hypotheses of the theorem are satisfied.

If for a specified range $(0, K_2)$, the condition $\theta \leqq -\frac{1}{2}\omega_0 + \frac{1}{2}\pi$ is not satisfied, the results of Theorem 3.3 may still be applied by either reducing the range $(0, K_2^* < K_2)$ or by increasing the sampling rate. In both cases, the value of ω_0 is reduced (see Examples 3.1 and 3.2).

By restricting the range of the nonlinear function, $F(\cdot) \in F_m[(K_1, K_2), (K_1, K_2)]$, $0 < K_1 < K_2 < \infty$, we may obtain a circle criterion from Theorem 3.3.

THEOREM 3.4. *Let the Nyquist plot of* $G(e^{i\bar{\omega}})$ *for all* $0 \leqq \bar{\omega} \leqq \pi$ *lie entirely outside a circle with center in the upper half-plane and passing through* $(-1/K_2, 0)$ *and* $(-1/K_1, 0)$. *Let the frequency* ω_0 *correspond to one of the*

points at which $G(e^{i\bar{\omega}_0})$ *intersects the circle having* $(-1/K_2, 0)$, $(-1/K_1, 0)$ *as its diameter* (see Fig. 4). *If* $G(e^{i\bar{\omega}})$ *for* $\bar{\omega} > \omega_0$ *lies outside the two circles and* $\operatorname{Im} G(e^{i\bar{\omega}}) \leqq 0$ *for* $\omega_0 \geqq \bar{\omega} \geqq 0$, *the system is asymptotically stable for all monotonic functions* $F(\cdot) \in F_m[(K_1, K_2), (K_1, K_2)]$ *in the feedback path if* $\theta \leqq -\frac{1}{2}\omega_0 + \frac{1}{2}\pi$, $0 \leqq \theta \leqq \frac{1}{2}\pi$, *where* θ *is the angle subtended by the centers of the two circles at* $(-1/K_1, 0)$.

If $\operatorname{Im} G(e^{i\bar{\omega}}) \geqq 0$ for $\omega_0 \geqq \bar{\omega} \geqq 0$, a similar theorem can be stated in terms of a circle with its center in the lower half-plane.

Proof. Case 1. $\frac{1}{2}\pi \geqq \theta \geqq 0$. Since the Nyquist plot $G(e^{i\bar{\omega}})$ in $0 \leqq \bar{\omega} \leqq \pi$ lies entirely outside a circle passing through the points $(-1/K_1, 0)$, $(-1/K_2, 0)$ (see Fig. 4) on the negative real axis, the following transformations may be carried out:

(a) Shifting the origin of coordinates to $(-1/K_2, 0)$ yields $G(e^{i\bar{\omega}}) + 1/K_2$ lying entirely outside a circle through the origin.

(b) Inverting with respect to the origin, $K_2/(1 + K_2G(e^{i\bar{\omega}}))$ lies entirely to the right of a straight line through $(-K_2K_1/(K_2 - K_1), 0)$.

Hence, proposing a multiplier $M(z) = (z - \eta)/(z - 1)$, we can show that $K_2/(1 + K_2G)$ is asymptotically stable for all monotonic functions in the sector $0 < (F(y_1) - F(y_2))/(y_1 - y_2) < (K_2 - K_1)/K_2K_1$ by using a method similar to that used in the proof of Theorem 3.3, or G is asymptotically stable for all monotonic functions satisfying $K_1 < (F(y_1) - F(y_2))/(y_1 - y_2) < K_2$.

Case 2. $-\frac{1}{2}\pi \leqq \theta \leqq 0$. A similar procedure can be used in this case but the proof is omitted here.

It should be noted that by restricting the monotonic nonlinearity in Theorems 3.3 and 3.4 to be an odd function, conditions on γ and θ can be relaxed since $M(z)$ can also have the form $(z + \eta)/(z + 1)$, where $0 \leqq \eta \leqq 1$ (see Theorem 3.2).

Example 3.1. If a second order sampled-data control system with a zero order hold circuit is given such that $G(s) = (1 - e^{-sT})/(s^2(s + 1))$, the corresponding z-transform with sampling period $T = 1$ is given by $G(z) = (e^{-1}(z + e - 2))/((z - e^{-1})(z - 1))$. This example has been considered by Jury and Lee [4], Szegö and Pearson [5] and Pearson and Gibson [6]. Drawing a Nyquist plot of $G(e^{i\bar{\omega}})$ for $0 \leqq \bar{\omega} \leqq \pi$, we obtain the Nyquist gain $K_N = 2.39$. Hence the system with constant gain K in the feedback path is asymptotically stable if $0 < K < K_N$ [8], [9]. If a straight line is drawn tangential to $G(e^{i\bar{\omega}})$ at $(1/K_N, 0)$, $\theta = 1.063$ rad and $\omega_0 = 1.324$ rad are obtained which do not satisfy the inequality (3.6). However, choosing $K_2 = 2.34$, one can find $\theta = 0.916$ rad and $\omega_0 = 1.305$ rad on the Nyquist plot of G, which satisfy the inequality (3.6) and belong to Case 1 of Theorem 3.3. Hence the system is asymptotically stable with the

TABLE 3.1
Gain K_2 obtained by various authors

	Example 3.1	Example 3.2
Tsypkin's criterion [2a] (any nonlinear time-varying function)	0.666	0.924
Stability theorem by Jury–Lee [4a], Szegö–Pearson [5] (nonlinear function with slope bounds)	2.0	2.0
Lyapunov function Pearson–Gibson [6] (monotonic nonlinear function)	1.9	
Experimental value Pearson–Gibson [6] (monotonic nonlinear function)	$\simeq$2.3	
Geometrical criterion Narendra–Cho	2.34[2]	3.39[2]
Nyquist gain [2a], [8], [9]	2.39	4.32

functions in the feedback path satisfying

$$0 < \frac{F(y_1) - F(y_2)}{y_1 - y_2} < 2.34.$$

Table 3.1 is given to compare the result obtained here with those obtained by previous authors.

Example 3.2. Consider next a second order system without a hold circuit, whose linear part is specified by $G(s) = 1/(s(s+1))$. Taking the z-transform, we get $G(z) = (1 - e^{-1})z/((z - e^{-1})(z - 1))$. From the Nyquist plot $G(e^{i\bar{\omega}})$, where $0 < \bar{\omega} \leqq \pi$, the Nyquist gain $K_N = 4.32$ is obtained (see [8, p. 243]). Choosing $K_2 = K_N$, we find that $\theta = 0.716$ rad and $\omega_0 = \pi$ rad which do not satisfy inequality (3.6).

However, $K_2 = 3.39$ yields $\theta = 0.55$ rad and $\omega_0 = 2.04$ rad. Hence, by Theorem 3.3, the system is asymptotically stable for all monotonic functions satisfying $0 < (F(y_1) - F(y_2))/(y_1 - y_2) < 3.39$.

Values of K_2 obtained by other authors are also given in Table 3.1.

4. Linear time-varying discrete systems. The results presented in this section are extensions to the discrete case of similar results obtained recently for continouus systems [7c]. The stability of a feedback system with a linear time-invariant operator $G(z)$ and a single time-varying gain $K(nT)$ belonging to the class defined in §2.3 is discussed in terms of the location of the singularities of a multiplier $M(z)$ which makes $G(z)M(z)$ passive.

[2] In a private correspondence it was brought to the attention of the authors that this gain for Example 3.1 was also obtained by Jury and Lee in a paper presented at the 1966 IFAC [4c]. However an algebraic criterion was used by those authors. Application of the same criterion to Example 3.2 also yields similar results. The geometrical criterion is, however, much simpler to apply and yields considerable insight into the nature of the stability of the discrete systems, in that it is readily apparent how to change the range of $F(\cdot)$ or the sampling rate to achieve stability.

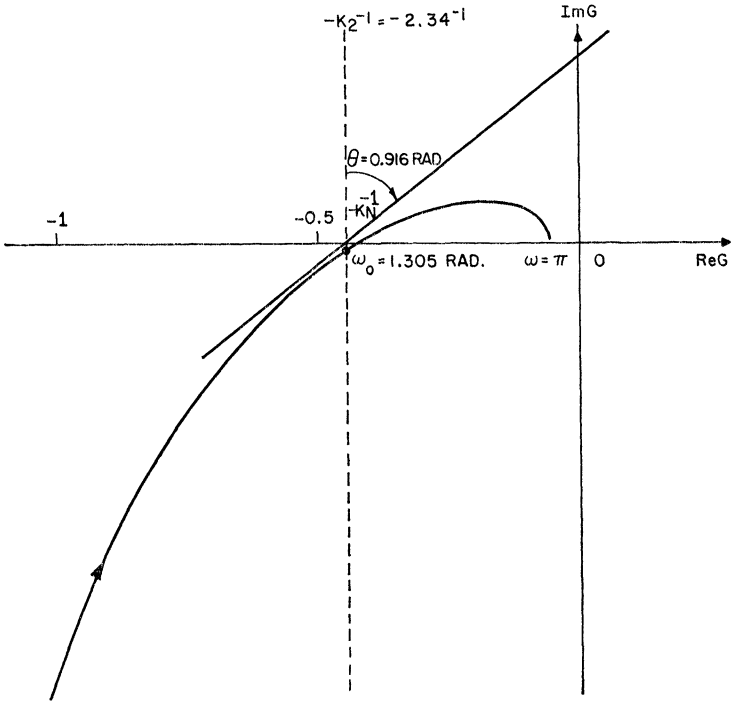


FIG. 5. *Example* 3.1, *frequency response plot of* $G(z) = e^{-1}(z + e - 2)/((z - e^{-1})(z - 1))$

Sufficient conditions for the stability of the time-varying system are also derived in terms of the root locus plot of a corresponding time-invariant feedback system.

4.1. General theorems.

THEOREM 4.1. *In the feedback system of Fig.* 1, *let* $G(z)$ *be a linear time-invariant discrete operator and* $K(nT)$ *a time-varying function in the interval* $0 < K(nT) < \infty$. *If a discrete operator* $M(z)$ *and positive real constants* β *and* δ *exist such that*:

(i) $M(z/\beta)$ *is passive*,

(ii)

$$\text{Re } G(z)M^{-1}(z) \geqq \delta > 0 \quad \textit{for all} \quad |z| = 1, \tag{4.1}$$

(iii)

$$\frac{K(nT + T)}{K(nT)} < \beta^2, \tag{4.2}$$

then the feedback system is asymptotically stable.

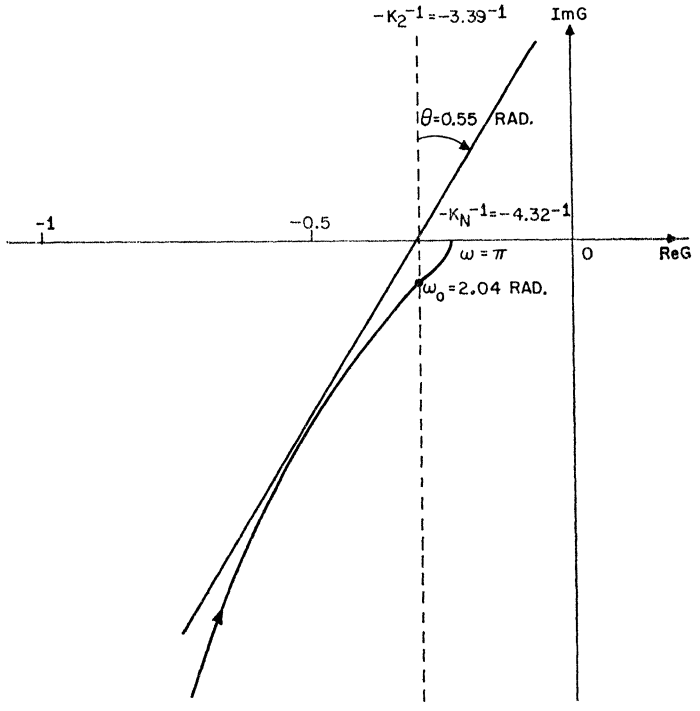


FIG. 6. *Example* 3.2, *frequency response plot of* $G(z) = (1 - e^{-1})z/((z - e^{-1})(z - 1))$

Proof. Since $M(z)$ followed by $K(nT)$ is a passive operator by Lemma 2.8, the condition (4.1) is sufficient for asymptotic stability of the system by Lemma 2.2 and Theorems 2.1 and 2.2.

THEOREM 4.2. *Under the same assumptions as in Theorem* 4.1 *and with the time-varying feedback gain* $K(nT)$ *in the range* $K_1 < K(nT) < K_2$ *for all nonnegative integers* n, *if a multiplier* $M(z)$ *exists such that* $M(z/\beta)$ *is passive and*

$$\text{(4.3)} \qquad \operatorname{Re} \frac{1 + K_2 G(z)}{1 + K_1 G(z)} \cdot M(z)^{-1} \geqq \delta > 0 \quad \textit{for all} \quad |z| = 1,$$

then the system is asymptotically stable for all $K(nT)$ *satisfying the inequality*

$$\text{(4.4)} \qquad \frac{K(nT + T) - K_1}{K(nT) - K_1} \cdot \frac{K(nT) - K_2}{K(nT + T) - K_2} < \beta^2.$$

Proof. The finite sector problem of Theorem 4.2 can be transformed to the infinite sector problem by defining

$$G_1(z) = \frac{G(z)}{1 + K_1 G(z)} + \frac{1}{K_2 - K_1} = \frac{1}{K_2 - K_1} \cdot \frac{1 + K_2 G(z)}{1 + K_1 G(z)}$$

and

$$K^*(nT) = \frac{(K_2 - K_1)(K(nT) - K_1)}{K_2 - K(nT)}.$$

By Theorem 4.1, for all $0 < K^*(nT) < \infty$ and $[K^*(nT + T)/K^*(nT)] < \beta^2$, the feedback system is asymptotically stable if hypothesis (4.3) is satisfied. However, the conditions on $K^*(nT)$ can be expressed in terms of $K(nT)$ to yield inequality (4.4) if $K_1 < K(nT) < K_2$.

The stability analysis of time-varying systems by the passive operator method consequently reduces to the determination of $M(z)$ for the given $G(z)$. Once $M(z)$ is known the value of β may be determined. If a multiplier is given as in (2.8), then $\beta = \min [1/\gamma_i, 1/\eta_i]$.

Example 4.1. In a time-varying sampled-data feedback control system (see Fig. 1) without a hold circuit, $G(s) = 1/(s(1 + T_m s))$ (where T_m is a positive real constant), and a sampler is located in the forward path. The time-varying feedback gain lies in the interval $0 < K(nT) < K_2$.

Taking z-transforms, we get

$$G(z) = \frac{(1 - e^{-T/T_m})z}{(z - 1)(z - e^{-T/T_m})}$$

and

$$1 + K_2 G(z) = \frac{z^2 + (1 - e^{-T/T_m})K_2 - (1 + e^{-T/T_m})z + e^{-T/T_m}}{(z - 1)(z - e^{-T/T_m})}. \tag{4.5}$$

If K_2 is chosen as the Nyquist gain K_N, (4.5) becomes

$$1 + K_N G(z) = \frac{(z + 1)(z + e^{-T/T_m})}{(z - 1)(z - e^{-T/T_m})}, \tag{4.6}$$

where $K_N = 2(1 + e^{-T/T_m})/(1 - e^{-T/T_m})$ (see [8, p. 243]).

Hence, if $M(z) = (z - e^{-T/T_m})/(z + e^{-T/T_m})$, $[1 + K_N G(z)]M(z)$ is a passive operator. From $M(z)$, it is easily determined that $\beta = e^{T/T_m}$ and hence the system is asymptotically stable by Theorem 4.2 if

$$\frac{K(nT + T)}{K(nT)} \cdot \frac{K(nT) - K_N}{K(nT + T) - K_N} < e^{2T/T_m}. \tag{4.7}$$

Note that if $K_2 = K_c = 2(1 - e^{-T/T_m})^2/(1 - e^{-2T/T_m})$, then $1 + K_c G(z)$ is passive, so in this case $M(z) = I$, and hence β can have infinite value. Therefore, by Theorem 4.2, the system is asymptotically stable for any time-varying gain $K(nT)$ in the range $0 < K(nT) < K_c$, and this is a special case of Tsypkin's circle criterion for time-varying sampled-data systems [2a], [2b].

The example demonstrates that $K(nT)$ can be made to lie in an increased sector for stability by sacrificing the rate of variation of the time-varying function.

4.2. Root locus stability criteria. The root locus method is well known in the analysis and design of linear time-invariant continuous and discrete systems [8], [9]. In this section stability criteria for linear time-varying sampled-data systems are derived from a root locus plot of the corresponding time-invariant system to predict the rate of variation of $K(nT)$ which assures the asymptotic stability of the system.

THEOREM 4.3. *If the root locus of the system shown in Fig. 1 for all values of constant gain K in the interval $0 < K < \infty$ lies inside a circle with center at the origin and radius $1/\beta(\leqq 1)$ in the z-plane, then the system with a time-varying feedback gain $K(nT)$ in the interval $0 < K(nT) < \infty$ is asymptotically stable if $K(nT + T)/K(nT) < \beta^2$.*

To prove this theorem, the following two lemmas are needed.

LEMMA 4.1. *There exists a discrete passive operator $M_r(z)$ whose phase alternates between $+\frac{1}{2}\pi$ and $-\frac{1}{2}\pi$ as $\bar{\omega}$ increases on the unit circle in the z-plane, or*

$$\arg M_r(e^{i\bar{\omega}}) = \begin{cases} \frac{1}{2}\pi & \text{for} \quad \omega_{2i} \leqq \bar{\omega} < \omega_{2i+1}, \\ -\frac{1}{2}\pi & \text{for} \quad \omega_{2i+1} \leqq \bar{\omega} < \omega_{2i+2}, \end{cases} \qquad i = 0, 1, 2, \cdots,$$

where $\omega_0 = 0$ and $0 \leqq \bar{\omega} \leqq \pi$.

Proof. If we choose a multiplier

$$M_r(z) = \frac{z - 1}{z + 1} \prod_{i=1}^{n} \frac{z^2 - 2z \cos \omega_{2i} + 1}{z^2 - 2z \cos \omega_{2i-1} + 1}, \tag{4.8}$$

then $\arg M_r(e^{i\bar{\omega}})$ will satisfy the conditions stated in the lemma.

LEMMA 4.2. *If a linear time-invariant discrete system with $G(z)$ in the forward path and a constant K in the feedback path is stable for all K in the interval $0 < K < \infty$, then it is possible to split the open loop of the system into two passive operators one of which is δ-passive or δ-M-passive.* (A similar result for the continuous case may be found in Theorem 2 of [13].)

Proof. By the Nyquist criterion, $(-\pi + \epsilon) \leqq \arg G(e^{i\bar{\omega}}) \leqq (\pi - \epsilon)$ for all $0 \leqq \bar{\omega} \leqq \pi$, where ϵ is an arbitrarily small positive real number. Hence there always exists a passive operator $M_r(z)$ or $M_r^{-1}(z)$ in Lemma 4.1 so that

$$-\tfrac{1}{2}\pi + \epsilon \leqq \arg G(e^{i\bar{\omega}}) M_r^{\pm 1}(e^{i\bar{\omega}}) \leqq \tfrac{1}{2}\pi - \epsilon$$

or

$$-\tfrac{1}{2}\pi \leqq \arg G(e^{i\bar{\omega}}) \cdot M_r^{\pm 1}(\alpha e^{i\bar{\omega}}) \leqq \tfrac{1}{2}\pi,$$

where $\alpha < 1$ is a positive real constant arbitrarily close to unity. Hence the open loop consists of a passive operator $G(z) \cdot M_r^{\pm 1}(\alpha z)$ (by Lemma 2.2) followed by a δ- (or δ-M-) passive operator $KM_r^{\mp 1}(\alpha z)$.

Proof of Theorem 4.3. Since $G(z/\beta)$ is an asymptotically stable operator with a constant feedback gain K in the interval $(0, \infty)$, there exists a

multiplier $M_r(z)$ so that $G(z/\beta) \cdot M_r(\alpha z)$ is a passive operator (where $\alpha < 1$ is a positive real constant; see Lemma 4.2). Hence $G(z) \cdot M_r(\alpha\beta z)$ is also a passive operator if $\beta \geqq 1$. Defining $M_r(\alpha\beta z) = M(z)$ in inequalities (4.1) and (4.2), we see that the stability condition reduces to $K(nT + T)/K(nT) < \beta^2$ by Theorem 4.1.

THEOREM 4.4. *If the root locus of the system* (see Fig. 1) *for all values of constant gain K in the interval $K_1 < K < K_2$ lies inside a circle with center at the origin and radius $1/\beta(\leqq 1)$ in the z-plane, then the system with a time-varying feedback gain $K(nT)$ satisfying $K_1 < K(nT) < K_2$ is asymptotically stable if*

$$\frac{K(nT + T) - K_1}{K(nT) - K_1} \cdot \frac{K(nT) - K_2}{K(nT + T) - K_2} < \beta^2.$$

Proof. The finite sector problem can be transformed to the equivalent infinite sector problem by defining

$$G_1(z) = \frac{G(z)}{1 + K_1 G(z)} + \frac{1}{K_2 - K_1}$$

and

$$K^*(nT) = \frac{(K_2 - K_1) \cdot (K(nT) - K_1)}{K_2 - K(nT)}.$$

In this case, $G_1(z/\beta)$ is an asymptotically stable operator for all constant feedback gains in the interval $0 < K^* < \infty$. Hence employing the same arguments used in the proofs of Theorems 4.2 and 4.3, we may prove the theorem.

Comments on Theorems 4.3 *and* 4.4. The stability criteria derived using the root locus plot yield sufficient but by no means necessary conditions. Since β is determined by the marginally passive operator $M_r(z)$, the value of β is usually a very conservative estimate in most cases.

Example 4.2. If a linear time-invariant discrete system (Fig. 1) is given such that $G(z) = (z + 0.6)^2/[(z - 1)(z - 0.36)]$, then the root locus of the system for the constant gain $2/300 \leqq K \leqq \infty$ lies inside the circle whose radius and center are 0.6 and (0, 0) (see [9, p. 187]). Hence by the root locus stability criterion developed in Theorem 4.4, the system is asymptotically stable for the linear time-varying gain function satisfying

$$\frac{2}{300} < K(nT) < \infty, \qquad \frac{K(nT + T) - 2/300}{K(nT) - 2/300} < \left(\frac{1}{0.6}\right)^2.$$

But by Theorem 4.1, defining $M(z) = (z - 0.36)/(z + 0.6)$, the time-varying gain $K(nT)$ satisfying $0 < K(nT) < \infty$ and $K(nT + T)/K(nT) < (1/0.6)^2$ also guarantees the stability of the system.

In general situations the determination of the multiplier $M(z)$ is usually a difficult problem. If, however, the root locus plot of the corresponding linear time-invariant feedback system is available and a conservative esti-

mate of $K(nT + T)/K(nT)$ is adequate, the problem can be by-passed completely using Theorem 4.3.

REFERENCES

[1] V. M. Popov, *Absolute stability of nonlinear systems of automatic control*, Automat. Remote Control, 22 (1962), pp. 857–875.

[2a] Ya. Z. Tsypkin, *The absolute stability of large-scale, nonlinear, sampled-data systems*, Dokl. Akad. Nauk SSSR, 145 (1962), pp. 52–55.

[2b] ———, *Frequency criteria for the absolute stability of nonlinear sampled-data systems*, Automat. Remote Control, 24 (1964), pp. 261–267.

[2c] ———, *Sampling Systems Theory*, vols. 1 and 2, Macmillan, New York, 1964.

[3] G. P. Szegö, *On the absolute stability of sampled-data control systems*, Proc. Nat. Acad. Sci. U.S.A., 50 (1963), pp. 558–560.

[4a] E. I. Jury and B. W. Lee, *On the stability of certain class of nonlinear sampled-data systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 51–61.

[4b] ———, *A note on the absolute stability of nonlinear sampled-data systems*, Ibid., AC-9 (1964), pp. 551–554.

[4c] ———, *A stability theory for multinonlinear control systems*, Proc. Third Congress of the International Federation of Automatic Control, London, 1966, pp. 28A.1–11.

[5] G. P. Szegö and J. B. Pearson, Jr., *On the absolute stability of sampled-data systems: the 'indirect control' case*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 160–163.

[6] J. B. Pearson and J. E. Gibson, *On the asymptotic stability of a class of saturating sampled-data systems*, IEEE Trans. Applications and Industry, 83 (1964), pp. 81–86.

[7a] K. S. Narendra and Y. S. Cho, *Stability of feedback systems containing a single odd monotonic nonlinearity*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 448–450.

[7b] ———, *An off-axis circle criterion for the stability of feedback systems with a monotonic nonlinearity*, Ibid., AC-13 (1968), to appear.

[7c] ———, *Stability of linear time-varying feedback systems*, Proc. Second Princeton Conference on Information Sciences and Systems, Princeton University, 1968, pp. 1–6.

[8] J. T. Tou, *Digital and Sampled-Data Control Systems*, McGraw-Hill, New York, 1959.

[9] B. C. Kuo, *Analysis and Synthesis of Sampled-Data Control Systems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[10] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge University Press, London, 1940.

[11] I. W. Sandberg, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell System Tech J., 44 (1965), pp. 891–894.

[12] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems I*, IEEE Trans. Automatic Control, AC-11 (1964), pp. 228–238.

[13] R. W. Brockett and J. L. Willems, *Frequency domain stability criteria I*, Ibid., AC-10 (1965), pp. 255–261.

[14] K. S. Narendra and C. P. Neuman, *Stability of a class of differential equations with a single monotone nonlinearity*, this Journal, 4 (1966), pp. 295–308.

[15] R. P. O'Shea, P. A. Kossey, M. Siman and M. I. Unis, *Some new results concerning the asymptotic stability and the response of nonlinear systems*, Rep. N67-25496, NASA, Huntsville, Alabama, 1967.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

ERRATUM: LAGRANGIAN SADDLE POINTS AND OPTIMAL CONTROL*

D. O. NORRIS

It has been pointed out by K. Tsujioka that the proof of Theorem 1 in this paper is not correct. He has given a simple counterexample (viz., $K = \{x \in R^1 : x \geqq 0\}$, $J(x) = x$ and $G(x) = 1 - \frac{1}{2}x$) which shows that the statement following (7) is incorrect.

The application made in the paper is still valid provided one works the problem in $L_\infty'[0, T]$, since in this case the positive cone has a nonvoid interior and Hurwicz's results apply (see [1, Theorem V.3.1]). This is no restriction on the problem because we are looking for a bounded controller. The application also follows from a result by Tsujioka [2, Lemma 7].

REFERENCES

[1] L. HURWICZ, *Programming in linear spaces*, Studies in Linear and Non-Linear Programming, K. J. Arrow, L. Hurwicz, H. Uzawa, eds., Stanford University Press, Stanford, 1958, pp. 38–102.

[2] K. TSUJIOKA AND N. MIYAMOTO, *A theorem on multiplication operator in certain function spaces*, Sci. Papers College Gen. Ed. Univ. Tokyo, 17 (1967), pp. 141–154.

* This Journal, 5 (1967), pp. 594–599. Received by the editors March 26, 1968.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

MULTIPLIER FUNCTIONS IN OPTIMAL CONTROL*

J. PONSTEIN†

Summary. By means of functional analytical tools, necessary optimality conditions are obtained for optimal control problems involving time dependent inequality constraints. Such problems require Lagrangian multiplier *functions* rather than numbers. The results are based on a lemma similar to Farkas' lemma in mathematical programming.

1. Introduction. The aim of this paper is to establish necessary optimality conditions for optimal control problems whose constraints contain inequalities which must hold for each individual time in a certain time interval, and which therefore are not necessarily integrated over time. A typical example is the following. Let R be the field of real numbers, and let E^k be the k-dimensional Euclidean space.

$$(1) \quad \text{Minimize} \qquad \int_{t_0}^{t_1} f(x(t), t)\, dt \Big| g(x(t), t) \leqq 0,$$

where $f(x(t), t) \in R$, $x(t) \in E^n$, $g(x(t), t) \in E^m$, and $t \in [t_0, t_1] \subset R$. The vector $x(t)$ is the control vector, so that it may seem that no state variables occur. These have been eliminated, however, by solving the ordinary or partial differential equations for them. As is illustrated in one of the examples below, this elimination need not be carried out explicitly.

For this kind of problem we shall use Lagrangian multiplier *functions* rather than the usual Lagrangian multipliers which are *numbers*. The latter are used, for instance, in papers by Halkin and Neustadt [2], [5], and they also occur in a paper by Ritter [6], who considers constraints such as $\int_a^b g(x(t), t)\, dt \leqq 0$. Very general kinds of multipliers, it is true, are considered in a fundamental paper by Hurwicz [3], but the kind of problem we have in mind is not worked out by him. Also it seems that the generalization of Farkas' lemma that we shall need is somewhat different from his, because we do not consider conjugate spaces as far as the multipliers are concerned. Our subject is also considered in a paper by Russell [7], but approached from another point of view. The most important difference seems to be that the assumptions which have to be verified are quite different: in his case the set of "feasible solutions" must be "linearly approximable"; in our case a certain set, namely S, must be closed and a constraint condition must hold. We agree with this author that in many

* Received by the editors February 1, 1968, and in revised form April 18, 1968.

† Computer Science Department, University of Kentucky, Lexington, Kentucky 40506.

problems it will be difficult indeed to verify the assumptions. This is true, in particular, in time optimal control problems, or when partial differential equations are involved. Another difference is that in the present paper the (nonnegative) multipliers are introduced right at the beginning, whereas in Russell's paper they appear at the end.

2. Reformulation as a functional optimization problem. Since we want to apply functional analytical tools let us rewrite (1) as a functional optimization problem. This will also enable us to generalize.

Let X and Y be some function spaces, to be specified in more detail later on. Let an element x of X map t onto $x(t) \in E^n$, and let an element y of Y map t onto $y(t) \in E^m$, $t \in [t_0, t_1]$. Define the following functionals on X and $X \times Y$:

$$Fx = \int_{t_0}^{t_1} f(x(t), t)\, dt, \tag{2}$$

$$G(x, y) = \int_{t_0}^{t_1} y^T(t) g(x(t), t)\, dt. \tag{3}$$

Here and elsewhere T denotes transposition. Further, let $y \geqq 0$ mean that the components of $y(t)$ are nonnegative for all t, $t \in [t_0, t_1]$. Then problem (1) can be written in the following form:

$$\text{(4)} \quad \text{Minimize} \quad Fx \,|\, G(x, y) \leqq 0 \quad \text{for all} \quad y \geqq 0, \quad x \in X, \quad y \in Y.$$

Clearly the multipliers y are functions. Another approach [5] would be to introduce the functional $\max g(x(t), t)$, where the maximum is taken over $[t_0, t_1]$. Then numbers would again appear as multipliers, which seem easier to handle than functions. The disadvantage of using functions, however, is more than balanced by the advantages one obtains by considering the functional (3) instead of $\max g(x(t), t)$.

Problem (4) also suggests the following generalization. Let X be any reflexive Banach space and let Y be any normed linear space, both over R. Let $F: x \to Fx$ be a functional on X and let $G: (x, y) \to G(x, y)$ be a functional on $X \times Y$, which is linear in y. Both F and G must have (bounded) Fréchet differentials with respect to x. Finally, we assume that Y is partially ordered by means of the relation $\geqq$, which must satisfy the following relations:

$$\text{(5a)} \qquad 0 \geqq 0;$$

$$\text{(5b)} \qquad \text{if} \quad y_1 \geqq 0 \quad \text{and} \quad y_2 \geqq 0, \quad \text{then} \quad y_1 + y_2 \geqq 0, \quad y_1, y_2 \in Y;$$

$$\text{(5c)} \qquad \text{if} \quad y \geqq 0 \quad \text{and} \quad \alpha \geqq 0, \quad \text{then} \quad \alpha y \geqq 0, \quad \alpha \in R, \quad y \in Y.$$

Now problem (4) covers a much wider class of problems, e.g., problems where t_1 is allowed to vary but still is subject to inequality constraints, or where time-independent constraints are involved, or the general problem of nonlinear mathematical programming.

3. A generalization of Farkas' lemma. Let X, Y, etc. be as defined in the previous section. Let $a: x \to ax$ be a bounded, real-linear functional on X, and let $B: (x, y) \to B(x, y)$ be a real-bilinear functional on $X \times Y$, such that the functionals on X which are induced by taking y fixed are bounded for each y. Let x^0 be a solution of (4) and consider the following sets:

$$S = \{s_y \mid s_y h = B(h, y) \text{ for some } y \in Y \text{ for which } y \geqq 0 \text{ and } G(x^0, y) = 0, \text{ and all } h,\ h \in X\}, \tag{6a}$$

$$H = \{h \mid B(h, y) \geqq 0 \text{ for all } y \in Y \text{ for which } y \geqq 0 \text{ and } G(x^0, y) = 0,\ h \in X\}. \tag{6b}$$

LEMMA. *If for all $h \in H$ we have that $ah \geqq 0$, and if S is closed in the conjugate space X^* of X, then $a \in S$. In other words, there exists a $y^0 \in Y$ for which $y^0 \geqq 0$ and $G(x^0, y^0) = 0$, and such that $ah = B(h, y^0)$ for all $h \in X$.*

Proof. It is easily shown that S is convex and that the elements of S are bounded, linear functionals on X, so that S is a subset of X^*. Suppose $a \notin S$. Since S is closed and convex, it follows from one of the separation theorems [1, V2.7, Theorem 10] that there exists a $p^{**} \in X^{**}$ such that

$$p^{**}a < \inf_{s \in S} p^{**}s. \tag{7}$$

Since X is reflexive there exists a $p \in X$ such that $p^{**}x^* = x^*p$ for all $x^* \in X^*$. Therefore,

$$ap < \inf \{s_y p \mid s_y \in S\}. \tag{8}$$

Taking $y = 0$ we find that $\inf \leqq 0$. If $\inf < 0$ there would exist a $\bar{y} \geqq 0$, such that $G(x^0, \bar{y}) = 0$ and such that $s_{\bar{y}}p < 0$, and hence, for $\alpha \geqq 0$, $s_{\alpha\bar{y}}p = \alpha s_{\bar{y}}p$ would tend to $-\infty$ if α would tend to $+\infty$. This contradicts $ap < \inf$, so that we must have $\inf = 0$. From this, however, it would follow that $B(p, y) \geqq 0$ for all $y \in Y$ for which $y \geqq 0$ and $G(x^0, y) = 0$, so that $p \in H$, and, according to one of our assumptions, $ap \geqq 0$. But this contradicts $ap < \inf = 0$. Hence our assumption concerning a is false and $a \in S$.

4. Necessary optimality conditions. In order to be able to derive necessary optimality conditions we have to assume that some sort of constraint condition holds. We shall take a form which is a straightforward generaliza-

tion of one of the forms of constraint qualification in mathematical programming, although, just as there, weaker forms are probably sufficient.

Let F_x and $G_x(y)$ be the Fréchet differentials of F and G, respectively, with respect to x. Let x^0 be a solution of the minimization problem (4). Then the constraint condition reads as follows:

$$F_{x^0}h \geqq 0 \quad \text{for all} \quad h \in X \quad \text{which satisfy} \quad G_{x^0}(y)h \leqq 0 \quad \text{for all} \quad y \in Y \quad \text{for which} \quad y \geqq 0 \quad \text{and} \quad G(x^0, y) = 0. \tag{9}$$

In adding the requirement that y must satisfy $G(x^0, y) = 0$, we do in essence the same as is done in mathematical programming when distinguishing between active and nonactive constraints.

Since Fréchet differentials are bounded, linear functionals, we can apply the lemma by putting

$$B(h, y) = -G_{x^0}(y)h \tag{10}$$

and

$$ah = F_{x^0}h, \tag{11}$$

so that

$$S = \{-G_{x^0}(y) \mid y \geqq 0, G(x^0, y) = 0, y \in Y\}, \tag{12}$$

and we find the following result.

THEOREM. *If X, Y, etc. are as defined above, if x^0 is a solution of* (4), *if the constraint condition* (9) *holds, and if S defined by* (12) *is closed in X^*, then there exists a $y^0 \in Y$ such that*

$$F_{x^0} + G_{x^0}(y^0) = 0, \tag{13a}$$

$$y^0 \geqq 0, \tag{13b}$$

$$G(x^0, y^0) = 0. \tag{13c}$$

Two crucial requirements for this theorem are, of course, that X must be reflexive and that S must be a closed subset of X^*.

In the applications of the next section, and in general in many cases, X will be a Hilbert space, so that X is reflexive. Furthermore, $B(x, y)$ will take the form of an inner product in X:

$$B(x, y) = (x, \bar{B}y), \tag{14}$$

where $\bar{B}: y \to \bar{B}y \in X$.

In order to show that S is closed, let $s^i \in S$, $i = 1, 2, \cdots$, converge to s. Then we have to show that $s \in S$. Since $s^i \in S$ there exist $y^i \in Y$, such that $y^i \geqq 0$, $G(x^0, y^i) = 0$ and

$$s^i x = (x, \bar{B}y^i) \quad \text{for all} \quad x \in X. \tag{15}$$

Further we have, since s is a bounded, linear functional on a Hilbert space, that there exists a $p \in X$ such that

$$sx = (x, p) \quad \text{for all} \quad x \in X. \tag{16}$$

Hence $\| s^i - s \| = \| \bar{B}y^i - p \|$ converges to zero, so that $\bar{B}y^i$ converges to p. For each of the applications we shall show that from this it follows that there exists a $y \in Y$ such that $y \geqq 0$, $G(x^0, y) = 0$ and $p = \bar{B}y$. From this and (14) and (16) it then follows that $sx = B(x, y)$; hence $s \in S$.

5. Two applications.

5.1. Let us first apply the results of the previous section to a simple example. Suppose x^0 solves the following problem:

$$\text{Minimize} \quad \int_0^3 x(t)\, dt \,\Big|\, g_1(x(t)) \leqq 0,\ g_2(x(t)) \leqq 0, \tag{17}$$

where $x(t) \in R$, $g_1(x(t)) = c_1(t) - x(t)$, $g_2(x(t)) = c_2(t) - x(t)$, and $c_1(t) = 1 - t$ if $t \in [0, 1]$, $c_1(t) = 0$ if $t \in [1, 3]$, $c_2(t) = 0$ if $t \in [0, 2]$, $c_2(t) = t - 2$ if $t \in [2, 3]$.

Then

$$Fx = \int_0^3 x(t)\, dt \tag{18}$$

and

$$G(x, y) = \int_0^3 \{y_1(t)g_1(x(t)) + y_2(t)g_2(x(t))\}\, dt, \tag{19}$$

so that

$$F_{x^0} h = \int_0^3 h(t)\, dt = Fh \tag{20}$$

and

$$G_{x^0}(y)h = -\int_0^3 (y_1(t) + y_2(t))h(t)\, dt. \tag{21}$$

We assume that x, y_1 and y_2 belong to $L_2[0, 3]$ and that $y \geqq 0$ means that $y_1(t) \geqq 0$ and $y_2(t) \geqq 0$ almost everywhere in $[0, 3]$. Clearly, the requirement that $G(x^0, y) = 0$ is equivalent to

$$y_1(t)g_1(x^0(t)) = y_2(t)g_2(x^0(t)) = 0, \quad t \in [0, 3], \tag{22}$$

if $y \geqq 0$, so that S becomes

$$S = \Big\{ s_y \,\Big|\, s_y h = \int_0^3 (y_1(t) + y_2(t))h(t)\, dt,\ y \geqq 0, \tag{23}$$
$$y_1(t)g_1(x^0(t)) = y_2(t)g_2(x^0(t)) = 0,\ t \in [0, 3] \Big\}.$$

Here and elsewhere we have dropped the clause "almost everywhere".

In order to show that S is closed we follow the procedure as outlined at the end of the previous section. Hence we are given a sequence y^i, such that $y^i \geqq 0$ and $G(x^0, y^i) = 0$, and a $p: t \to p(t) \in R$ such that

$$\int_0^3 \{y_1^i(t) + y_2^i(t) - p(t)\}^2 \, dt \tag{24}$$

converges to zero. We have to show that we can find $y_1(t)$, $y_2(t) \geqq 0$ such that (22) holds and such that $p = \bar{B}y = y_1 + y_2$, or that $p(t) = y_1(t) + y_2(t)$.

From (24) it follows that $p(t) \geqq 0$ for all t, and from (24) and $G(x^0, y^i) = 0$, that $p(t) = 0$ if $g_1(x^0(t)) < 0$ and $g_2(x^0(t)) < 0$ in a subset of $[0, 3]$. In the latter case we must have that $y_1(t) = y_2(t) = 0$. If $g_1(x^0(t)) = g_2(x^0(t)) = 0$, then there are no restrictions on $y_1(t)$ and $y_2(t)$, as far as (22) is concerned, so that we can always determine them in such a way that $p(t) = y_1(t) + y_2(t)$. If $g_1(x^0(t)) < 0$ and $g_2(x^0(t)) = 0$, then $y_1(t) = 0$, but there are no restrictions on $y_2(t)$, so that we can put $y_2(t) = p(t)$. Similarly, if $g_1(x^0(t)) = 0$ and $g_2(x^0(t)) < 0$ we have that $y_2(t) = 0$ and we can put $y_1(t) = p(t)$. In all cases we clearly have that $p(t) = y_1(t) + y_2(t)$. Hence S is closed.

Let us now show that the constraint condition (9) is satisfied. Let h be such that $G_{x^0}(y)h \leqq 0$ for all y for which $y \geqq 0$ and $G(x^0, y) = 0$. Since x^0 is an optimum of the minimization problem (17), it is easily checked that for all t at least one of $g_1(x^0(t))$ and $g_2(x^0(t))$ is zero, because otherwise the solution could be improved. Hence the sum $y_1(t) + y_2(t)$ may be taken positive, so that $h(t) \geqq 0$ for all t. Therefore, $x^0 + h$ is feasible, so that $F(x^0 + h) \geqq Fx^0$, or in view of (20), $F_{x^0}h \geqq 0$.

We can now apply the theorem and it follows that there must exist a y^0 such that

$$1 - y_1^0(t) - y_2^0(t) = 0, \tag{25a}$$

$$y_1^0(t) \geqq 0, \quad y_2^0(t) \geqq 0, \tag{25b}$$

$$y_1^0(t)(c_1(t) - x^0(t)) = y_2^0(t)(c_2(t) - x^0(t)) = 0. \tag{25c}$$

If $t < 1$, then $g_2(x^0(t)) < 0$ so that $y_2^0(t) = 0$, $y_1^0(t) = 1$ and $x^0(t) = c_1(t)$. If $t > 2$, then $g_1(x^0(t)) < 0$ so that $y_1^0(t) = 0$, $y_2^0(t) = 1$ and $x^0(t) = c_2(t)$. Finally, if $1 \leqq t \leqq 2$, then $c_1(t) = c_2(t) = 0$, so that $(y_1^0(t) + y_2^0(t))x^0(t) = 0$ and $x^0(t) = 0$. In this interval we can take $y_1^0(t) = 2 - t$ and $y_2^0(t) = t - 1$. Obviously the x^0 found is indeed the optimum of problem (17).

5.2. As our second application we take a problem considered by Mangasarian [4] involving a partial differential equation. The treatment of this example is not wholly satisfactory, because there are several assumptions concerning the solvability of the differential equations involved for which

we do not specify sufficient conditions. Some of these assumptions are not even stated explicitly.

Let us consider the following problem:

Minimize

$$\int_0^1 \phi^2(s, T)\, ds \Big| \phi_{ss}(s, t) - \phi_t(s, t) = 0, \phi(s, 0) = a(s), \tag{26}$$

$$\phi(1, t) = 0, \phi(0, t) = x(t), \mid x(t) \mid \leqq 1, \qquad t \in [0, T].$$

Here T is a fixed terminal time, $a(s)$ is a prescribed function of s, which, for instance, is the distance along a rod, and $x: t \to x(t)$ is the control. ϕ itself, for instance, is the gradient of the temperature of the rod, so that we want to minimize the heat dissipation at time $t = T$.[1]

We assume that $\phi(s, t)$ can be solved from the equality constraints in terms of s, t and x, so that we can put

$$\phi(s, t) = f(s, t, x). \tag{27}$$

Then

$$Fx = \int_0^1 f^2(s, T, x)\, ds \tag{28}$$

and

$$G(x, y) = \int_0^T \{y_1(t)(x(t) - 1) - y_2(t)(x(t) + 1)\}\, dt. \tag{29}$$

Now let x^0 be a solution of (26) and let $\phi^0(s, t) = f(s, t, x^0)$. Then from (28) we find that

$$F_{x^0} h = \int_0^1 2\phi^0(s, T)\theta(s, T)\, ds, \tag{30}$$

where $\theta(s, t) = f(s, t, x^0 + h) - f(s, t, x^0)$ so that

$$\theta_{ss}(s, t) - \theta_t(s, t) = 0,$$

$$\theta(s, 0) = 0, \qquad \theta(1, t) = 0, \qquad \theta(0, t) = h(t). \tag{31}$$

From (29) we find that

$$G_{x^0}(y)h = \int_0^T (y_1(t) - y_2(t))h(t)\, dt, \tag{32}$$

from which we see that $G_{x^0}(y)$ is bounded. As before we let x, y_1 and y_2

[1] This interpretation of problem (26) was given by H. L. Beckers of Koninklijke/Shell-Laboratorium, Amsterdam.

belong to $L_2[0, T]$ and mean by $y \geqq 0$ that $y_1(t) \geqq 0$ and $y_2(t) \geqq 0$ for (almost) all $t \in [0, T]$.

In order to show that F_{x^0} is bounded we assume that $\psi^0(s, t)$ can be solved from

$$\begin{gathered} \psi^0_{ss}(s, t) + \psi_t{}^0(s, t) = 0, \\ \psi^0(s, T) = 2\phi^0(s, T), \qquad \psi^0(1, t) = 0, \qquad \psi^0(0, t) = 0 \end{gathered} \tag{33}$$

and that $\psi_s{}^0(0, t) \in L_2[0, T]$. Then it follows, if we may apply the divergence theorem to the differential form $-\theta(s, t)\psi^0(s, t)\, ds + \{\theta(s, t)\psi_s{}^0(s, t) - \theta_s(s, t)\psi^0(s, t)\}\, dt$, that

$$F_{x^0} h = \int_0^1 2\phi^0(s, T)\theta(s, T)\, ds = \int_0^T \psi_s{}^0(0, t)h(t)\, dt. \tag{34}$$

From this we see that F_{x^0} is bounded too.

Since $G(x^0, y) = 0$ for $y \geqq 0$ is equivalent to

$$y_1(t)(x^0(t) - 1) = y_2(t)(x^0(t) + 1) = 0, \tag{35}$$

it follows that

$$\begin{aligned} S = \Big\{ s_y \mid s_y x = -\int_0^T (y_1(t) - y_2(t))x(t)\, dt, \, y \geqq 0, \\ y_1(t)(x^0(t) - 1) = y_2(t)(x^0(t) + 1) = 0 \Big\}. \end{aligned} \tag{36}$$

As before, in order to show that S is closed, let a sequence y^i be given such that $y^i \geqq 0$ and $G(x^0, y^i) = 0$, and let $p : t \to p(t) \in R$ be given such that

$$\int_0^T \{y_1{}^i(t) - y_2{}^i(t) + p(t)\}^2\, dt \tag{37}$$

converges to zero. Then we have to show that we can find $y_1(t)$, $y_2(t) \geqq 0$ satisfying (35) and such that $p(t) = -y_1(t) + y_2(t)$. If $|x^0(t)| < 1$ in a subset of $[0, 3]$, then $y_1{}^i(t) = y_2{}^i(t) = 0$, and from (37) it follows that $p(t) = 0$. It also follows that $y_1(t) = y_2(t) = 0$. If $x^0(t) = 1$, then $y_2{}^i(t) = 0$, and from (37) it follows that $p(t) \leqq 0$. Since in this case $y_2(t) = 0$, but there are no restrictions on $y_1(t)$ as far as (35) is concerned, we can put $y_1(t) = -p(t)$. Similarly, if $x^0(t) = -1$, we have that $p(t) \geqq 0$, that $y_1(t) = 0$ and that we can put $y_2(t) = p(t)$. Hence in all cases we have that $p(t) = -y_1(t) + y_2(t)$, so that S is closed.

In order to show that the constraint condition (9) is satisfied, let us first suppose that a function h and an $\epsilon_0 > 0$ exist such that

$$|x^0(t) + \epsilon h(t)| \leqq 1 \tag{38}$$

for all $\epsilon \in [0, \epsilon_0)$ and $t \in [0, T]$. This means that $x^0 + \epsilon h$ is feasible, so that

$$\int_0^1 \{\phi^0(s, T) + \epsilon\theta(s, T)\}^2 \, ds \geqq \int_0^1 \{\phi^0(s, T)\}^2 \, ds,$$

where $\theta(s, t)$ still satisfies (31). Letting ϵ tend to zero, we have from this inequality that

$$F_{x^0} h = \int_0^T \psi_s{}^0(0, t)h(t) \, dt = \int_0^1 2\phi^0(s, T)\theta(s, T) \, ds \geqq 0.$$

Secondly, let h be bounded; hence let there be a β such that $|h(t)| \leqq \beta$, $t \in [0, T]$, and let $G_{x^0}(y)h \leqq 0$ for all $y \geqq 0$, $G(x^0, y) = 0$. From this together with (32) and (35) it follows that if $x^0(t) = 1$ in a subset of $[0, T]$, then $y_2(t) = 0$ and $h(t) \leqq 0$, and if $x^0(t) = -1$, then $h(t) \geqq 0$. Hence if we define

$$U = \{t \,|\, |x^0(t)| = 1, t \in [0, T]\},$$

then (38) holds for all ϵ such that $0 \leqq \epsilon < \epsilon_0 = 2/\beta$, and $t \in U$. Now let

$$V_n = \{t \,|\, |x^0(t)| \leqq 1 - 1/n, t \in [0, T]\}, \quad n = 1, 2, \cdots,$$

and let $\epsilon_n = 1/(n\beta) < \epsilon_0$. Then (38) holds if $0 \leqq \epsilon < \epsilon_n$ and $t \in U \cup V_n$, $n = 1, 2, \cdots$. Further define

$$h_n(t) = \begin{cases} h(t) & \text{if } t \in U \cup V_n, \\ 0 & \text{if } t \notin U \cup V_n, \end{cases} \qquad n = 1, 2, \cdots.$$

From this we see that

$$|x^0(t) + \epsilon h_n(t)| \leqq 1$$

if $0 \leqq \epsilon < \epsilon_n$ and $t \in [0, T]$, and as before it follows that $F_{x^0}h_n \geqq 0, n = 1, 2, \cdots$. Since $h_n(t)$ tends to $h(t)$ if n tends to ∞, $t \in [0, T]$; and since $|h_n(t)| \leqq \beta$ and $\int_0^T \beta^2 \, dt < \infty$, it follows that h_n tends to h in $L_2[0, T]$ (see [8, p. 234]). From this we see that $F_{x^0}h_n$ tends to $F_{x^0}h$, because F_{x^0} is a bounded, linear functional on $L_2[0, T]$. Hence $F_{x^0}h \geqq 0$. Finally, let h be unbounded. Then we define

$$\bar{h}_n(t) = \begin{cases} h(t) & \text{if } |h(t)| \leqq n, \\ 0 & \text{if } |h(t)| > n, \end{cases} \qquad n = 1, 2, \cdots.$$

From the foregoing it then follows that $F_{x^0}\bar{h}_n \geqq 0$, $n = 1, 2, \cdots$. Since $\bar{h}_n(t)$ tends to $h(t)$ if n tends to ∞, $t \in [0, T]$; and since $|\bar{h}_n(t)| \leqq |h(t)|$ and $|h| \in L_2[0, T]$, it follows that $\bar{h}_n$ tends to h in $L_2[0, T]$ (see [8]). Hence $F_{x^0}\bar{h}_n$ tends to $F_{x^0}h$, so that again $F_{x^0}h \geqq 0$. Therefore, if h is such that $G_{x^0}(y)h \leqq 0$ for all $y \geqq 0$, $G(x^0, y) = 0$, then $F_{x^0}h \geqq 0$; hence the constraint condition is satisfied.

From the theorem it now follows that there must exist a y^0 such that

$$\psi_s{}^0(0, t) + y_1{}^0(t) - y_2{}^0(t) = 0, \tag{39a}$$

$$y_1{}^0(t) \geqq 0, \qquad y_2{}^0(t) \geqq 0, \tag{39b}$$

$$y_1{}^0(t)(x^0(t) - 1) = y_2{}^0(t)(x^0(t) + 1) = 0. \tag{39c}$$

Here $\psi_s{}^0(s, t)$ has to be found from

$$\begin{gathered} \psi^0_{ss}(s, t) + \psi_t{}^0(s, t) = 0, \\ \psi^0(s, T) = 2\phi^0(s, T), \qquad \psi^0(1, t) = 0, \qquad \psi^0(0, t) = 0; \end{gathered} \tag{33}$$

and $\phi^0(s, t)$ from

$$\begin{gathered} \phi^0_{ss}(s, t) - \phi_t{}^0(s, t) = 0, \\ \phi^0(s, 0) = a(s), \qquad \phi^0(1, t) = 0, \qquad \phi^0(0, t) = x^0(t), \end{gathered} \tag{40}$$

and x^0 must of course satisfy $| x^0(t) | \leqq 1$.

It was shown by Mangasarian [4] that these conditions are sufficient for x^0 to be a solution of problem (26). In the case when $a(s)$ is given such that problem (26) has an optimal solution x^0 and the assumptions mentioned at the beginning are satisfied, and thus in particular (33) and (40) can be solved such that $\psi_s{}^0(0, t) \in L_2[0, T]$, then these conditions are also necessary. Hence we may expect that a class of functions $a(s)$ exists such that the conditions are both necessary and sufficient.

Acknowledgments. The author would like to thank C. de Ridder and P. J. A. Lekkerkerker both of Koninklijke/Shell-Laboratorium, Amsterdam, for their criticisms when he was in the employ of this laboratory. It was the former who expected conditions (39), (33) and (40) to be necessary for x^0 to be a solution of problem (26), and he outlined a proof based on variational analytical tools, which proof is quite different from ours. The latter, as well as the referee, suggested that the restriction to piecewise continuous functions, which originally was imposed in both examples, could be avoided. The referee also indicated how to do this. The author is also indebted to J. E. Simpson of the University of Kentucky for pointing out to him the most convenient form of the separation theorem for proving the lemma.

REFERENCES

[1] N. Dunford and J. T. Schwartz, *Linear Operators I*, John Wiley, New York, 1964.

[2] H. Halkin and L. W. Neustadt, *General necessary conditions for optimization problems*, Proc. Nat. Acad. Sci. U.S.A., 56 (1966), pp. 1066–1071.

[3] L. Hurwicz, *Programming in linear spaces*, Studies in Linear and Non-linear Programming, K. J. Arrow, L. Hurwicz, H. Uzawa, eds., Stanford University Press, Stanford, California, 1958, Chap. 4.

[4] O. L. MANGASARIAN, Private communication.
[5] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems, I: General theory*, this Journal, 4 (1966), pp. 505–527; *II: Applications*, this Journal, 5 (1967), pp. 90–137.
[6] K. RITTER, *Duality for nonlinear programming in a Banach space*, SIAM J. Appl. Math., 15 (1967), pp. 294–302.
[7] D. L. RUSSELL, *The Kuhn-Tucker conditions in Banach space with an application to control theory*, J. Math. Anal. Appl., 15 (1966), pp. 200–212.
[8] M. E. MUNROE, *Measure and Integration*, Addison-Wesley, Reading, Massachusetts, 1956.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

ON THE STRUCTURE THEORY OF LINEAR DIFFERENTIAL SYSTEMS*

LEONARD WEISS†

1. Introduction. Consider a linear system described by the equations

$$\begin{aligned} \frac{dx}{dt} &= F(t)x(t) + G(t)u(t), \\ y(t) &= H(t)x(t), \end{aligned} \tag{1}$$

where $x(t) \in R^n$, $u(t) \in R^p$, $y(t) \in R^r$, and $F(\cdot)$, $G(\cdot)$,$H(\cdot)$ are matrices which, for convenience, are assumed to be continuous functions of time. (Notation: We shall occasionally denote such a system by "the system $\{F(\cdot), G(\cdot), H(\cdot)\}$.")

A few years ago, Kalman [1], [2] proved a structure theorem for linear time-invariant systems of the form (1) which was motivated by some work of Gilbert [3]. An extension to the time-varying case was stated (without proof) by Kalman in [2] and, based on that, a further extension was subsequently stated (also without proof) by the writer [4], [5] using the concepts of "anticausal" controllability and observability.

The crux of the original statement of the theorem and of the subsequent extensions was the assertion that *given a fixed instant* t of time, there exists a coordinate transformation which converts the coefficient matrices of (1) into a special form *valid at the fixed instant* t.

It is generally recognized that a more satisfying result than the one described above would consist of the ability to assert existence of a "continuously varying" coordinate transformation which effects a structural decomposition of (1) *valid for all* t (or, at least, for all $t \geqq t'$ for some fixed t') into interconnected component subsystems whose mathematical representation and system-theoretic properties are similar to those indicated in the early theorems on structure. Some progress has been made in this direction (see, for instance, [6]), based essentially on a procedure suggested by the writer (see proof of Theorem 9 in [5]) for obtaining globally reduced weighting patterns (and therefore minimal realizations)[1] of linear systems from given nonreduced weighting patterns. However, the suggested procedure involves performing an initial coordinate transformation, so as to

* Received by the editors June 8, 1967, and in revised form April 5, 1968.

† Electrical Engineering Department, University of Maryland, College Park, Maryland 20740. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant 68-1346, and in part by the Alfred P. Sloan Foundation in the form of a research fellowship to the author.

[1] See §5 for definitions.

put the matrix $F(t)$ into a special initial form, which almost inevitably leads to a time-varying structural decomposition for (1) *even when* (1) *is time-invariant.*

In the sequel, we present a set of general structure theorems for (1) which yields the result mentioned in the first sentence of the preceding paragraph, while avoiding certain difficulties inherent in past contributions to the subject. The proofs, which follow a logical pattern similar to that given by Kalman for the time-invariant case, can be algorithmized and show that the structural decomposition obtained for (1) will be time-invariant if (1) is time-invariant.

A precise system-theoretic interpretation of the results is given which clarifies a number of issues which have arisen from past contributions.

Some special applications and corollaries of the main results are discussed.

2. Controllability, observability, reachability, determinability. Define the *adjoint system* to (1) as the system

$$\begin{aligned} \frac{dz}{dt} &= -F'(t)z(t) + H'(t)\tilde{u}(t), \\ \tilde{y}(t) &= G'(t)z(t), \end{aligned} \tag{2}$$

where the prime indicates transpose. It is easily shown that if $\Phi(t, \tau)$ is the transition matrix for (1), i.e., if $\Phi(t, \tau)$ satisfies

$$\begin{aligned} \frac{d\Phi(t, \tau)}{dt} &= F(t)\Phi(t, \tau), \\ \Phi(t, t) &= I, \end{aligned} \tag{3}$$

then $\Phi'(\tau, t)$ is the transition matrix for (2).

DEFINITION 1. A state x_0 of the system (1) is *controllable from time* τ if there exists $t > \tau$, t finite, and a control segment $u_{[\tau,t]}$ such that the phase (τ, x_0) is transferred to the phase $(t, 0)$. Otherwise the state is uncontrollable from τ. If every (no) state is controllable from time τ, the *system is controllable* (*uncontrollable*) *from* τ. Controllability (uncontrollability) of the system from all τ is denoted by *complete controllability* (*uncontrollability*). If $t - \tau$ in the preceding definition can be made arbitrarily small, we speak of *differential controllability* [7].

DEFINITION 2. A state x_0 of the system (1) is *reachable at time* τ (alternate terminology: *anticausal controllable*) if there exist $t < \tau$, t finite, and a control segment $u_{[t,\tau]}$ such that the phase $(t, 0)$ is transferred to the phase (τ, x_0). Otherwise the state is unreachable at τ. If every (no) state is reachable at time τ, the *system is reachable* (*unreachable*) *at* τ. Reachability (unreachability) of the system at all τ is denoted by *complete reachability*

(*unreachability*). If $|t - \tau|$ can be made arbitrarily small, we speak of *differential reachability*.

DEFINITION 3. A state x_0 of the system (1) is *observable from time* τ if, with respect to the adjoint system, that state is controllable from time τ. Otherwise the state is unobservable from τ. Remaining definitions of system observability, unobservability, and complete observability follow dually from Definition 1.

DEFINITION 4. A state x_0 is *determinable at time* τ if, with respect to the adjoint system, that state is reachable at time τ. Otherwise the state is undeterminable at τ. Remaining definitions of system determinability, undeterminability. and complete determinability follow dually from Definition 2.

Remark 1. The terminology used in the above definitions follows, with the exception of "determinability," that adopted in [8]. The above defined concepts of reachability and observability are precisely those described in [5] under the headings of "anticausal" controllability and "anticausal" observability, while determinability corresponds to the older definition of observability.

Remark 2. It will be apparent later on that the four concepts defined above can be used to state four equally valid structure theorems for linear systems. For convenience we restrict attention to the concepts of controllability and determinability.

Define the "controllability" and "determinability" matrices, respectively:

$$C(t, \sigma) = \int_t^\sigma \Phi(t, \eta)G(\eta)G'(\eta)\Phi'(t, \eta)\, d\eta,$$

$$D(t, \sigma) = \int_t^\sigma \Phi'(\eta, t)H'(\eta)H(\eta)\Phi(\eta, t)\, d\eta.$$

Then the following statements are easily proven. (See [9] for the proofs of Theorems 1 and 2. Theorems 3 and 4 then follow by the definitions of observability and determinability.)

THEOREM 1. *A state* x_0 *of* (1) *is controllable from* (*reachable at*) *time* τ *if and only if there exists* $t_1 > \tau$ ($t_1 < \tau$), t_1 *finite, such that* $x_0 \in \mathcal{R}[C(\tau, t_1)]$, *where* $\mathcal{R}[\cdot]$ *denotes* Range $[\cdot]$.

COROLLARY 1.1. *A system* (1) *is controllable from* (*reachable at*) *time* τ *if and only if there exists* $t_1 > \tau$ ($t_1 < \tau$), t_1 *finite, such that* rank $[C(\tau, t_1)] = n$.

THEOREM 2. *Let* $\mathcal{P}_c(t)$ ($\mathcal{P}_r(t)$) *denote the set of states controllable from* (*reachable at*) *time* t. *If* $x \in \mathcal{P}_c(t)$ ($\mathcal{P}_r(t)$) *and* $\tau \leqq t$ ($\tau \geqq t$), *then* $\Phi(\tau, t)x \in \mathcal{P}_c(\tau)$ ($\mathcal{P}_r(\tau)$).

THEOREM 3. *Consider the system* (1) *with* $u(t) \equiv 0$. *Then a state* x_0 *of* (1)

is determinable at (observable from) time τ if and only if there exists some $\sigma < \tau$ ($\sigma > \tau$) such that $x_0 \in \mathcal{R}[D(\tau, \sigma)]$.

COROLLARY 3.1. *A system (1) is determinable at (observable from) time τ if and only if there exists $\sigma < \tau$ ($\sigma > \tau$) such that* rank $[D(\tau, \sigma)] = n$.

THEOREM 4. *Let $\mathcal{Q}_d(t)$ ($\mathcal{Q}_0(t)$) denote the set of states which are determinable at (observable from) time t. If $x \in \mathcal{Q}_d(t)$ ($x \in \mathcal{Q}_0(t)$) and $\tau \geqq t$ ($\tau \leqq t$), then $\Phi'(t, \tau)x \in \mathcal{Q}_d(\tau)$ ($\mathcal{Q}_0(\tau)$).*

A result which plays an important role in our development is given by the following theorem.

THEOREM 5. *The range of $C(t, \sigma)$, $\sigma \geqq t$, is monotone nondecreasing with increasing σ.*

Proof. $C(t, \sigma)$ is a Gramian matrix and has the property that, for any $x \in R^n$, $0 \leqq x'C(t, \sigma_1)x \leqq x'C(t, \sigma_2)x$ for $\sigma_1 \leqq \sigma_2$. Hence,

$$x \in K[C(t, \sigma_2)] \quad \text{implies} \quad x \in K[C(t, \sigma_1)],$$

where $K[\cdot]$ denotes Kernel $[\cdot]$. This implies $K[C(t, \sigma_1)] \supseteq K[C(t, \sigma_2)]$ and therefore, by orthogonal complementation, $\mathcal{R}[C(t, \sigma_1)] \subseteq \mathcal{R}[C(t, \sigma_2)]$.

COROLLARY 5.1. *There exists a positive function $\mu(t)$ such that* $\bigcup_{\sigma > t} \mathcal{R}[C(t, \sigma)] = \mathcal{R}[C(t, t + \mu(t))]$.

COROLLARY 5.2. *The range of $C(t, \sigma)$, $\sigma \leqq t$, is monotone nondecreasing with decreasing σ.*

Obviously, Theorem 5 and Corollaries 5.1, 5.2 hold with $C(t, \sigma)$ replaced by $D(t, \sigma)$. We can therefore write that there exist functions $\mu(t)$, $\nu(t)$, $\omega(t)$, $\rho(t)$ such that

$$\begin{aligned}
\mathcal{P}_c(t) &= \mathcal{R}[C(t, t + \mu(t))], \\
\mathcal{P}_r(t) &= \mathcal{R}[C(t, t - \nu(t))], \\
\mathcal{Q}_0(t) &= \mathcal{R}[D(t, t + \rho(t))], \\
\mathcal{Q}_d(t) &= \mathcal{R}[D(t, t - \omega(t))].
\end{aligned}$$

3. The main structure theorems. One of the primary tools in our derivation of the structure of (1) is the following theorem, which is due to Doležal [10]. (See [11] for an alternate proof plus other applications.)

THEOREM 6. *Let $A(t)$ be an $n \times n$ matrix of C^k functions defined for all t and suppose there is a nonnegative integer $r < n$ such that* rank $A(t) = r$ *for all t. Then there exists an $n \times n$ matrix of C^k functions, $M(t)$, nonsingular for all t, such that $A(t)M(t) = [B(t) \quad 0]$, where $B(t)$ is $n \times r$ and* rank $B(t) = r$ *for all t.*

The principal applications of Theorem 6 in this paper occur by way of the following corollary.

COROLLARY 6.1. *Let $A(t)$ be as in Theorem 6 with the additional property that it is symmetric. Then there exists an $n \times n$ matrix of C^k functions, $T(t)$,*

nonsingular for all t, such that

$$T(t)A(t)T'(t) = \begin{bmatrix} \tilde{A}(t) & 0 \\ 0 & 0 \end{bmatrix} \quad \textit{for all } t,$$

where $\tilde{A}(t)$ *is* $r \times r$ *and* $\operatorname{rank} \tilde{A}(t) = r$ *for all* t.

Proof. Apply Theorem 6 to $A(t)$ with $M(t) = T'(t)$. The result then follows from the observation that $T(t)A(t)T'(t)$ is symmetric.

We now begin development of our main results on the structure of linear systems. The results consist of a series of four theorems, namely, Theorems 7, 8, 9 and 10 in the sequel.

THEOREM 7. *Consider the system* (1) *with controllability matrix* $C(t, t + \mu(t))$ *and suppose* $\operatorname{rank} C(t, t + \mu(t)) = r_c < n$ *for all* t. *Then there exists a diffeomorphic coordinate transformation of the state space of* (1) *with respect to which* (1) *takes on the form*

$$\begin{aligned} \dot{x}_1(t) &= F_{11}(t)x_1(t) + F_{12}(t)x_2(t) + G_1(t)u(t), \\ \dot{x}_2(t) &= F_{22}(t)x_2(t), \\ y(t) &= H_1(t)x_1(t) + H_2(t)x_2(t), \end{aligned} \tag{4}$$

valid for all time, where $x_1(t)$ *is an* r_c*-vector. Moreover, the system*

$$\{F_{11}(\cdot), G_1(\cdot), H_1(\cdot)\}$$

is completely controllable.

Proof. Application of Corollary 6.1 to $C(t, t + \mu(t))$ shows existence of a continuously differentiable $n \times n$ matrix, $T(t)$, which is nonsingular for all t, such that

$$T(t)C(t, t + \mu(t))T'(t) = \begin{bmatrix} \tilde{C}(t) & 0 \\ 0 & 0 \end{bmatrix}, \tag{5}$$

where $\tilde{C}(t)$ is $r_c \times r_c$ and is symmetric, and $\operatorname{rank} \tilde{C}(t) = r_c$ for all t. The right side of (5) represents the controllability matrix for (1) after the transformation $\tilde{x}(t) = T(t)x(t)$ is made. Hence, by Theorem 1, a controllable state in the transformed system at any time has the form $\begin{pmatrix} \tilde{x}_1 \\ 0 \end{pmatrix}$, where $\tilde{x}_1$ is an r_c-vector. From Theorem 2 we have that the transformed transition matrix $\tilde{\Phi}$ has the form (independent of arguments)

$$\tilde{\Phi} = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ 0 & \Phi_{22} \end{bmatrix}, \tag{6}$$

where Φ_{11} is $r_c \times r_c$. It then follows by (3) that regardless of t, $\tilde{F}$ has the form

$$\tilde{F} = \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix}.$$

The above transformed quantities are related to the original by the equations

$$\tilde{\Phi}(t, \tau) = T(t)\Phi(t, \tau)T^{-1}(\tau),$$
$$\tilde{F}(t) = T(t)F(t)T^{-1}(t) + \dot{T}(t)T^{-1}(t).$$

Also, (5) implies that

$$T(t)\Phi(t, \eta)G(\eta)G'(\eta)\Phi'(t, \eta)T'(t) = \begin{bmatrix} K(t, \eta)K'(t, \eta) & 0 \\ 0 & 0 \end{bmatrix}$$

for all t and all $\eta \in [t, t + \mu(t)]$, where $K(t, \eta)$ is $r_c \times p$. Choosing $\eta = t$, it is clear that the above equation implies

$$T(t)G(t) = \begin{bmatrix} K(t, t) \\ 0 \end{bmatrix} \quad \text{for all } t.$$

Using the notation

$$\tilde{G}(t) = T(t)G(t) = \begin{bmatrix} G_1(t) \\ G_2(t) \end{bmatrix},$$

where $G_1(t)$ is $r_c \times p$, we see that $K(t, t) = G_1(t)$ and $G_2(t) = 0$ for all t, which proves the main part of the theorem. The remainder follows by a trivial observation.

THEOREM 8. *Consider* (1) *with determinability matrix* $D(t, t - \omega(t))$ *and let* rank $D(t, t - \omega(t)) = r_d < n$ *for all* t. *Then there exists a diffeomorphic coordinate transformation of* (1) *such that under this transformation,*

$$\begin{aligned} \dot{x}_1(t) &= F_{11}(t)x_1(t) + G_1(t)u(t), \\ \dot{x}_2(t) &= F_{21}(t)x_1(t) + F_{22}(t)x_2(t) + G_2(t)u(t), \\ y(t) &= H_1(t)x_1(t), \end{aligned} \tag{7}$$

valid for all t, *where* $x_1(t)$ *is an* r_d*-vector. Furthermore, the system*

$$\{F_{11}(\cdot), G_1(\cdot), H_1(\cdot)\}$$

is completely determinable.

Proof. Using a completely analogous argument to that in the proof of Theorem 7, but applied to the adjoint system (2), we see that there exists a transformation which takes the transition matrix Φ^A of (2) into the form

$$\Phi^A = \begin{bmatrix} \Phi_{11}^A & \Phi_{12}^A \\ 0 & \Phi_{22}^A \end{bmatrix}.$$

By transposition, the transformed transition matrix of (1) then has the

form

$$\tilde{\Phi} = \begin{bmatrix} \Phi_{11} & 0 \\ \Phi_{12} & \Phi_{22} \end{bmatrix}, \tag{8}$$

and by following the remainder of the proof of Theorem 7, we obtain the desired result.

THEOREM 9. *Consider* (1) *and let the hypotheses of Theorem* 7 *hold so that* (1) *can be transformed into* (4) *with the transformed transition matrix given by* (6). *Define the determinability matrices* D_1 *and* D_2 *by*

$$D_1(t, \sigma) = \int_t^\sigma \Phi_{11}'(\eta, t) H_1'(\eta) H_1(\eta) \Phi_{11}(\eta, t)\, d\eta, \tag{9}$$

$$D_2(t, \sigma) = \int_t^\sigma \Phi_{22}'(\eta, t) H_2'(\eta) H_2(\eta) \Phi_{22}(\eta, t)\, d\eta, \tag{10}$$

and let $\omega_i(t)$, $i = 1, 2$, *be continuously differentiable functions such that*

$$\mathfrak{R}[D_i(t, t - \omega_i(t))] = \bigcup_{\sigma < t} \mathfrak{R}[D_i(t, \sigma)], \qquad i = 1, 2,$$

and let

$$\operatorname{rank} D_1(t, t - \omega_1(t)) = r_{d_1} < r_c \quad \textit{for all } t,$$

$$\operatorname{rank} D_2(t, t - \omega_2(t)) = r_{d_2} < n - r_c \quad \textit{for all } t.$$

Then there exists a diffeomorphic coordinate transformation, $\hat{x}(t) = T(t)x(t)$, *defined for all* t, *which converts* (1) *into a form in which the coefficient matrices have the form*

$$F(t) \to \begin{bmatrix} F^{aa}(t) & 0 & F^{ac}(t) & F^{ad}(t) \\ F^{ba}(t) & F^{bb}(t) & F^{bc}(t) & F^{bd}(t) \\ 0 & 0 & F^{cc}(t) & 0 \\ 0 & 0 & F^{dc}(t) & F^{dd}(t) \end{bmatrix},$$

$$G(t) \to \begin{bmatrix} G^a(t) \\ G^b(t) \\ 0 \\ 0 \end{bmatrix}, \tag{11}$$

$$H(t) \to [H^a(t) \quad 0 \quad H^c(t) \quad 0].$$

Proof. Let $T_1(t)$ be the diffeomorphism which transforms (1) into (4). With the resulting transition matrix given by (6), the determinability

matrix for the transformed system is (with arguments omitted)

$$D(t,\sigma) = \int_t^\sigma \begin{bmatrix} \Phi_{11}' H_1' H_1 \Phi_{11} & \Phi_{11}' H_1' H_1 \Phi_{12} + \Phi_{11}' H_1' H_2 \Phi_{22} \\ \Phi_{12}' H_1' H_1 \Phi_{11} + \Phi_{22}' H_2' H_1 \Phi_{11} & \Phi_{12}' H_1' H_1 \Phi_{12}' + \Phi_{12}' H_1' H_2 \Phi_{22} + \Phi_{22}' H_2' H_1 \Phi_{12} + \Phi_{22}' H_2' H_2 \Phi_{22} \end{bmatrix} ds$$

$$= \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}.$$

From (9) and (10) we have $D_{11} = D_1$, and the last term in D_{22} is D_2. Clearly, D_1 is the determinability matrix for the system $\{F_{11}(\cdot), G_1(\cdot), H_1(\cdot)\}$ in (4). By hypothesis, rank $D_1(t, t - \omega_1(t)) = r_{d_1} < r_c$ for all t. Hence, by Corollary 6.1 there exists a continuously differentiable $r_c \times r_c$ nonsingular matrix $T_2(t)$ such that

$$T_2'(t) D_1(t, t - \omega_1(t)) T_2(t) = \begin{bmatrix} \hat{D}_1(t) & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } t,$$

where $\hat{D}_1(t)$ is $r_{d_1} \times r_{d_1}$ and rank $\hat{D}_1(t) = r_{d_1}$ for all t.

Hence, by Theorem 8, the transformation

$$\begin{bmatrix} x^a(t) \\ x^b(t) \end{bmatrix} = T_2(t)^{-1} x_1(t)$$

transforms (4) into a form in which

$$F_{11}(t) \to \begin{bmatrix} F^{aa}(t) & 0 \\ F^{ba}(t) & F^{bb}(t) \end{bmatrix},$$

$$G_1(t) \to \begin{bmatrix} G^a(t) \\ G^b(t) \end{bmatrix},$$

$$H_1(t) \to [H^a(t) \quad 0].$$

Thus, (4) becomes

$$\begin{bmatrix} x^a(t) \\ x^b(t) \end{bmatrix} = \begin{bmatrix} F^{aa}(t) & 0 \\ F^{ba}(t) & F^{bb}(t) \end{bmatrix} \begin{bmatrix} x^a(t) \\ x^b(t) \end{bmatrix} + T_2(t)^{-1} F_{12}(t) x_2(t) + \begin{bmatrix} G^a(t) \\ G^b(t) \end{bmatrix} u(t), \tag{12}$$

$$\dot{x}_2(t) = F_{22}(t) x_2(t),$$

$$y(t) = H^a(t) x^a(t) + H_2(t) x_2(t),$$

where we have dim $x^a(t) = r_{d_1}$ and dim $x^b(t) = r_c - r_{d_1}$.

Now consider the system $\{F_{22}(\cdot), 0, H_2(\cdot)\}$ in (12). The determinability matrix for this system is D_2. By hypothesis, rank $D_2(t, t - \omega_2(t)) = r_{d_2} < n - r_c$ for all t. Then there exists an $(n - r_c) \times (n - r_c)$ continuously differentiable nonsingular matrix $T_3(t)$ such that

$$T_3'(t)D_2(t, t - \omega_2(t))T_3(t) = \begin{bmatrix} \hat{D}_2(t) & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } t,$$

where $\hat{D}_2(t)$ is $r_{d_2} \times r_{d_2}$ and rank $\hat{D}_2(t) = r_{d_2}$ for all t. The coordinate transformation

$$\begin{bmatrix} x^c(t) \\ x^d(t) \end{bmatrix} = T_3(t)^{-1}x_2(t)$$

transforms (12) into the desired canonical form in which

$$T_2(t)^{-1}F_{12}(t)T_3(t) = \begin{bmatrix} F^{ac}(t) & F^{ad}(t) \\ F^{bc}(t) & F^{bd}(t) \end{bmatrix},$$

$$\dot{T}_3(t)^{-1}T_3(t) + T_3(t)^{-1}F_{22}(t)T_3(t) = \begin{bmatrix} F^{cc}(t) & 0 \\ F^{dc}(t) & F^{dd}(t) \end{bmatrix},$$

$$H_2(t)T_3(t) = [H^c(t) \quad 0],$$

i.e., the system now becomes:

$$\begin{aligned} \dot{x}^a(t) &= F^{aa}(t)x^a(t) + F^{ac}(t)x^c(t) + F^{ad}(t)x^d(t) + G^a(t)u(t), \\ \dot{x}^b(t) &= F^{ba}(t)x^a(t) + F^{bb}(t)x^b(t) + F^{bc}(t)x^c(t) + F^{bd}(t)x^d(t) \\ &\quad + G^b(t)u(t), \\ \dot{x}^c(t) &= F^{cc}(t)x^c(t), \\ \dot{x}^d(t) &= F^{dc}(t)x^c(t) + F^{dd}(t)x^d(t), \\ y(t) &= H^a(t)x^a(t) + H^c(t)x^c(t), \end{aligned} \tag{13}$$

valid for all t, where $\dim x^c(t) = r_{d_2}$ and $\dim x^d(t) = n - r_c - r\,d_2$. This completes the proof of Theorem 9 in which the overall coordinate transformation $T(t)$ is given by

$$T(t) = \begin{bmatrix} T_2(t)^{-1} & 0 \\ 0 & T_3(t)^{-1} \end{bmatrix} T_1(t).$$

Our final theorem in this section, when taken together with Theorems 7, 8 and 9, yields the most general form of the structural decomposition of a given system (1). It is motivated by the possibility that certain state variables in x^d may be determinable as a result of the connection F^{ad}, i.e., the system associated with $\begin{bmatrix} x^a \\ x^d \end{bmatrix}$ may contain a determinable subsystem.

THEOREM 10. *Consider the system* (1) *and let the hypotheses of Theorem* 9 *hold so that* (1) *can be transformed into* (13). *Consider the system*

$$\left\{\begin{bmatrix} F^{aa}(\cdot) & F^{ad}(\cdot) \\ 0 & F^{dd}(\cdot) \end{bmatrix}, \begin{bmatrix} G^{a}(\cdot) \\ 0 \end{bmatrix}, [H^{a}(\cdot) \quad 0]\right\} \tag{14}$$

and call the corresponding determinability matrix $D_3(t, \sigma)$.

Let $\omega_3(t)$ *be a continuously differentiable function such that* $\Re[D_3(t, t - \omega_3(t))] = \bigcup_{\sigma<t} \Re[D_3(t, \sigma)]$, *and let* rank $D_3(t, t - \omega_3(t)) = r_{d_3} < n - r_c - r_{d_2} + r_{d_1}$ *for all* t.

Then there exists a diffeomorphic coordinate transformation which converts (13) *into the form*

$$\begin{aligned}
\dot{\mathbf{x}}^a(t) &= \mathbf{F}^{aa}(t)\mathbf{x}^a(t) + \mathbf{F}^{ac}(t)\mathbf{x}^c(t) + \mathbf{G}^a(t)u(t), \\
\dot{\mathbf{x}}^b(t) &= \mathbf{F}^{ba}(t)\mathbf{x}^a(t) + \mathbf{F}^{bb}(t)\mathbf{x}^b(t) + \mathbf{F}^{bc}(t)\mathbf{x}^c(t) + \mathbf{F}^{bd}(t)\mathbf{x}^d(t) \\
&\qquad + \mathbf{G}^b(t)u(t), \\
\dot{\mathbf{x}}^c(t) &= \mathbf{F}^{cc}(t)\mathbf{x}^c(t), \\
\dot{\mathbf{x}}^d(t) &= \mathbf{F}^{dc}(t)\mathbf{x}^c(t) + \mathbf{F}^{dd}(t)\mathbf{x}^d(t), \\
y(t) &= \mathbf{H}^a(t)\mathbf{x}^a(t) + \mathbf{H}^c(t)\mathbf{x}^c(t),
\end{aligned} \tag{15}$$

valid for all time, where $\dim \mathbf{x}^a(t) = r_{d_1}$, $\dim \mathbf{x}^b(t) = r_c - r_{d_1}$, $\dim \mathbf{x}^c(t) = r_{d_2} + r_{d_3} - r_{d_1}$, $\dim \mathbf{x}^d(t) = n - r_c - r_{d_2} - r_{d_3} + r_{d_1}$.

Note. The general form differs from (13) in that, by means of a further diffeomorphic transformation of coordinates plus a regrouping of state variables ($\dim \mathbf{x}^c(t) > \dim x^c(t)$ and $\dim \mathbf{x}^d(t) < x^d(t)$), the feedback coefficient from the system associated with $\mathbf{x}^d$ to that associated with $\mathbf{x}^a$ becomes identically zero.

Proof of Theorem 10. Let the transition matrix for the system (14) be given by

$$\begin{bmatrix} \Phi^{aa} & \Phi^{ad} \\ 0 & \Phi^{dd} \end{bmatrix},$$

where Φ^{aa} corresponds to F^{aa} and has dimension $r_{d_1} \times r_{d_1}$. Then, omitting arguments in the integrand below, we have

$$D_3(t, \sigma) = \int_t^\sigma \begin{bmatrix} \Phi^{aa'}H^{a'}H^a\Phi^{aa} & \Phi^{aa'}H^{a'}H^a\Phi^{ad} \\ \Phi^{ad'}H^{a'}H^a\Phi^{aa} & \Phi^{ad'}H^{a'}H^a\Phi^{ad} \end{bmatrix} d\eta$$

or

$$D_3(t,\sigma) = \begin{bmatrix} D^{aa}(t, \sigma) & R(t, \sigma) \\ R'(t, \sigma) & Q(t, \sigma) \end{bmatrix}, \tag{16}$$

where $D^{aa}(t, \sigma)$ is the determinability matrix for $\{F^{aa}(\cdot), G^a(\cdot), H^a(\cdot)\}$,

so that rank $D^{aa}(t, t - \omega_1(t)) = r_{d_1}$ for all t. Let $\omega_4(t)$ be a continuously differentiable function such that

$$\mathfrak{R}[Q(t, t - \omega_4(t))] = \bigcup_{\sigma<t} \mathfrak{R}[Q(t, \sigma)]$$

and let $\omega(t)$ be likewise continuously differentiable such that

$$\omega(t) > \max_i (\omega_i(t)), \qquad i = 1, 2, 3, 4,$$

for all t. It is easy to show that

$$\mathfrak{R}[D^{aa}(t, t - \omega(t))] \supseteq \mathfrak{R}[R(t, t - \omega(t))]$$

(e.g., by showing that the above holds under orthogonal complementation). Hence there exists a matrix $K(t)$ such that

$$D^{aa}(t, t - \omega(t))K(t) = R(t, t - \omega(t)).$$

By Theorem 6 it follows that $K(\cdot) \in C^1$.

Now define the $(n - r_c - r_{d_2} + r_{d_1}) \times (n - r_c - r_{d_2} + r_{d_1})$ matrix $T_4(t)$ by the formula

$$T_4(t) = \begin{bmatrix} I_{r_{d_1}} & -K(t) \\ 0 & I_{n-r_c-r_{d_2}} \end{bmatrix}, \tag{17}$$

where $I_{r_{d_1}}$ denotes the $r_{d_1} \times r_{d_1}$ identity matrix. Clearly, $T_4(t)$ has the same smoothness properties as $K(t)$. From (16) and (17) we obtain (omitting arguments on the right-hand side)

$$T_4'(t)D_3(t, t - \omega(t))T(t) = \begin{bmatrix} D^{aa} & 0 \\ 0 & Q_1 \end{bmatrix}, \tag{18}$$

where $Q_1 = Q - R'K$ is nonnegative definite, and it follows by hypothesis that rank $Q_1(t, t - \omega(t)) = r_{d_3} - r_{d_1}$ *for all* t.

Applying Corollary 6.1, let $T_5(t)$ be an $(n - r_c - r_{d_2}) \times (n - r_c - r_{d_2})$ continuously differentiable nonsingular matrix such that

$$T_5'(t)Q_1(t, t - \omega(t))T_5(t) = \begin{bmatrix} P_1(t) & 0 \\ 0 & 0 \end{bmatrix}, \tag{19}$$

where $P_1(t)$ is $(r_{d_3} - r_{d_1}) \times (r_{d_3} - r_{d_1})$ and rank $Q_1(t, t - \omega(t)) = \text{rank } P_1(t)$ for all t. Then the coordinate transformation

$$\begin{bmatrix} \mathbf{x}^a(t) \\ \mathbf{x}^d(t) \end{bmatrix} = \hat{T}(t) \begin{bmatrix} x^a(t) \\ x^d(t) \end{bmatrix}, \tag{20}$$

with

$$\hat{T}(t) = \begin{bmatrix} I_{r_{d_1}} & 0 \\ 0 & T_5(t)^{-1} \end{bmatrix} T_4(t)^{-1} = \begin{bmatrix} I_{r_{d_1}} & K \\ 0 & T_5^{-1} \end{bmatrix}, \tag{21}$$

has the effect of transforming the determinability matrix D_3 for the system (14) into the form

$$\left[\begin{array}{c|cc} D^{aa} & 0 & \\ \hline 0 & \left[\begin{array}{cc} P_1 & 0 \\ 0 & 0 \end{array}\right] & \end{array}\right] \tag{22}$$

for all t. It is easy to check that $\hat{T}(t)^{-1}$ has the same form as $\hat{T}(t)$; in fact,

$$\hat{T}(t)^{-1} = \begin{bmatrix} I_{r_{d_1}} & -KT_5 \\ 0 & T_5 \end{bmatrix} \tag{23}$$

for all t, where the dimensions of the "0" are $n - r_c - r_{d_2} \times r_{d_1}$.

Therefore, the transformed state coefficient matrix of (14) is given by

$$\hat{F}_{a,d}(t) = \hat{T}(t) \begin{bmatrix} F^{aa}(t) & F^{ad}(t) \\ 0 & F^{dd}(t) \end{bmatrix} \hat{T}(t)^{-1} + \dot{\hat{T}}(t)\hat{T}(t)^{-1}$$

and has the form

$$\hat{F}_{a,d}(t) = \begin{bmatrix} \hat{F}^{aa}(t) & \hat{F}^{ad}(t) \\ 0 & \hat{F}^{dd}(t) \end{bmatrix},$$

and the corresponding transition matrix $\hat{\Phi}_{a,d}(t, \tau)$ also has this form, i.e.,

$$\hat{\Phi}_{a,d}(t, \tau) = \begin{bmatrix} \Phi^{aa}(t, \tau) & \hat{\Phi}^{ad}(t, \tau) \\ 0 & \hat{\Phi}^{dd}(t, \tau) \end{bmatrix} \tag{24}$$

for all t, τ.

Now partition $\hat{\Phi}^{ad}$ and $\hat{\Phi}^{dd}$ as follows:

$$\hat{\Phi}^{ad} = [\hat{\Phi}_1^{ad}\ \hat{\Phi}_2^{ad}],$$

where $\hat{\Phi}_1^{ad}$ is $r_{d_1} \times n - r_c - r_{d_2} - r_{d_3} + r_{d_1}$, and

$$\hat{\Phi}^{dd} = \begin{bmatrix} \hat{\Phi}_{11}^{dd} & \hat{\Phi}_{12}^{dd} \\ \hat{\Phi}_{21}^{dd} & \hat{\Phi}_{22}^{dd} \end{bmatrix},$$

where $\hat{\Phi}_{12}^{dd}$ is $(n - r_c - r_{d_2} - r_{d_3} + r_{d_1}) \times (r_{d_3} - r_{d_1})$ and the remaining matrices are conformable with this.

This partition corresponds to a partition of the vector x^d as $\begin{bmatrix} x_1^d \\ x_2^d \end{bmatrix}$, where $\dim x_1^d = r_{d_3} - r_{d_1}$. Then the transpose of (24) becomes

$$\hat{\Phi}'_{a,d} \begin{bmatrix} \Phi^{aa\prime} & 0 & 0 \\ \hat{\Phi}_1^{ad\prime} & \hat{\Phi}_{11}^{dd\prime} & \hat{\Phi}_{21}^{dd\prime} \\ \hat{\Phi}_2^{ad\prime} & \hat{\Phi}_{12}^{dd\prime} & \hat{\Phi}_{22}^{dd\prime} \end{bmatrix}. \tag{25}$$

By (22), states which are determinable at any fixed time t must, under the new coordinate system, have the form

$$\begin{bmatrix} \hat{x}_1(t) \\ 0 \end{bmatrix},$$

where $\dim \hat{x}_1(t) = r_{d_3}$. From Theorem 4 and (25) it follows that

$$[\hat{\Phi}_2^{ad'}(t)\ \hat{\Phi}_{12}^{dd'}(t)] = 0$$

for all t. Transposing, and using (3), we find that

$$\hat{F}_{a,d} = \begin{bmatrix} F^{aa} & \hat{F}_1^{ad} & 0 \\ 0 & \hat{F}_{11}^{dd} & 0 \\ 0 & \hat{F}_{21}^{dd} & \hat{F}_{22}^{dd} \end{bmatrix}$$

for all t.

Before giving the final regrouping of terms, it remains to find the "output" coefficient matrix of (14) under the new coordinate system. This is given by

$$\hat{H}(t) = [H^a(t)\ 0]\hat{T}(t)^{-1}$$

and so, from (23) (with arguments omitted),

$$\hat{H} = [H^a - H^a K T_5],$$

and from Theorem 8 it is clear that $\hat{H}$ takes on the form

$$\hat{H} = [H^a\ H_1{}^d\ 0],$$

where $H_1{}^d$ is $r \times r_{d_3} - r_{d_1}$.

We now define the following quantities:

$$\mathbf{x}^a = x^a, \qquad \mathbf{x}^b = x^b,$$

$$\mathbf{x}^c = \begin{bmatrix} x^c \\ x_1{}^d \end{bmatrix}, \quad \mathbf{x}^d = x_2{}^d,$$

$$\mathbf{F}^{cc} = [F^{cc}\ \hat{F}_{11}^{dd}], \quad \mathbf{H}^c = [H^c\ H_1{}^d],$$

$$\mathbf{F}^{dc} = [F^{dc}\ \hat{F}_1{}^{ad}], \quad \mathbf{F}^{dd} = \hat{F}_{22}^{dd},$$

$$\mathbf{F}^{ac} = [F^{ac}\ \hat{F}_1{}^{ad}], \quad \mathbf{F}^{aa} = F^{aa}, \quad \mathbf{F}^{bb} = F^{ba},$$

and if we write

$$F^{bd}\ x^d = [F_1{}^{bd}\ F_2{}^{bd}] \begin{bmatrix} x_1{}^d \\ x_2{}^d \end{bmatrix},$$

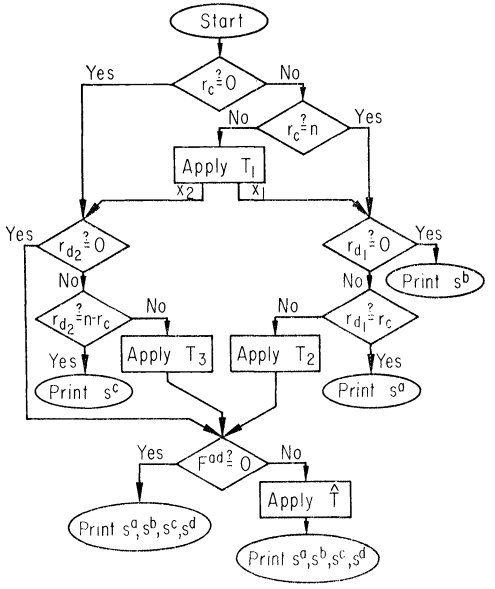


FIG. 1. *Flow chart for structural decomposition*

then we define

$$\mathbf{F}^{bc} = [F^{bc}\, F_1^{bd}], \quad \mathbf{F}^{bd} = [F_2^{bd}].$$

Finally, $\mathbf{F}^{bb} = F^{bb}, \mathbf{H}^a = H^a, \mathbf{G}^a = G^a, \mathbf{G}^b = G^b$. The theorem is thus proved.

The procedure we have given for obtaining the structural decomposition of (1) can be algorithmized as indicated by the flow diagram in Fig. 1.

It should be emphasized that our procedure for structural decomposition is "symmetric" from a number of points of view. For example, just as Theorem 8 is a dual result to Theorem 7, we could have given a completely dual procedure for obtaining a form consonant with (15). That is, one can easily write the dual to Theorems 9 and 10 which would begin with the application of Theorem 8 and would replace the matrices D_1, D_2, D_3 with matrices C_1, C_2, C_3, etc.

If, from the point of view of procedural logic, the order of steps in the proof of the dual theorem is the same as in the original, then the coefficient matrices in this case would take on the form

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}^{aa} & \mathbf{F}^{ab} & 0 & 0 \\ 0 & \mathbf{F}^{bb} & 0 & 0 \\ \mathbf{F}^{ca} & \mathbf{F}^{cb} & \mathbf{F}^{cc} & \mathbf{F}^{cd} \\ 0 & \mathbf{F}^{db} & 0 & \mathbf{F}^{dd} \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}^a \\ 0 \\ \mathbf{G}^c \\ 0 \end{bmatrix},$$

$$\mathbf{H} = [\mathbf{H}^a \ \mathbf{H}^b \ 0 \ 0].$$

In addition to all this "dual" symmetry, Theorems 1–5 indicate that the same type of structural decomposition is obtained if "controllability" is replaced by "reachability" and/or "determinability" is replaced by "observability." Hence, Theorems 9 and 10 as well as their duals are each representatives for a set of *four* structural decomposition theorems.

To avoid confusion in the sequel, our discussion and interpretation of the results of this section are given only with reference to the actual procedure adopted in Theorems 9 and 10 to obtain (15). On the basis of our comments above, the reader can easily supply the interpretations for all the remaining approaches.

Remark 3. The overall coordinate transformation which produces the general structural decomposition of an arbitrary system (1) is represented by the matrix

$$T(t) = \begin{bmatrix} I_{r_{d_1}} & 0 \\ 0 & T_5(t)^{-1} \end{bmatrix} T_4(t)^{-1} \begin{bmatrix} T_2(t)^{-1} & 0 \\ 0 & T_3(t)^{-1} \end{bmatrix} T_1(t).$$

Remark 4. For the special case when (1) is *time-invariant*, all applications of Corollary 6.1 will involve time-invariant transformations so that the procedure given in the proofs of Theorems 9 and 10 clearly leads to a *time-invariant* structural decomposition.

Remark 5. If $X(t)$ is any fundamental matrix solution of $\dot{x}(t) = F(t)x(t)$, then the coordinate transformation $\hat{x}(t) = [X(t)]^{-1}x(t)$, followed by our (now vastly simplified) decomposition procedure, leads to a structure in which $\mathbf{F}(\cdot) \equiv 0$. However, since $X(t)$ is constant with time only when $F(\cdot) \equiv 0$, the transformed system obtained this way *will almost always be time-varying even when the original system* (1) *is not.*

Remark 6. Since our structural decomposition holds for all time, the original statement in [2] on structural decomposition of time-varying systems at each fixed instant of time, with an appropriate modification, is an immediate consequence of our results.

Remark 7. With respect to the question of alternative approaches to structural decomposition, one can easily show that the method detailed in [6] can be adapted to obtain the structure theorems in this paper. A prime advantage of the present method which emphasizes Doležal's theorem, apart from its merits based on algorithmic considerations alone, stems from

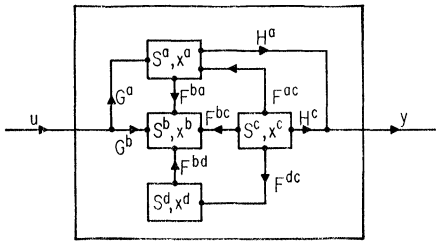


FIG. 2. *Block diagram of structural components*

the fact that Doležal's theorem is also applicable to the problem of obtaining structural decomposition independent of knowledge of the transition matrix. Hence, our approach points the way to and is part of a unified theory of structural decomposition under diverse hypotheses.

4. System-theoretic interpretation of structural decomposition. Pictorially, the decomposition (15) can be viewed as in Fig. 2, which shows four interconnected systems S^a, S^b, S^c, S^d enclosed in "boxes" labeled with associated state vectors. If, as is natural, *we view the interconnecting lines inside the large "box" as input and output lines for the structural components*, then the following result is readily discernible from individual examination of each structural component in (15) plus reference to the proofs of Theorems 9 and 10.

COROLLARY 10.1.

(i) S^a *is completely controllable and completely determinable,*

(ii) S^b *is completely controllable and completely undeterminable,*

(iii) S^c *is completely uncontrollable and completely determinable,*

(iv) S^d *is completely uncontrollable and completely undeterminable.*

Remark 8. If the matrices $F(\cdot)$, $G(\cdot)$, $H(\cdot)$ in (1) are analytic functions of time, the ranks of $C(t, t + \mu(t))$, $D_i(t, t - \omega_i(t))$, $i = 1, 2, 3$, will be constant everywhere in the t-domain. Hence, the system-theoretic interpretation of the structural decomposition of a system with analytic coefficients is given by Corollary 10.1. This provides the proof for assertions concerning analytic systems which were made in [2] and [5].

Remark 9. It is readily apparent that the series of coordinate transformations leading to (15) can be associated with projections onto the ranges or null spaces of the matrices C and D_i, $i = 1, 2, 3$. We can therefore view the decomposition (15) as being associated with a direct sum decomposition of the state space

$$x(t) = x_1(t) \oplus x_2(t) \oplus x_3(t) \oplus x_4(t)$$

in which, under the overall coordinate transformation, the various com-

ponents have the form

$$x_1(t) \to \begin{bmatrix} \mathbf{x}^a(t) \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad x_2(t) \to \begin{bmatrix} 0 \\ \mathbf{x}^b(t) \\ 0 \\ 0 \end{bmatrix},$$

$$x_3(t) \to \begin{bmatrix} 0 \\ 0 \\ \mathbf{x}^c(t) \\ 0 \end{bmatrix}, \qquad x_4(t) \to \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathbf{x}^d(t) \end{bmatrix}.$$

5. Minimal realizations, weighting patterns, impulse responses and causality. Our results on structure form a basis for clarification of a number of results on realization of systems for input/output quantities, as discussed in [5].

We begin with the solution of (1), written in the form

$$y(t) = H(t)\Phi(t, t_0)x_0 + \int_{t_0}^{t} W(t, \tau)u(\tau)\, d\tau, \tag{26}$$

where x_0 is the state of the system at time t_0, and

$$W(t, \tau) = H(t)\Phi(t, \tau)G(\tau) \quad \text{for all} \quad t, \tau \tag{27}$$

and is called the weighting pattern [4] for (1). We shall consider some aspects of the problem of realizing a given function $W(t, \tau)$ by a system (1) and relating $W(t, \tau)$ to the structure of that system. We shall also discuss, later on, similar problems with respect to the "causal impulse response" function [5] $W_c(t, \tau)$ defined by

$$W_c(t, \tau) = \begin{cases} W(t, \tau) & \text{for} \quad t \geqq \tau, \\ 0 & \text{for} \quad t < \tau. \end{cases} \tag{28}$$

(The distinction between W_c and W from a system-theoretic point of view is important and was first recognized, apparently, by the writer (see [4], [5]).)

Consider the following definitions (introduced, with the exception of Definition 7, in [4], [5]).

DEFINITION 5. A weighting pattern $W(t, \tau)$ is in *reduced form on an interval* (α_1, α_2) if the rows of $\Phi(t_0, \cdot)G(\cdot)$ and the columns of $H(\cdot)\Phi(\cdot, t_0)$ are linearly independent functions on the interval (α_1, α_2) independent of t_0.

DEFINITION 6. A weighting pattern is *globally reduced* if it is in reduced form on the entire interval (in this case $(-\infty, \infty)$) of definition of the system (1).

DEFINITION 7. A weighting pattern *has the property DCDO on an interval* (α_1, α_2) if it is in reduced form on every subinterval of (α_1, α_2) of positive length.

DEFINITION 8. The *order* of a globally reduced weighting pattern (27) is the number of columns of $H(\cdot)\Phi(\cdot, t_0)$ = number of rows of $\Phi(t_0, \cdot)G(\cdot)$.

DEFINITION 9. A *realization* of a globally reduced weighting pattern $W(t, \tau)$ is a dynamical system (1) whose weighting pattern can be reduced to $W(t, \tau)$.

DEFINITION 10. If the dimension of the state space of the realization equals the order of $W(t, \tau)$, the realization is *globally reduced* or is *minimal.*

The following three lemmas were proved in [4], [5] (Lemma 1 was first proved in [2] for the case of causal impulse responses).

LEMMA 1. *An $r \times p$ matrix function of two variables, $W(t, \tau)$, is a weighting pattern for an n-dimensional system* (1) *if and only if W can be factored as*

$$W(t, \tau) = \Phi(t)\Theta(\tau) \quad \text{for all} \quad t, \tau.$$

LEMMA 2. *Every weighting pattern has a globally reduced form.*

LEMMA 3. *A minimal realization of a globally reduced weighting pattern $W(t, \tau)$ has the lowest dimension of all globally reduced realizations of $W(t, \tau)$.*

In order to discuss the significance of minimal realizations from the point of view of the concepts of controllability, determinability, etc., the following definition is given (see [2]).

DEFINITION 11. Two n-dimensional linear systems S_1, S_2 of the form (1) are *algebraically equivalent* if there exists an $n \times n$ nonsingular, continuously differentiable matrix $T(t)$ such that

$$\begin{aligned} F_{S_1}(t) &= T(t)^{-1}F_{S_2}(t)T(t) - T(t)^{-1}\dot{T}(t), \\ G_{S_1}(t) &= T(t)^{-1}G_{S_2}(t), \\ H_{S_1}(t) &= H_{S_2}(t)T(t) \end{aligned}$$

for all t.

Algebraic equivalence implies that the two systems are related by a coordinate transformation $x_{S_2}(t) = T(t)x_{S_1}(t)$ and is of interest because of the following results.

LEMMA 4. *Weighting patterns are invariant under algebraic equivalence.*

Proof. The proof follows from (27) and Definition 11.

LEMMA 5. *Points of time from which a system is controllable (or observable) or at which a system is reachable (or determinable) are invariant under algebraic equivalence.*

Proof (for controllability only; the remainder follows analogously). Under algebraic equivalence, we have the correspondence

$$C(t, t + \mu(t)) \to T(t)C(t, t + \mu(t))T'(t) = \hat{C}(t, t + \mu(t))$$

and so, rank $C(t, t + \mu(t)) = n$ implies rank $\hat{C}(t, t + \mu(t)) = n$.

The above result coupled with the well-known theorem below (see [5] for statement of theorem and [6] for published proof) shows that all minimal realizations of a given globally reduced weighting pattern have essentially the same behavior from the point of view of controllability, determinability, etc.

THEOREM 11. *Any two minimal realizations of a given globally reduced weighting pattern are algebraically equivalent.*

We are now in a position to relate the results above to the earlier ones on structural decomposition. The first fact of significance is given by the following lemma.

LEMMA 6. *Consider a system* (1) *with structural decomposition* (15). *Then the weighting pattern* (27) *of* (1) *is given by*

$$W(t, \tau) = \mathbf{H}^a(t)\mathbf{\Phi}^{aa}(t, \tau)\mathbf{G}^a(\tau), \tag{29}$$

where $\mathbf{\Phi}^{aa}$ *is the transition matrix corresponding to* $\mathbf{F}^{aa}$ *in* (15).

Proof. The proof follows from (27) and (15) plus the fact that the transition matrix $\mathbf{\Phi}$ for (15) must have the same pattern of zeros as does F.

LEMMA 7. *The subsystem* S^a *in Fig. 2 is a minimal realization of the weighting pattern for the overall system.*

Proof. The right side of (29) is the weighting pattern for S^a and is globally reduced. It is easy to check that the order of this weighting pattern is the dimension of $\mathbf{x}^a$.

From Theorem 11 and Lemma 5 we obtain the two theorems below.

THEOREM 12. *All minimal realizations of a given globally reduced weighting pattern are controllable* (*or observable*) *from all* $t < t'$ *for some sufficiently small* t' *and are reachable* (*or determinable*) *at all* $t > t''$ *for some sufficiently large* t''.

THEOREM 13. *All minimal realizations of a weighting pattern with the property DCDO are differentially controllable, reachable, determinable, observable.*

Our objective now is to discuss the concept of impulse response and its relationship to the preceding material in this section. The psychological causes for past confusion between the impulse response concept (which arises from the incorporation of the principle of causality into a system's input/output relation constructed from postulating the superposition principle) and the weighting pattern (which is generated from the solution of a differential equation and is therefore further removed from the process of model building) are not pertinent to the present discussion. We simply note that if one insists upon using differential equations as the basic mathematical model of a system, then the weighting pattern is a fundamental object for study, and that the impulse response is of interest primarily because the only part of $W(t, \tau)$ that can be experimentally determined (or approximated) is that for the case $t \geqq \tau$.

The causal impulse response was defined earlier. It will be useful to have also the following concepts.

DEFINITION 12. The *anticausal impulse response* of a system (1) with weighting pattern $W(t, \tau)$ is a function $W_a(t, \tau)$ such that

$$W_a(t, \tau) = \begin{cases} W(t, \tau) & \text{for } t \leqq \tau, \\ 0 & \text{for } t > \tau. \end{cases}$$

DEFINITION 13. A *realization* of a causal (anticausal) impulse response $W_c(t, \tau)$ ($W_a(t, \tau)$) is a system (1) whose causal (anticausal) impulse response is $W_c(t, \tau)$ ($W_a(t, \tau)$).

By the definition of impulse response, it is clear from Lemma 1 that any $W_c(t, \tau)$ can be factored as

$$W_c(t, \tau) = \begin{cases} \Psi(t)\Theta(t) & \text{for } t \geqq \tau, \\ 0 & \text{for } t < \tau, \end{cases} \tag{30}$$

and a similar result (with reversed ordering for t and τ on the right side of (30)) holds for $W_a(t, \tau)$. We then have the following definitions.

DEFINITION 14. A causal impulse response (30) is *globally reduced* if, for some $t_1 > -\infty$, the rows of $\Theta(\cdot)$ are linearly independent over $(-\infty, t_1]$ while the columns of $\Psi(\cdot)$ are linearly independent over $[t_1, \infty)$.

DEFINITION 15. An anticausal impulse response is *globally reduced* if, for some $t_2 < \infty$, the rows of $\Theta(\cdot)$ are linearly independent over $[t_2, \infty)$ while the columns of $\Psi(\cdot)$ are linearly independent over $(-\infty, t_2]$.

DEFINITION 16. The function $W(t, \tau) = \Psi(t)\Theta(\tau)$, defined for all t, τ, is the *naturally induced weighting pattern associated with the globally reduced impulse response* $W_c(t, \tau)$ (or $W_a(t, \tau)$) given by (30).

DEFINITION 17. A realization of a globally reduced impulse response is *minimal* if its dimension equals the order of the weighting pattern naturally induced by that impulse response.

Remark 10. It should be emphasized that many different weighting patterns may be associated with a given (causal or anticausal) impulse response. (For example, let $\Theta(\tau) = 0$ for all $\tau \leqq 0$ in (30). Then $\Psi(t)$ can be arbitrarily chosen for $t \leqq 0$ without affecting $W_c(t, \tau)$.) It is therefore obvious that two minimal realizations of a given globally reduced impulse response may not be algebraically equivalent.

The condition under which algebraic equivalence is preserved is given by the result below, which follows from Theorem 11.

THEOREM 14. *Two minimal realizations of a globally reduced causal (anticausal) impulse response are algebraically equivalent if and only if they have the same anticausal (causal) impulse response.*

As far as system-theoretic properties of minimal realizations of impulse responses are concerned, we have the following from Definitions 14–17 and Theorem 12.

THEOREM 15. (i) *A minimal realization of a globally reduced causal impulse response is reachable at all* $t > t_1$ *and is observable from all* $t < t_1$, *where* t_1 *is as given in Definition* 14.

(ii) *A minimal realization of a globally reduced anticausal impulse response is controllable from all* $t < t_2$ *and is determinable at all* $t > t_2$, *where* t_2 *is as given in Definition* 15.

Finally, we briefly discuss the subject of realizations of weighting patterns on a finite time interval.

DEFINITION 18. If a globally reduced weighting pattern W of a system defined on an interval (α_1, α_2) is not in reduced form on some subinterval $(\beta_1, \beta_2) \subset (\alpha_1, \alpha_2)$, the reduced form of W on (β_1, β_2) is called a *local reduction* of W on (β_1, β_2).

DEFINITION 19. A minimal realization of a local reduction of weighting pattern W on (β_1, β_2) is a *local minimal realization* of W on (β_1, β_2).

LEMMA 8. *The order of a local reduction of* W *is* $\leqq$ *that of a global reduction of* W. (*Hence, a local minimal realization of* W *has a state space whose dimension is* $\leqq$ *that of a global minimal realization.*)

Proof. The proof follows from (27) plus the fact that functions which are linearly independent on an interval I may be linearly dependent on a subinterval $J \subset I$.

THEOREM 16. *Let* S_i, $i = 1, 2$, *be respective minimal realizations of the same globally reduced weighting pattern. Let* $S_i \mid [\gamma, \delta]$ *be the system* S_i *restricted to a fixed time interval* $[\gamma, \delta]$. *Let* $n_i(t, \gamma, \delta)$ *be the dimension of* $S_i \mid [\gamma, \delta]$. *Then* $n_1(t, \gamma, \delta) = n_2(t, \gamma, \delta)$ *for all* t, γ, δ.

Proof. Suppose there exist γ_1, δ_1, t_1 such that $n_1(t_1, \gamma_1, \delta_1) \neq n_2(t_1, \gamma_1, \delta_1)$. Then, by continuity, there exists $(\mu, \nu) \subset [\gamma_1, \delta_1]$ with $t_1 \in (\mu, \nu)$ such that $n_1(t, \gamma_1, \delta_1) \neq n_2(t, \gamma_1, \delta_1)$ for all $t \in (\mu, \nu)$. But this implies that the globally reduced weighting pattern associated with S_1, S_2 has two local reductions on (μ, ν) of *different order*, and it follows from Theorem 11 that this is impossible.

Further discussion about impulse responses and their realizations may be found in [5] and also in [12].

Acknowledgment. The author wishes to thank Professor P. L. Falb and W. M. Wonham for a number of valuable discussions on an earlier version of this paper.

REFERENCES

[1] R. E. KALMAN, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci. U. S. A., 48 (1962), pp. 596–600.

[2] ———, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[3] E. G. GILBERT, *Controllability and observability in multivariable control systems*, this Journal, 1 (1963), pp. 128–151.

[4] L. Weiss, *Weighting patterns and the controllability and observability of linear systems*, Proc. Nat. Acad. Sci. U. S. A., 51 (1964), pp. 1122–1127.

[5] L. Weiss and R. E. Kalman, *Contributions to linear system theory*, Internat. J. Engrg. Sci., 3 (1965), pp. 141–171.

[6] D. C. Youla, *The synthesis of linear dynamical systems from prescribed weighting patterns*, SIAM J. Appl. Math., 14 (1966), pp. 527–549.

[7] L. Weiss, *The concepts of differential controllability and differential observability*, J. Math. Anal. Appl., 10 (1965), pp. 442–449; *Correction and addendum*, Ibid., 13 (1966), pp. 577–578.

[8] R. E. Kalman, P. L. Falb and M. A. Arbib, *Lecture Notes on Modern System Theory*, ASEE–NASA Summer Faculty Institute, Stanford University, 1966.

[9] R. E. Kalman, Y. C. Ho and K. S. Narendra, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1962), pp. 189–213.

[10] V. Doležal, *The existence of a continuous basis of a certain linear subsp ace of E_r which depends on a parameter*, Časopis Pěst. Mat., 89 (1964), pp. 4 66–468.

[11] L. Weiss and P. L. Falb, *Doležal's theorem, linear algebra with continuously parametrized elements, and time-varying systems*, Math. Systems Theory to appear.

[12] C. A. Desoer and P. P. Varaiya, *The minimal realization of a nonanticipative impulse response matrix*, SIAM J. Appl. Math., 15 (1967), pp. 754–764.

SIAM J. CONTROL
Vol. 6, No. 4, 1968
Printed in U.S.A.

ON A MATRIX RICCATI EQUATION OF STOCHASTIC CONTROL*

W. M. WONHAM†

1. Introduction. The object of this paper is to discuss a generalized version of the matrix Riccati and matrix quadratic equations, which arise in problems of stochastic control and filtering. The properties obtained include existence, uniqueness and asymptotic behavior, and contain as special cases some (but not all) of the results reported in [1], [2]. We refer in particular to [2] for a detailed review and bibliography of the "standard" equation for the linear regulator problem.

The present generalization consists in the addition of a linear positive operator to the linear terms of the standard Riccati equation and, in certain instances, a weakening of the usual hypothesis of complete observability to observability of unstable modes (detectability).

The proofs given here are simple applications of Bellman's principle of quasi-linearization ("approximation in policy space") [5] and of a known monotone convergence property of symmetric matrices. In this way the discussion becomes unified and straightforward. Applications to control and filtering are indicated in §6.

2. Notation and summary. In the following, all vectors and matrices have real elements except where otherwise stated. A, B, C, K are matrices of dimension respectively $n \times n$, $n \times m$, $p \times n$, and $m \times n$; N, P, Q denote symmetric matrices of dimension respectively $m \times m$, $n \times n$, and $n \times n$; it will always be assumed that N is positive definite. A' denotes the transpose of A, and I is the identity matrix. Matrix functions of time t which are assumed as data are Lebesgue measurable and bounded in norm on every finite subinterval of their domain of definition. In particular $N(t)^{-1}$ is so bounded.

If P is positive (semi-)definite, we write $P > 0$ ($P \geqq 0$); $P > Q$ means $P - Q > 0$, etc. If P is symmetric, the Euclidean norm $|P|$ is the absolute value of the numerically largest eigenvalue of P; thus $-|P|I \leqq P \leqq |P|I$.

Π will denote a (possibly t-dependent) positive linear map of the class of symmetric $n \times n$ matrices into itself: that is, $\Pi = \Pi(t, P)$ is measurable in (t, P), linear in P, and $P \geqq 0$ implies $\Pi(t, P) \geqq 0$. In the case where A, B, K

* Received by the editors November 3, 1967, and in revised form March 8, 1968.

† National Aeronautics and Space Administration, Electronics Research Center, Cambridge, Massachusetts 02139. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant AF-AFOSR-693-67, and in part by the National Science Foundation, Engineering, under Grant GK-967.

and Π are independent of t, the condition

$$\inf_K \left| \int_0^\infty e^{t(A-BK)'} \Pi(I) e^{t(A-BK)}\, dt \right| < 1 \tag{2.1}$$

will be important later. It expresses the fact that Π is not too large.

Let A, B be constant. The *controllability matrix* of (A, B) is the $n \times mn$ matrix

$$\Gamma(A, B) = [B, AB, \cdots, A^{n-1}B].$$

The pair (A, B) is *controllable* if the rank of Γ is n. If (A, B) is controllable, so is $(A - BK, B)$ for every matrix K.

The pair of constant matrices (A, B) is *stabilizable* if there exists a contant matrix K such that $A - BK$ is stable (i.e., all its eigenvalues have negative real parts). Let the minimum polynomial $\phi(\lambda)$ of A be factored as $\phi(\lambda) = \phi^+(\lambda)\phi^-(\lambda)$, where all zeros of $\phi^+(\lambda)$ lie in the closed right-half complex plane and all zeros of $\phi^-(\lambda)$ lie in the open left-half plane. It is well known that n-space E can be written as a direct sum $E = E_A{}^+ \oplus E_A{}^-$, where

$$E_A{}^+ = \{x : \phi^+(A)x = 0\}, \qquad E_A{}^- = \{x : \phi^-(A)x = 0\}.$$

$E_A{}^+$ thus represents the "unstable modes" of A. It is known [3] that (A, B) is stabilizable if and only if the range of $\Gamma(A, B)$ contains $E_A{}^+$.

Dual to the concept of controllability is that of observability: the pair of constant matrices (C, A) is *observable* if (A', C') is controllable. A weaker but useful property is that (at least) the unstable modes of A be observable. Precisely, (C, A) is *detectable* if (A', C') is stabilizable.

Controllability and observability are well-known concepts (cf. [7]); stabilizability is discussed in [3]; detectability in its present meaning originates here.

Of primary interest will be the Riccati equation

$$\begin{aligned} &\frac{dP(t)}{dt} + A(t)'P(t) + P(t)A(t) + \Pi[t, P(t)] \\ &\quad - P(t)B(t)N(t)^{-1}B(t)'P(t) + C(t)'C(t) = 0, \end{aligned} \tag{2.2a}$$

$$t_0 \leqq t \leqq T,$$

subject to the terminal condition

$$P(T) = P_T \geqq 0. \tag{2.2b}$$

In the constant parameter case we also consider the quadratic equation

$$A'P + PA + \Pi(P) - PBN^{-1}B'P + C'C = 0. \tag{2.3}$$

The main result is the following theorem.

THEOREM 2.1. *There exists a matrix $P(t)$ with the following properties:*

(i) *P is defined and absolutely continuous on $[t_0, T]$ and satisfies* (2.2a) (*almost everywhere*) *and* (2.2b).

(ii) *$P(t) \geqq 0$, $t_0 \leqq t \leqq T$, and $P(t)$ is the unique solution of* (2.2a, b).

(iii) (Minimum property). *Let $\tilde{K}(t)$ be an arbitrary* (*bounded measurable*) *$m \times n$ matrix defined on $[t_0, T]$ and let $\tilde{P}(t)$ be the solution of the linear equation*

$$\frac{d\tilde{P}(t)}{dt} + [A(t) - B(t)\tilde{K}(t)]'\tilde{P}(t) + \tilde{P}(t)[A(t) - B(t)\tilde{K}(t)] + \Pi[t, \tilde{P}(t)] + C(t)'C(t) + \tilde{K}(t)'N(t)\tilde{K}(t) = 0, \tag{2.4a}$$

$$\tilde{P}(T) = P_T. \tag{2.4b}$$

If $P(t)$ is the solution of (2.2a, b), *then $P(t) \leqq \tilde{P}(t)$, $t_0 \leqq t \leqq T$.*

(iv) *Let A, B, C, N and Π be independent of t and consider* (2.2) *with $t_0 = -\infty$, $T = 0$. If (A, B) is stabilizable and Π satisfies* (2.1), *then $P(t)$ is bounded on $(-\infty, 0]$. If in addition (C, A) is observable, then*

$$P_\infty = \lim P(t), \qquad t \to -\infty,$$

exists and is positive definite. In that case P_∞ is the unique positive semidefinite solution of (2.3), *and the matrix*

$$A - BN^{-1}B'P_\infty$$

is stable.

A proof is given in §3–§5. Results for the quadratic equation (2.3) are summarized in Theorem 4.1.

3. Existence, uniqueness and minimality.

LEMMA 3.1 (Monotone convergence). *Let $\{P_\nu ; \nu = 1, 2, \cdots\}$ be a sequence of $n \times n$ symmetric matrices such that $P_1 \leqq P_2 \leqq \cdots$ and $P_\nu \leqq P$, $\nu = 1, 2, \cdots$, for some P. Then $P_\infty = \lim P_\nu$, $\nu \to \infty$, exists and $P_\infty \leqq P$.*

The lemma is a special case of a result for positive operators in Hilbert space [4, p. 189]; the result holds also for a monotone decreasing sequence which is bounded below.

We turn to a proof of assertions (i) and (ii) of Theorem 2.1. Let

$$\psi(P, K) = (A - BK)'P + P(A - BK) + \Pi(t, P)$$

and

$$K(t) = N(t)^{-1}B(t)'P(t). \tag{3.1a}$$

Then (2.2a, b) become

$$\frac{dP(t)}{dt} + \psi[P(t), K(t)] + C(t)'C(t) + K(t)'N(t)K(t) = 0, \tag{3.1b}$$

$$t_0 \leqq t \leqq T,$$

$$P(T) = P_T. \tag{3.1c}$$

The key to solving (3.1a, b, c) is the observation that (3.1b) is linear in P and that the expression (3.1a) minimizes the left side of (3.1b), regarded as a function of K (cf. [5]). The latter statement results from the identity

$$\begin{aligned}(A - BK_0)'P + P(A - BK_0) + K_0'NK_0 \\ \equiv (A - BK)'P + P(A - BK) \\ + K'NK - (K - K_0)'N(K - K_0),\end{aligned} \tag{3.2}$$

where $K_0 = N^{-1}B'P$.

Now let $K(t)$ be arbitrary and let $\Phi(t, s)$ be the fundamental matrix associated with the matrix $A(t) - B(t)K(t)$, that is, Φ is determined by the equations

$$\frac{\partial\Phi(t, s)}{\partial t} = [A(t) - B(t)K(t)]\Phi(t, s), \qquad t_0 \leqq s, t \leqq T, \tag{3.3}$$

$$\Phi(t, t) = I.$$

Recall that $\Phi(s, t) = \Phi(t, s)^{-1}$, whence

$$\frac{\partial\Phi(s, t)}{\partial t} = -\Phi(s, t)[A(t) - B(t)K(t)]. \tag{3.4}$$

It is then easily checked that (3.1b) and (3.1c) are equivalent to the equation

$$\begin{aligned}P(t) = \Phi(T, t)'P_T\Phi(T, t) + \int_t^T \Phi(s, t)'\{\Pi[s, P(s)] \\ + C(s)'C(s) + K(s)'N(s)K(s)\}\Phi(s, t)\, ds, \qquad t_0 \leqq t \leqq T.\end{aligned} \tag{3.5}$$

The Volterra equation (3.5) has a unique integrable solution $P(t)$ which can be found by successive approximation. Consider the approximation sequence $\{P_\nu : \nu = 1, 2, \cdots\}$ with $P_1(t) \equiv 0$. Recalling the positivity of Π we see that $P_\nu(t) \geqq P_1(t)$ for all t; hence $P(t) = \lim P_\nu(t) \geqq 0$.

We solve the simultaneous equations (3.1a) and (3.5) as follows: Denote the right side of (3.5) by $\mathfrak{I}(K, P, t)$. Choosing K_1 arbitrarily, define P_1 to be the unique solution of

$$P_1(t) = \mathfrak{I}(K_1, P_1, t), \qquad t_0 \leqq t \leqq T.$$

Having defined $K_1, \cdots, K_\nu$, let P_ν be the solution of

$$P_\nu(t) = \mathfrak{I}(K_\nu, P_\nu, t), \qquad t_0 \leqq t \leqq T, \tag{3.6}$$

and define

$$K_{\nu+1}(t) = N(t)^{-1}B(t)'P_\nu(t). \tag{3.7}$$

From what was said previously, the matrices K_ν and P_ν are well-defined, measurable and bounded on $[t_0, T]$. Next we exploit the minimum property (3.2). For brevity let us write (3.1b) as

$$\frac{dP(t)}{dt} + \Psi\{P(t), K(t)\} = 0,$$

where

$$\Psi(P, K) = \psi(P, K) + C'C + K'NK.$$

Then (3.2), (3.6) and (3.7) yield

$$\begin{aligned}\frac{dP_\nu(t)}{dt} + \Psi\{P_\nu(t), K_{\nu+1}(t)\} &\leqq \frac{dP_\nu(t)}{dt} + \Psi\{P_\nu(t), K_\nu(t)\} \\ &= 0 \\ &= \frac{dP_{\nu+1}}{dt}(t) + \Psi\{P_{\nu+1}(t), K_{\nu+1}(t)\}, \\ &\qquad t_0 \leqq t \leqq T.\end{aligned}$$

Setting $Q(t) = P_\nu(t) - P_{\nu+1}(t)$, we have

$$\frac{dQ(t)}{dt} + \psi[Q(t), K_{\nu+1}(t)] + R(t) = 0$$

for a suitable matrix $R(t) \geqq 0$; and from this we obtain, as before, $Q(t) \geqq 0$. It follows that for each $t \in [t_0, T]$ the sequence of nonnegative matrices $\{P_\nu(t)\}$ is monotone nonincreasing and hence, by Lemma 3.1,

$$P(t) = \lim_{\nu \to \infty} P_\nu(t)$$

exists. Since

$$|P_\nu(t)| \leqq \sup\{|P_1(s)| : t_0 \leqq s \leqq T\}, \qquad \nu = 1, 2, \cdots,$$

it follows by (3.7) that the sequence $\{|K_\nu(t)|\}$ is uniformly bounded. Let $\Phi_A(t, s)$ be the fundamental matrix (cf. (3.3)) determined by A.

Then (3.6) is equivalent to

$$\begin{aligned} P_\nu(t) = \Phi_A(T,t)'P_T\Phi_A(T,t) + \int_t^T \Phi_A(s,t)'\{\Pi[s,P_\nu(s)] \\ - K_\nu(s)'B(s)'P_\nu(s) - P_\nu(s)B(s)K_\nu(s) + C(s)'C(s) \\ + K_\nu(s)'N(s)K_\nu(s)\}\Phi_A(s,t)\,ds. \end{aligned} \tag{3.8}$$

Applying the dominated convergence theorem to the integral in (3.8), we conclude that (3.8) holds with P_ν, K_ν replaced by P, K, where

$$\begin{aligned} K(t) &= \lim_{\nu\to\infty} K_\nu(t) \\ &= N(t)^{-1}B(t)'P(t). \end{aligned} \tag{3.9}$$

Equations (3.8) (with $P_\nu = P$, $K_\nu = K$) and (3.9) are equivalent to (3.1 a, b, c); hence existence of an absolutely continuous solution of (2.2 a, b) is established.

Uniqueness of the solution results from the fact that the function

$$\Psi(P,K) = \Psi(P, N^{-1}B'P)$$

satisfies a uniform Lipschitz condition in P in every domain $t_0 \leqq t \leqq T$, $|P| < \text{const.}$

Assertions (i) and (ii) of Theorem 2.1 have now been proved.

The minimum property (iii) is proved by using (3.2) in the same manner as before. Thus from the inequality

$$\Psi[P(t),K(t)] \leqq \Psi[P(t),\tilde{K}(t)]$$

together with (2.4 a, b) and (3.2), there follows

$$0 = \frac{d\tilde{P}(t)}{dt} + \Psi[\tilde{P}(t),\tilde{K}(t)] \leqq \frac{dP(t)}{dt} + \Psi[P(t),\tilde{K}(t)].$$

If $Q(t) = \tilde{P}(t) - P(t)$, then

$$\frac{dQ(t)}{dt} + \Psi[\tilde{P}(t),\tilde{K}(t)] - \Psi[P(t),\tilde{K}(t)] \leqq 0; \tag{3.10}$$

hence for a suitable matrix $R(t) \geqq 0$,

$$\frac{dQ(t)}{dt} + \psi[Q(t),\tilde{K}(t)] + R(t) = 0, \tag{3.11a}$$

$$Q(T) = 0. \tag{3.11b}$$

If (3.11 a, b) are written as an integral equation (cf. (3.5)) and solved, as before, by successive approximation, we obtain a sequence $Q_\nu(t)$ such

that $Q_\nu(t) \to Q(t)$, $\nu \to \infty$. Since we can choose $Q_0(t) \equiv 0$ it is easily seen that $Q_\nu(t) \geqq 0$ for all t, ν. This completes the proof of Theorem 2.1 (iii).

4. Solution of the quadratic equation (2.3). In this section we assume that the parameter matrices A, B, C, N and the operator Π are independent of t, and consider exclusively the quadratic equation (2.3).

THEOREM 4.1. *If (A, B) is stabilizable and (C, A) is detectable, and if Π satisfies condition (2.1), then (2.3) has at least one solution $\bar{P}$ in the class of positive semidefinite matrices. The matrix $A - BN^{-1}B'\bar{P}$ is stable. If in addition (C, A) is observable, then $\bar{P}$ is unique and $\bar{P} > 0$.*

For the proof we need three auxiliary results.

LEMMA 4.1. *Let*

$$C'C + D'D = F'F$$

and let G be an arbitrary matrix of suitable dimension.

(i) *If (C, A) is observable, then $(F, A + GD)$ is observable.*

(ii) *If (C, A) is detectable, then $(F, A + GD)$ is detectable.*

Proof. Let $\{\cdot\}$ denote the range of a matrix and $\mathfrak{N}(\cdot)$ the null space. It is easily seen that $\{\Gamma(A' + \hat{F}', F')\} = \{\Gamma(A', F')\}$ whenever $\{\hat{F}'\} \subset \{F'\}$. Also, if $x'F'Fx = 0$, then $x'C'Cx = x'D'Dx' = 0$, so that $\mathfrak{N}(F) \subset \mathfrak{N}(C) \cap \mathfrak{N}(D)$; taking orthogonal complements, we have

$$\{C'\} + \{D'\} \subset \{F'\}.$$

Thus $\{D'G'\} \subset \{F'\}$, so that

$$\{\Gamma(A' + D'G', F')\} = \{\Gamma(A', F')\} \supset \{\Gamma(A', C')\},$$

proving (i). For (ii), write $\hat{D}' = D'G'$ and let $A' + C'R'$ be stable. Since $\{C'R' - \hat{D}'\} \subset \{F'\}$, a matrix S' can be chosen such that

$$A' + \hat{D}' + F'S' = A' + C'R'.$$

The proof is complete.

The following remark will be useful: if (C, A) is detectable, then either A is stable or the matrix

$$W_t(A, C) = \int_0^t e^{sA'}C'Ce^{sA}\, ds$$

is unbounded on $0 \leqq t < \infty$. For if A is not stable, let λ be an eigenvalue of A with $\operatorname{Re}\lambda \geqq 0$ and eigenvector ξ. If ξ^* is the conjugate transpose of ξ,

$$\xi^* W_t(A, C)\xi = \int_0^t e^{2s\operatorname{Re}\lambda} \mid C\xi \mid^2 ds.$$

Suppose the integral is bounded. Then $C\xi = 0$, i.e., $CA^{\nu-1}\xi = \lambda^{\nu-1}C\xi = 0$,

$\nu = 1, \cdots, n$, so that

$$\operatorname{Re}\xi, \operatorname{Im}\xi \in \mathfrak{N}(\Gamma(A', C')').$$

If (C, A) is detectable, $\{\Gamma(A', C')\} \supset E_{A'}^{+}$ and taking orthogonal complements, we have

$$\mathfrak{N}\{\Gamma(A', C')'\} \subset (E_{A'}^{+})^{\perp} = E_{A}^{-}.$$

Thus, $\operatorname{Re}\xi, \operatorname{Im}\xi \in E_{A}^{+} \cap E_{A}^{-}$, namely $\xi = 0$, a contradiction.

LEMMA 4.2. *Let (C, A) be detectable and suppose the equation*

$$A'P + PA + \Pi(P) + C'C = 0 \tag{4.1}$$

has a solution $P \geqq 0$. Then A is stable. Let $\mathfrak{I}(R)$ be defined by

$$\mathfrak{I}(R) = \int_0^\infty e^{tA'}\Pi(R)e^{tA}\,dt,$$

and put $\mathfrak{I}^\nu(R) = \mathfrak{I}(\mathfrak{I}^{\nu-1}(R))$, $\mathfrak{I}^0(R) = R$, $\nu = 1, 2, \cdots$. If (C, A) is observable, then the series

$$\sum_{\nu=0}^{\infty} \mathfrak{I}^\nu(R) = (\mathfrak{s} - \mathfrak{I})^{-1}(R) \tag{4.2}$$

converges for every symmetric $n \times n$ matrix R. In that case, the solution P of (4.1) is unique and is given by

$$P = (\mathfrak{s} - \mathfrak{I})^{-1}\left(\int_0^\infty e^{tA'}C'Ce^{tA}\,dt\right).$$

Here $\mathfrak{s}$ denotes the identity operator.

Proof. From (4.1) there results the identity

$$P = e^{tA'}Pe^{tA} + \int_0^t e^{sA'}[\Pi(P) + C'C]e^{sA}\,ds, \qquad t \geqq 0. \tag{4.3}$$

Since (C, A) is detectable, the integral

$$Q(t) = \int_0^t e^{sA'}C'Ce^{sA}\,ds, \qquad t \geqq 0,$$

is bounded only if A is stable. Since $P - Q(t) \geqq 0$, the first assertion is proved. To prove the second, let $t \to \infty$ in (4.3) and write $Q = Q(\infty)$ to obtain $P = \mathfrak{I}(P) + Q$. Hence

$$P = \mathfrak{I}^{\nu+1}(P) + \sum_{k=0}^{\nu} \mathfrak{I}^k(Q), \qquad \nu = 1, 2, \cdots. \tag{4.4}$$

Since $Q \geqq 0$ there follows $\mathfrak{I}^k(Q) \geqq 0$; thus the last written sum is dominated by P, and the series of nonnegative matrices converges. Now suppose

(C, A) is observable. Then $Q > 0$, and convergence of the series with arbitrary $R \geqq 0$ in place of Q follows by linearity of $\mathfrak{I}$ and the fact that $R \leqq \rho Q$ for some $\rho < \infty$. Finally, every symmetric matrix R can be written in the form $R = R^+ - R^-$, where $R^+ \geqq 0$, $R^- \geqq 0$.[1] Since the operators $\mathfrak{I}^\nu$ are linear, convergence of (4.2) for arbitrary symmetric R is established. Uniqueness of P follows by invertibility of $\mathfrak{s} - \mathfrak{I}$.

LEMMA 4.3 (Minimum property). *Let $P \geqq 0$ satisfy (2.3). Let $Q \geqq 0$ and suppose that for some matrix J,*

$$(A - BJ)'Q + Q(A - BJ) + \Pi(Q) + C'C + J'NJ = 0. \tag{4.5}$$

If (C, A) is detectable, then $A - BJ$ is stable, and if (C, A) is observable, then $P \leqq Q$.

Proof. Let $C'C + J'NJ = F'F$. By Lemma 4.1 (with $D = N^{1/2}J$ and $G = -BN^{-1/2}$), the pair $(F, A - BJ)$ is detectable. Applying Lemma 4.2, we conclude that $A - BJ$ is stable. Setting $Q - P = V$ and using (3.2) we obtain

$$(A - BJ)'V + V(A - BJ) + \Pi(V) + S = 0 \tag{4.6}$$

for some $S \geqq 0$. Then

$$V = \mathfrak{I}(V) + R, \tag{4.7}$$

where $\mathfrak{I}$ is defined as in Lemma 4.2 (with $A - BJ$ in place of A, and $R = \int_0^\infty e^{\alpha(A-BJ)'} S e^{\alpha(A-BJ)}\, d\alpha$). Again, by Lemma 4.1, observability of (C, A) implies observability of $(F, A - BJ)$, and then Lemma 4.2 applied to (4.5) shows that

$$\sum_0^\infty \mathfrak{I}^\nu(I)$$

converges. Since

$$-|V|\mathfrak{I}^\nu(I) \leqq \mathfrak{I}^\nu(V) \leqq |V|\mathfrak{I}^\nu(I),$$

there follows $\mathfrak{I}^\nu(V) \to 0$, $\nu \to \infty$. Hence (4.7) yields

$$V = \mathfrak{I}^{\nu+1}(V) + \sum_{k=0}^{\nu} \mathfrak{I}^k(R), \qquad \nu = 1, 2, \cdots,$$

$$\to \sum_{k=0}^{\infty} \mathfrak{I}^k(R) \quad \text{as} \quad \nu \to \infty.$$

---

[1] Choose T so that $T'RT = D$, where $D = \operatorname{diag}(d_1, \cdots, d_n)$. Define $d_i^+ = \max(d_i, 0)$, $d_i^- = -\min(d_i, 0)$, $i = 1, \cdots, n$, $D^+ = \operatorname{diag}(d_i^+)$, $D^- = \operatorname{diag}(d_i^-)$. Then $R^+ = TD^+T'$, $R^- = TD^-T'$.

Thus $V \geqq 0$, and Lemma 4.3 is proved.

Turning to the proof of Theorem 4.1 we use, as before, quasi-linearization and successive approximations. Equation (2.3) is equivalent to the pair of equations

$$K = N^{-1}B'P, \tag{4.8a}$$

$$(A - BK)'P + P(A - BK) + \Pi(P) + C'C + K'NK = 0. \tag{4.8b}$$

If $K_0 = N^{-1}B'P$ and K is arbitrary, (3.2) yields the inequality

$$\begin{aligned}(A - BK_0)'P + P(A - BK_0) + K_0'NK_0 \\ \leqq (A - BK)'P + P(A - BK) + K'NK.\end{aligned} \tag{4.9}$$

First we solve (4.8b) for a suitable fixed matrix K. If $A - BK$ is stable, then (4.8b) is equivalent to

$$P = \int_0^\infty e^{t(A-BK)'}[\Pi(P) + C'C + K'NK]e^{t(A-BK)}\,dt. \tag{4.10}$$

Denote the right side of (4.10) by $f(K, P)$. Since

$$-|P|\Pi(I) \leqq \Pi(P) \leqq |P|\Pi(I),$$

condition (2.1) implies that for some K,

$$\left|\int_0^\infty e^{t(A-BK)'}\Pi(P)e^{t(A-BK)}\,dt\right| \leqq \theta\,|P|, \tag{4.11}$$

where $\theta \in (0, 1)$ is independent of P. Hence for this K the function $f(K, P)$ is a contraction mapping in P, and so (4.10) has a unique solution. For later reference we note that the approximating sequence $\{P_\nu\}$, defined by

$$P^{(1)} = 0, \qquad P^{(\nu)} = f(K, P^{(\nu-1)}), \qquad \nu = 2, 3, \cdots,$$

is monotone nondecreasing.

We can now solve the pair of equations (4.8a, b). By assumption there exists K_1 such that $A - BK_1$ is stable and (4.11) is true. Let P_1 be the solution of $P = f(K_1, P)$ and define $K_2 = N^{-1}B'P_1$. Next solve the equation

$$P = f(K_2, P)$$

by successive approximations. To see that this is possible observe that (4.9) and Lemma 4.3 imply that $A - BK_2$ is stable; hence $f(K_2, P)$ is defined. Now set

$$P_2^{(1)} = 0, \qquad P_2^{(\kappa+1)} = f(K_2, P_2^{(\kappa)}), \qquad \kappa = 1, 2, \cdots.$$

As before it follows that $P_2^{(\kappa)} \geqq 0$, $\kappa = 2, 3, \cdots$, and $\{P_2^{(\kappa)}\}$ is nondecreas-

ing. We shall show that

$$P_2^{(\kappa)} \leqq P_1 . \tag{4.12}$$

The inequality (4.9) (with $K_0 = K_2$ and $P = P_1$) implies

$$f(K_2, P_2^{(\kappa)}) \leqq -\int_0^\infty e^{t(A-BK_2)'}[(A - BK_2)'P_1 + P_1(A - BK_2) + \Pi(P_1) - \Pi(P_2^{(\kappa)})]e^{t(A-BK_2)}\, dt$$

$$= P_1 - \int_0^\infty e^{t(A-BK_2)'}\Pi(P_1 - P_2^{(\kappa)})e^{t(A-BK_2)}\, dt.$$

Thus if $P_2^{(\kappa)} \leqq P_1$, then $P_2^{(\kappa+1)} = f(K_2, P_2^{(\kappa)}) \leqq P_1$; since $P_2^{(1)} = 0$, (4.12) is true. It follows by Lemma 3.1 that the limit

$$P_2 = \lim_{\kappa\to\infty} P_2^{(\kappa)}$$

exists, and $0 \leqq P_2 \leqq P_1$.

Repeating this procedure we obtain sequences $\{K_\mu\}$, $\{P_\mu\}$ with $K_{\mu+1} = N^{-1}B'P_\mu$ and $0 \leqq P_{\mu+1} \leqq P_\mu$. Then

$$\bar{P} = \lim_{\mu\to\infty} P_\mu$$

exists. If

$$\bar{K} = \lim_{\mu\to\infty} K_\mu = N^{-1}B'\bar{P},$$

it is clear that $\bar{K}, \bar{P}$ satisfy (4.8a, b), and (4.10) shows that $\bar{P} \geqq 0$.

Lemma 4.3 implies that $A - B\bar{K}$ is stable. If (C, A) is observable, uniqueness of $\bar{P}$ in the class $P \geqq 0$ is an immediate result of the minimum property. Finally, Lemma 4.1 shows that $((C'C + K'NK)^{1/2}, A - BK)$ is observable if (C, A) is observable; then it is clear from (4.10) that $\bar{P} > 0$. Theorem 4.1 is proved.

5. Proof of Theorem 2.1 (iv): asymptotic behavior of the solution. In this section we prove assertion (iv) of Theorem 2.1. As in §4, the parameter matrices A, B, C, N and the operator Π are independent of t.

LEMMA 5.1. *Let (C, A) be observable, let $P_0 \geqq 0$, and suppose $P(t)$ satisfies the differential equation*

$$\frac{dP(t)}{dt} + A'P(t) + P(t)A + \Pi[P(t)] + C'C = 0, \qquad t \leqq 0, \tag{5.1a}$$

$$P(0) = P_0 . \tag{5.1b}$$

If there exists a constant matrix $P^* \geqq 0$ *such that*

$$A'P^* + P^*A + \Pi(P^*) + C'C = 0, \tag{5.2}$$

then

$$P(t) \to P^* \quad as \quad t \to -\infty. \tag{5.3}$$

Proof. By Lemma 4.2, P^* is the unique solution of (5.2). Let $\mathfrak{L}$ denote the operator defined by

$$\mathfrak{L}(P) = A'P + PA + \Pi(P).$$

From Lemma 4.2 we know that $\mathfrak{L}$ (regarded as a linear transformation on the $n \times n$ symmetric matrices) is nonsingular, and that $-\mathfrak{L}^{-1}(Q) \geqq 0$ if $Q \geqq 0$.

Since $\mathfrak{L}$ is linear and independent of t, it is enough to consider the homogeneous equation

$$\frac{dQ(t)}{dt} + \mathfrak{L}[Q(t)] = 0, \qquad t \leqq 0, \tag{5.4}$$

$$Q(0) = P_0,$$

and to show that

$$Q(t) \to 0 \quad \text{as} \quad t \to -\infty. \tag{5.5}$$

For this let

$$R(t) = \int_t^0 Q(s)\, ds, \qquad t \leqq 0,$$

and let $\bar{R}$ be the (unique) solution of

$$\mathfrak{L}(\bar{R}) + P_0 = 0. \tag{5.6}$$

It will be shown that $0 \leqq R(t) \uparrow \bar{R}$ $(t \downarrow -\infty)$. In fact, $Q(t) \geqq 0$ by (3.5), so that $R(t)$ is nondecreasing as t decreases. Integration of (5.4) yields

$$\frac{dR(t)}{dt} + \mathfrak{L}[R(t)] + P_0 = 0. \tag{5.7}$$

Setting $F(t) = \bar{R} - R(t)$, we obtain from (5.6) and (5.7),

$$\frac{dF(t)}{dt} + \mathfrak{L}[F(t)] = 0, \qquad t \leqq 0, \tag{5.8}$$

$$F(0) = \bar{R}.$$

Since $\bar{R} = -\mathfrak{L}^{-1}(P_0) \geqq 0$, it is clear from (5.8) (cf. (3.5)) that $F(t) \geqq 0$, $t \leqq 0$, that is, $0 \leqq R(t) \leqq \bar{R}$ and $R(-\infty) = \lim R(t)$, $t \to -\infty$, exists.

Next, (5.7) shows that dR/dT is bounded and then that d^2R/dt^2 is bounded. Since

$$R(-\infty) = \int_{-\infty}^{0} \frac{dR(t)}{dt}\, dt,$$

it follows that $Q(t) = dR(t)/dt \to 0$ as $t \to -\infty$. The proof is complete.

Next, consider the Riccati equation

$$\frac{dP(t)}{dt} + [A - BK(t)]'P(t) + P(t)[A - BK(t)] + \Pi[P(t)] + C'C + K(t)'N(t)K(t) = 0, \qquad t \leqq 0, \tag{5.9a}$$

$$K(t) = N^{-1}B'P(t), \tag{5.9b}$$

$$P(0) = P_0 \geqq 0. \tag{5.9c}$$

From §3 we know that (5.9a, b, c) have a unique solution $P(t) \geqq 0$.

LEMMA 5.2. (i) *If* (A, B) *is stabilizable and if* Π *satisfies* (2.1), *then the solution* $P(t)$ *of* (5.9a, b, c) *is bounded on* $(-\infty, 0]$.

(ii) *If* $P_0 = 0$, *then* $P(t)$ *is monotone nondecreasing as* t *decreases.*

Proof. Let $\hat{K}$ be a constant matrix such that $\hat{A} = A - B\hat{K}$ is stable, and let $\hat{P}(t)$ be the solution of (5.9a) and (5.9c) with $K(t) = \hat{K}$. By the minimum property (Theorem 2.1(iii)), $P(t) \leqq \hat{P}(t)$. It will be shown that $\hat{P}(t)$ is bounded for suitable $\hat{K}$. Now

$$\hat{P}(t) = e^{-t\hat{A}'}P_0\, e^{-t\hat{A}} + \int_t^0 e^{-(t-s)\hat{A}'}\{\Pi[\hat{P}(s)] + C'C + \hat{K}'N\hat{K}\}e^{-(t-s)\hat{A}}\, ds. \tag{5.10}$$

We solve (5.10) by successive approximation, setting $\hat{P}_0(t) \equiv 0$. Then

$$\hat{P}_{\nu+1}(t) \leqq \gamma I + \int_t^0 e^{-(t-s)\hat{A}'}\Pi[\hat{P}_\nu(s)]e^{-(t-s)\hat{A}}\, ds, \qquad \nu = 0, 1, \cdots, \tag{5.11}$$

where

$$\gamma = \sup_{t \leqq 0} | e^{-t\hat{A}'}P_0\, e^{-t\hat{A}} | + \left| \int_0^\infty e^{s\hat{A}'}(C'C + \hat{K}'N\hat{K})e^{s\hat{A}}\, ds \right|.$$

By (2.1), $\hat{K}$ can be chosen so that

$$\left| \int_0^\infty e^{s\hat{A}'}\Pi(I)e^{s\hat{A}}\, ds \right| = \theta < 1.$$

With this choice (5.11) yields, on iteration,

$$| \hat{P}_\nu(t) | \leqq \gamma(1 + \theta + \cdots + \theta^{\nu-1}) \leqq \gamma(1 - \theta)^{-1}, \qquad t \leqq 0, \quad \nu = 1, 2, \cdots.$$

Hence,

$$\hat{P}(t) = \lim_{\nu \to \infty} \hat{P}_\nu(t)$$

is a bounded function of t, and (i) follows.

To prove (ii) set $P_0 = 0$ and let $\Phi(t, s)$ denote the fundamental matrix associated with $A - BK(t)$ (cf. (3.3)). Then

$$P(t) = \int_t^0 \Phi(s, t)'\{\Pi[P(s)] + C'C + K(s)'N(s)K(s)\}\Phi(s, t)\, ds. \tag{5.12}$$

Let $\tau \geqq 0$ be fixed, and define

$$\tilde{K}(t) = K(t - \tau), \qquad t \leqq 0.$$

If $\tilde{\Phi}(t, s)$ is the fundamental matrix determined by $A - B\tilde{K}(t)$, then clearly

$$\tilde{\Phi}(t, s) = \Phi(t - \tau, s - \tau).$$

Let $\tilde{P}(t)$ be the solution of (5.9a) with K replaced by $\tilde{K}$ and $\tilde{P}(\infty) = 0$. Again by the minimum property (Theorem 2.1(iii)), $P(t) \leqq \tilde{P}(t)$, or

$$\begin{aligned} P(t) &\leqq \int_t^0 \tilde{\Phi}(s, t)'\{\Pi[\tilde{P}(s)] + C'C + \tilde{K}(s)'N(s)\tilde{K}(s)\}\tilde{\Phi}(s, t)\, ds \\ &= \int_{t-\tau}^{-\tau} \tilde{\Phi}(s + \tau, t)'\{\Pi[\tilde{P}(s + \tau)] + C'C \\ &\qquad + \tilde{K}(s + \tau)'N(s)\tilde{K}(s + \tau)\}\tilde{\Phi}(s + \tau, t)\, ds \qquad (5.13) \\ &\leqq \int_{t-\tau}^0 \Phi(s, t - \tau)'\{\Pi[\tilde{P}(s + \tau)] + C'C \\ &\qquad + K(s)'N(s)K(s)\}\Phi(s, t - \tau)\, ds, \end{aligned}$$

where we have set $\tilde{P}(s) \equiv 0$ for $s \geqq 0$. Now

$$\begin{aligned} \tilde{P}(s + \tau) &= \int_{s+\tau}^0 \tilde{\Phi}(\sigma, s + \tau)'\{\Pi[\tilde{P}(\sigma)] + C'C + \tilde{K}(\sigma)'N(\sigma)\tilde{K}(\sigma)\}\tilde{\Phi}(\sigma, s + \tau)\, d\sigma \\ &= \int_{s+\tau}^0 \Phi(\sigma - \tau, s)'\{\Pi[\tilde{P}(\sigma)] + C'C \\ &\qquad + K(\sigma - \tau)'N(\sigma)K(\sigma - \tau)\}\Phi(\sigma - \tau, s)\, d\sigma \\ &\leqq \int_s^0 \Phi(\sigma, s)'\{\Pi[\tilde{P}(\sigma + \tau)] + C'C + K(\sigma)'N(\sigma)K(\sigma)\}\Phi(\sigma, s)\, d\sigma. \end{aligned}$$

Writing $Q(s) = P(s) - \tilde{P}(s + \tau)$ and using (5.12), we see that

$$Q(s) \geqq \int_s^0 \Phi(\sigma, s)'\Pi[Q(\sigma)]\Phi(\sigma, s)\, d\sigma. \tag{5.14}$$

Denote the right side of (5.14) by $\mathcal{S}Q(s)$. Defining $\mathcal{S}^\nu$ by iteration, there results

$$Q(s) \geqq \mathcal{S}^\nu Q(s)$$

and

$$|\mathcal{S}^\nu Q(s)| \leqq \alpha(\beta|s|)^\nu/\nu!, \qquad \nu = 1, 2, \cdots,$$

for suitable constants $\alpha > 0$, $\beta > 0$ (which may depend on s). Letting $\nu \to \infty$ we obtain $Q(s) \geqq 0$, or

$$\tilde{P}(s+\tau) \leqq P(s), \qquad s \leqq 0.$$

Substituting this result in (5.13) and comparing with (5.12), we conclude that

$$P(t) \leqq P(t-\tau), \qquad t \leqq 0, \quad \tau \geqq 0,$$

and the proof is complete.

LEMMA 5.3. *If (A, B) is stabilizable and (C, A) is detectable, if Π satisfies (2.1), and if $P_0 = 0$, then the solution $P(t)$ of (5.9a, b, c) has the property*

$$\lim_{t \to -\infty} P(t) = \tilde{P}, \tag{5.15}$$

where $\tilde{P}$ is a positive semidefinite solution of (2.3).

Proof. By Lemmas 3.1 and 5.2, the limit in (5.15) exists and is positive semidefinite. Since $P(t)$ is bounded, (5.9a, b, c) show that the same is true of $dP(t)/dt$ and $d^2P(t)/dt^2$; then convergence of the integral

$$\int_{-\infty}^{0} \frac{dP(t)}{dt} \tag{5.16}$$

shows that $dP/dt \to 0$ as $t \to \infty$. The conclusion follows by inspection of (2.3) and (5.9a, b, c).

We turn to a proof of assertion (iv) of Theorem 2.1. Set $T = 0$. Boundedness of $P(t)$ follows from Lemma 5.2. If (C, A) is observable, then, by Theorem 4.1, (2.3) has a unique solution $\tilde{P} \geqq 0$. Set $\bar{K} = N^{-1}B'\tilde{P}$, $\bar{A} = A - B\bar{K}$; Theorem 4.1 implies that $\bar{A}$ is stable. Denote by $P^*(t)$ the solution of (5.9a) and (5.9c) with $K(t) = \bar{K}$. By the minimum property (Theorem 2.1 (iii)), the solution $P(t)$ of (5.9a, b, c) satisfies

$$P(t) \leqq P^*(t), \qquad t \leqq 0, \tag{5.17}$$

and by Lemma 5.1 (with A replaced by $\bar{A}$, P^* by $\tilde{P}$, and $C'C$ by $C'C + \bar{K}'N\bar{K}$),

$$P^*(t) \to \tilde{P} \quad \text{as} \quad t \to -\infty. \tag{5.18}$$

On the other hand, if $P_*(t)$ denotes the solution of (5.9a) and (5.9b)

with $P_*(0) = 0$, then by Lemma 5.3,

$$P_*(t) \to \bar{P} \quad \text{as} \quad t \to -\infty. \tag{5.19}$$

Hence the desired result will follow if we show that

$$P(t) \geqq P_*(t), \qquad t \leqq 0. \tag{5.20}$$

For this observe from (3.5) that

$$P(t) \geqq \int_t^0 \Phi(s, t)'\{\Pi[P(s)] + C'C + K(s)'N(s)K(s)\}\Phi(s, t)\, ds \tag{5.21}$$

Denote the right side of (5.21) by $Q(t)$. It will be shown that $Q(t) \geqq P_*(t)$. Write $K_*(t) = N^{-1}B'P_*(t)$ and $\Phi_*(t, s)$ for the fundamental matrix associated with $A - BK_*(t)$; and let $\tilde{P}(t)$ be the solution of (5.9a) with $K(t)$ as before and $\tilde{P}(0) = 0$. Then, by the minimum property,

$$P_*(t) \leqq \tilde{P}(t) \tag{5.22}$$

and

$$\tilde{P}(t) = \int_t^0 \Phi(s, t)'\{\Pi[\tilde{P}(s)] + C'C + K(s)'N(s)K(s)\}\Phi(s, t)\, ds. \tag{5.23}$$

Then (5.21) and (5.23) yield

$$P(t) - \tilde{P}(t) \geqq \int_t^0 \Phi(s, t)'\Pi[P(s) - \tilde{P}(s)]\Phi(s, t)\, ds$$

and this shows, as in the proof of Lemma 5.2 (ii), that

$$P(t) - \tilde{P}(t) \geqq 0, \qquad t \leqq 0. \tag{5.24}$$

Inequalities (5.22) and (5.24) yield (5.20). Combining (5.17) and (5.20), we have that

$$P_*(t) \leqq P(t) \leqq P^*(t), \qquad t \leqq 0. \tag{5.25}$$

Since the extreme terms of the inequality (5.25) both tend to $\bar{P}$ as $t \to -\infty$, the desired result is established.

6. Applications.

6.1. Stochastic control. An equation of type (2.2a) arises in optimal control of a linear system with state-dependent white noise and quadratic cost (cf. [6], where time-invariant control was discussed, leading to (2.3)). In this problem,

$$[\Pi(t, P)]_{ij} = \operatorname{tr}\{G_i(t)'PG_j(t)\}, \quad i, j = 1, \cdots, n,$$

for certain G_i. We mention that an obvious generalization to include

control-dependent white noise leads to (2.2a) with Π replaced by

$$\Pi + PB(\Gamma + N)^{-1}\Gamma(\Gamma + N)^{-1}B'P. \tag{6.1}$$

In (6.1), $\Gamma = \Gamma(t, P)$ is a function of the same type as Π. This case can be discussed in exactly the same way, if (3.1a) is replaced by

$$K = (\Gamma + N)^{-1}B'P.$$

6.2. Linear filtering. A well-known linear filtering scheme [7] leads to the following equation for the covariance matrix:

$$\frac{dP}{dt} = AP + PA' + FF' - (PC' + FG')(GG')^{-1}(PC' + FG')', \qquad t_1 \leqq t \leqq t_2, \tag{6.2}$$

$$P(t_1) \equiv P_0 \geqq 0.$$

It is clear that (6.2) is equivalent to (2.2a, b) after replacing in (6.2) t, t_1, t_2 by $T + t_0 - t$, t_0, T, respectively, setting $\Pi = 0$, and redefining matrices. Thus Theorem 2.1 shows that (6.2) uniquely determines the covariance matrix. If the parameter matrices are constants, then the limit property of Theorem 2.1 (iv) holds if $(A' - C'(GG')^{-1}GF', C')$ is stabilizable and $(H, A' - C'(GG')^{-1}GF')$ is observable, where

$$FF' - FG'(GG')^{-1}GF' = H'H.$$

In particular this is true if (C, A) is detectable, (A, F) is controllable, and

$$FF' - FG'(GG')^{-1}GF' \geqq \rho FF'$$

for some $\rho \in (0, 1]$. The latter result under strengthened hypotheses was reported in [7, §13.33].

REFERENCES

[1] J. E. Potter, *A matrix equation arising in statistical filter theory*, Rep. RE-9, Experimental Astronomy Laboratory, Massachusetts Institute of Technology, Cambridge, 1965.

[2] D. L. Kleinman, *On the linear regulator problem and the matrix Riccati equation*, Rep. ESL-R-271, Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, 1966.

[3] W. M. Wonham, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[4] L. V. Kantorovich and G. P. Akilov, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.

[5] R. Bellman, *Functional equations in the theory of dynamic programming, positivity and quasilinearity*, Proc. Nat. Acad. Sci. U.S.A., 41 (1955), pp. 743–746.

[6] W. M. Wonham, *Optimal stationary control of a linear system with state-dependent noise*, this Journal, 5 (1967), pp. 486–500.

[7] R. E. Kalman, *New methods in Wiener filtering theory*, Proc. First Symposium on Engineering Applications of Random Function Theory and Probability, John Wiley, New York, 1963, pp. 270–388.